Part 2

MEDIATION ANALYSIS IN RETs

9

Mediation Analysis in RETs: Basic Approaches

I do not seek answers, but rather to understand the question

- YOUNG CAINE – Kung Fu TV Series

INTRODUCTION

Mediation analysis is key to randomized explanatory trials. For purposes of program evaluation, mediation analysis can be divided into three segments. First, one wants to determine if the program meaningfully impacts the outcome(s) of interest. If it does not, the question becomes why not. If it does, the question becomes how can one make the program even more impactful. Second, one wants to determine if the mediators that a program targets are indeed relevant to the outcome(s). A secondary goal might be to identify new mediators that program designers can target. Finally, one wants to determine which mediators a program successfully affects. If the program fails to meaningfully affect a mediator, then the program activities need to be strengthened.

CLASSIC MEDIATION ANALYSIS

Popular approaches to mediation analysis include (a) the method of Baron and Kenny (1987), (b) the product coefficient approach, (c) Hayes' (2018) PROCESS approach, (d) the MacArthur Network approach, (e) causal mediation analysis and (f) SEM. Although they can be applied to more complicated scenarios, the methods are often described using one distal variable (T), one mediator (M), and one outcome (O) as shown in Figures 9.1 and 9.2 where T represents the treatment condition a person is assigned to. In Figure 9.1, the path for the effect of T on O is referred to as the **total effect** of the program on the outcome and is designated as *c*. It reflects the effect of the program on the outcome ignoring the mediator. In Figure 9.2, the mediator is added to the model and the path for the direct effect of T on the outcome is designated as *c*'. It reflects the effect of the treatment on the outcome *holding constant any effects the program has on the mediator and, in turn, that the mediator has on the outcome*. For example, T might be a program for weight reduction that reduces weight by 10 pounds, on average, over two months relative to a control group (path *c*). One mechanism the program might address to bring about weight loss is how much people exercise. This is the mediator in Figure 9.2. Path *c'*

is the effect of the program on weight loss when one renders this mechanism moot by statistically holding it constant. The value of path *c*' might now equal 6 pounds, indicating the mediator accounts for 4 pounds of the 10 pound weight loss.

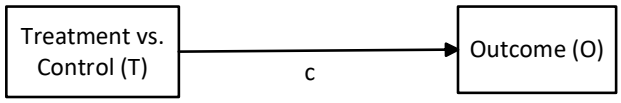


FIGURE 9.1. Total effect of treatment on outcome

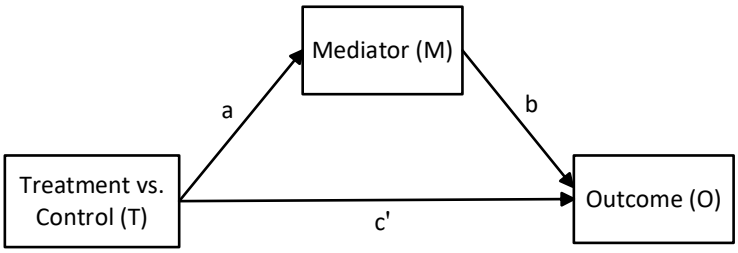


FIGURE 9.2. Single mediator, single outcome model

In this framework, mediation is conceptualized as the difference between the values of *c* and *c′*. If the two values are the same or close in value, then the supposed mediator, M, does not account for much of the effect of treatment on the outcome. If the two values are disparate, then M is said to mediate some of the effect of the program on the outcome. Estimates of *c* and *c′* can be obtained using standard regression methods, but I do not delve into the mathematics of doing so because, as it turns out, the strategy of comparing *c* and *c′* is limited. It is one way of thinking conceptually about mediation but it does not have much utility in practice unless one only has a single mediator in a three-variable system, which typically is not the case in RETs. Nevertheless, you will encounter mediation conceptualized in this fashion and it does indeed have intuitive appeal.

THE BARON AND KENNY METHOD

An early approach to mediation analysis is a three-step strategy developed by Baron and

Kenny (1987).[1] The steps are as follows:

Step 1: Evaluate if the distal variable is correlated with the outcome. Use O as the criterion and T as a predictor in a regression equation and estimate path *c* in Figure 9.1. This step establishes that there is a total effect worth explaining via mediation analysis.

Step 2: Evaluate if the distal variable is correlated with the mediator. Use M as the criterion variable and T as a predictor in a regression equation to estimate and test path *a* in Figure 9.2.

Step 3: Evaluate if the mediator affects the outcome variable. Use O as the criterion and T and M as predictors in a regression equation to estimate and test path *b* in Figure 9.2.

If affirmative results are obtained at each step (T is related to O, T is related to M, and M is related to O holding T constant), then mediation on the part of M is implied.

According to Baron and Kenny, it is not sufficient to simply correlate the mediator with the outcome at Step 3 to determine mediator relevance. The problem with such a strategy is that the mediator and outcome share T as an artificial common cause. T therefore must be controlled to assess the true M→O relationship by removing this artificial common cause of M and O. Although common cause confounding by T is indeed likely for simplistic single mediator models, it is not always the case that path *c′* should be included when evaluating mediation. In most RETs, there are multiple mediators that have been carefully mapped onto program structure and it is not unreasonable to believe that the mediators, taken as a whole, fully account for the program effect on the outcome. In such cases, path *c′* would be zero or trivial in magnitude so it can reasonably be omitted. For example, if a program uses an internet-based program to raise the discretionary monthly income of low-income families by educating people about (a) how to budget better and (b) how to use credit cards more effectively, it is not unreasonable to believe that any treatment-control difference in average monthly discretionary income is a function of only these two mediators, assuming proper random assignment. If one does not believe path *c′* is viable, then the path should not be part of the model (see James & Brett, 1984, for elaboration).

Another reason to be cautious about including path *c′* in an RET mediation model is when the strength of path *a* is strong, i.e., the program has a strong effect on the mediator, a result we hope to achieve when we design programs. In this case, when I regress O onto both M and T, the predictors M and T will be highly correlated because T has a strong impact on M. The result will be high multi-collinearity in the regression analysis when I

[1] A fourth step is often included to test for complete versus partial mediation, but I do not consider it here.

regress O onto both M and T. This multicollinearity can inflate coefficient standard errors, reduce statistical power, and inflate margins of error for the regression coefficients for M and T. Indeed, if the path *c′* is zero or near zero, the result can be an artifactual suppressor effect for T that disrupts interpretation. The bottom line is that in RETs, you should include path *c′* if you truly believe it is non-zero and meaningful. In my experience, in many RETs it is not needed because it is reasonable to assume the treatment only affects the outcome through the mediators the treatment targets.

Many analysts believe that the essential steps of Baron and Kenny for establishing mediation are Steps 2 and 3; that Step 1 is not required. One reason for the non-relevance of Step 1 is that mediation dynamics can operate in multi-mediator situations even when the estimated total effect of T on O (path *c* in Figure 9.1) is zero. This can happen if one of the mediators has a positive influence on the outcome and the other mediator has a negative influence on the outcome such that their opposing effects cancel each other, a case known as **opposing mediation**. An example is shown in Figure 9.3. It focuses on women in shelters for abused women and their intentions to leave their abusive partner as a function of the amount of physical abuse they have experienced from the partner in the past 6 months. One mediator is the belief on the part of the woman that if she stays in the relationship, she will be hurt again. Another mediator is the belief that if she tries to leave the relationship, her partner will hurt her. Both of these mediators are increased by past levels of abuse by the partner (paths *d* and *e*; I place positive signs on the paths to indicate the direction of influence). However, their respective effects on the outcome is an increased intention to leave the relationship via path *f* coupled with a decreased intention to leave the relationship via path *g*. These opposing dynamics might cancel each other out, producing no effect of past partner abuse on the intention to leave the relationship. Despite the zero or low correlation between past physical abuse and the intention to leave the relationship, we still want to learn about and address these important dynamics. Step 1 of the Baron-Kenny framework requires modification.

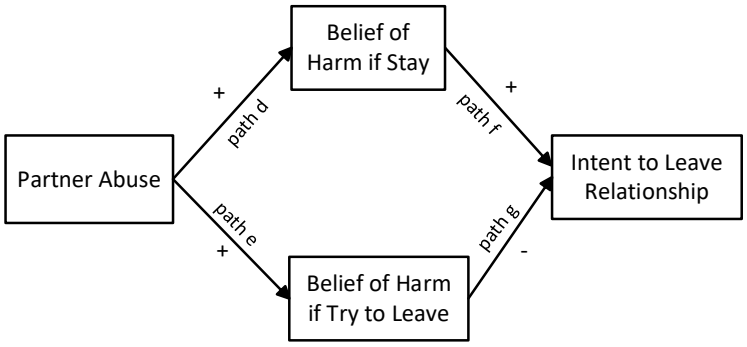


FIGURE 9.3. Example of opposing mediators

Another reason that Step 1 might be unnecessary is that sometimes the test of path *c* in Figure 9.1 will be statistically non-significant even though the tests of both path *a* and path *b* in Figure 9.2 are statistically significant. This represents conflicting signals; there appears to be no overall effect of T on O when the data are analyzed one way (the test of *c* in Figure 9.1), but there is an effect when the data are analyzed a different way (by testing paths *a* and *b* jointly in Figure 9.2; logically, if T impacts M and M impacts O, then T must influence O). It turns out that such conflicting results can happen because the two tests have different statistical power, with the joint test of paths having more power than the test of path *c* alone (Shrout & Bolger, 2002; Zhao, Lynch & Chen, 2010).

My own view is that in an adequately powered RET, even if the overall total effect of the program on the outcome is weak or non-significant, I still want to understand why this is the case. A careful analysis of the various links in the relevant mediational chains will give me insight into the underlying dynamics, as outlined in Chapter 1. I may not obtain an overall program effect because the mediators are not relevant to the outcome, contrary to program assumptions. Or, I may not obtain an overall program effect because, although the mediators are relevant, the program did not bring about sufficient change in those mediators. I do not want a failed Step 1 result to stop me from digging deeper into the operative causal dynamics that link the program to the presumed mediators and the presumed mediators to the outcome.

In sum, although the Baron and Kenny formulation has done much to advance work with mediation and the contribution of their seminal article has been important, I believe a stronger approach to the analysis of RETs is to use the SEM frameworks outlined in this book, especially since there invariably will be multiple mediators, multiple covariates, possible causal relationships among mediators, and possible correlated disturbances at work.

THE COEFFICIENT PRODUCT METHOD

A second popular strategy for testing mediation evaluates the product of all coefficients in a mediational chain, a process I mentioned in Chapter 5. In Figure 9.2, this approach tests the null hypotheses that the product of path *a* times path *b* is equal to zero in the population. If the test yields a statistically significant result, then one concludes that some mediation is present. Note that if any one of the links in the mediational chain is "broken," i.e., equals zero, then the product of the coefficients will equal zero.

A feature of the coefficient product approach is that it not only evaluates the null hypothesis of no mediation through a given mediational chain, it also produces an estimate of the *magnitude* of the omnibus mediational effect through that chain. For

example, the product of paths *a* and *b* in Figure 9.2 tells us for every one unit that the distal variable, T, changes, how many units the outcome variable is predicted to change through the mediational chain in question. In an RET, T is a dummy variable, so the coefficient product is the mean outcome difference between the treatment and control groups through the mediational chain in question.

Suppose a program seeks to reduce the number of packs of cigarettes people smoke per week by increasing the number of supportive friends people have to help them reduce smoking. Suppose the program increases the number of supportive friends (the mediator), on average, by 2 friends relative to the control group. This is path *a* in Figure 9.2. Suppose that for every additional friend that is added, the number of packs smoked per week declines by -0.50. This is path *b* in Figure 9.2. The product of the two paths is (2)(-0.50) = -1.0. This product is the mean outcome difference between the treatment and control condition due to this particular mediational chain; at posttest, those in the treatment condition smoked, on average, one less pack of cigarettes per week than those in the control group vis-à-vis the impact of the treatment on the number of supportive friends. If the test of this coefficient product is statistically significant, we conclude that the number of supportive friends mediates some of the effect of the program on smoking.

Many methods have been proposed for calculating standard errors, p values, and confidence intervals for the product of the coefficients. Three of the more common approaches are (1) traditional maximum likelihood (or OLS) in the form of the classic Sobel test, (2) a Sobel test but with robust maximum likelihood based on Huber-White estimation in place of the maximum likelihood estimator, and (3) bootstrapping (Bollen & Stine, 1990). For mathematical details of the three approaches see Muthén, Muthén and Asparouhov (2016). The latter two methods are generally superior to the first method because they do not rely on assumptions of normality and homoscedasticity. For the bootstrapping approach, there are different forms of bootstrapping that can be used. Simulation studies have compared percentile bootstrapping with bias corrected bootstrapping, with most studies concluding in favor of the use of the percentile method (e.g., Shrout & Bolger, 2002; Biesanz, Falk & Savalei, 2010; Falk & Biesanz, 2015). However, there are exceptions (e.g., Williams & MacKinnon, 2008). An advantage of the bootstrap approach over the robust maximum likelihood method is that the bootstrap approach can accommodate asymmetric confidence intervals. This can be important because the sampling distribution of the product of coefficients is not always symmetrical, in which case, the confidence interval should be asymmetric. In addition to these three approaches, other approaches include Bayesian methods (with either informative or uninformative priors using the highest posterior density method), bias corrected-accelerated bootstrap methods, semi-parametric bootstraps, residual-based

bootstrapping, profile-likelihood methods, Monte Carlo confidence interval methods, and model-based constrained optimization (Tofighi & Kelley, 2020), among others. The choice between these different methods is not straightforward. There have been many simulations to develop guidelines for method choice, but the recommendations are quite variable (e.g., Hayes & Scharkow, 2013; Huang, 2018; Leth-Steensen & Gallitto, 2016; MacKinnon et al., 2002; Mallinckrodt, Abraham, Wei & Russell, 2006;' Taylor, MacKinnon & Tein, 2008; Thoemmes, MacKinnon & Reiser, 2010; Tofighia & Kelley, 2020; Williams & MacKinnon, 2008; Valente, Gonzalez, Miocevic & MacKinnon, 2016). My reading of this literature is that authors often are quick to recommend a "best" method even when the performance in a simulation between the "best" method and the "next best" method is small, sometimes amounting to power differences like 0.82 versus 0.80 or Type I error rates of 0.044 versus 0.055 for the favored versus unfavored approach. Preacher (2015) in his *Annual Review* chapter concluded that the "methods agree more often than they disagree … thus, which method is chosen is often of little consequence." The best way to know if a given method will be appropriate for your particular RET is to conduct a localized simulation that maps onto the facets of the RET you have conducted or plan to conduct. I show how to conduct such simulations using Mplus in Chapter 28. Having said that, probably the most popular method for statistical significance testing of the product of coefficients and formulating confidence intervals for them is percentile bootstrapping.

The product coefficient method is reasonably straightforward with continuous mediators and continuous outcomes in the context of linear modeling. However, complications occur when the mediational chain has links that are mixtures of binary, ordinal, and/or nominal variables and when the ultimate outcome is binary, ordinal, or nominal. I outline these complications in future chapters and discuss how to address them.

THE JOINT SIGNFICANCE TEST

The joint significance test (JST) assumes a variable is a mediator of the effect of T on O if each link in a given mediational chain from the program to the outcome through the mediator is non-zero. By testing the statistical significance of each path in the chain, one makes an inference that all paths are non-zero if they are each statistically significant. One then concludes for mediation. If one of the paths is statistically non-significant, then the link is said to be "broken," leading to lack of evidence for mediation. The JST evaluates a null versus alternative hypothesis of no mediation versus mediation. It provides no information about the *magnitude* of the mediation. By contrast, the product

coefficient method provides information about both the null hypothesis test and the magnitude of the effect. For this reason, some social scientists prefer it to the JST. Counterarguments to this preference are that (a) the JST often does as good or a better job of testing the null hypothesis of no mediation than the product coefficient method in terms of Type I and Type II errors, (b) the JST is easier to apply when the mediational chain has links that are mixtures of binary, ordinal, and/or nominal variables, and (c) there are better ways of documenting mediation effect sizes than through the product of coefficients method; see Chapter XX.

For RETs, I personally emphasize a focus on the individual links in a mediational chain. I seek to determine if any of these links are "broken" and then make a judgment if the broken link is "repairable." I document the strength of each separate link using effect size indices and standards described in Chapter 10. If all the links in a mediational chain are statistically significant and reasonably strong, then I conclude for meaningful mediation by the target mediator. The overall omnibus effect size of the full mediational chain, as documented by the coefficient product approach, is of less utility for program evaluation purposes, a point I elaborate in Chapters XX and XX. The work needed to improve a program is usually determined from a link-by-link analysis, not an overall omnibus effect. Yzerbyt et al. (2018) report that omnibus mediation indices dominate reporting of mediational effects and lament that this "unfortunately means that researchers may not even look at, let alone test, the components of the indirect effect" (p. 940). To me, omnibus tests of mediators of program effects typically are of lesser import in RETs.

There is considerable support for the performance of the JST as a test of the null hypothesis of mediation and tests of moderated mediation (MacKinnon et al., 2002; Thoemmes, MacKinnon & Reiser, 2010; Huang, 2018; Leth-Steensen & Gallitto, 2016; Valente et al., 2016). Yzerbyt et al. (2018) conducted extensive simulations of the JST and different product coefficient methods for null hypothesis tests and concluded that "the joint-significance method constitutes the best compromise between Type I error rate and power and ought to be the method of choice" (p. 940). Biesanz et al. (2010) conducted a comparative simulation and also found reasonable performance of the JST relative to other approaches across normal and non-normal conditions.

One issue relevant to the application of the JST in SEM is that the path coefficients in the mediational chain can be dependent (MacKinnon, 2008; Valente et al., 2016). To elaborate, the logic of the JST derives, in part, from the idea that the joint probability of two events *a* and *b* equal the product of the probability of *a* times the probability of *b*. This relationship holds when *a* and *b* are independent. However, in some mediation modeling that uses SEM software, this will not be the case. Some methodologists argue

that failure to take the dependency into account might affect JST performance in certain contexts. Valente et al. (2016) used SEM to conduct a comparative simulation of the JST with a bootstrap version of the product coefficient method. They replicated a study by Leth-Steensen and Gallitto (2016) but they used a more comprehensive design. The tested models had negative dependencies in the coefficients comprising the target mediational chain, ranging from dependency correlations of -0.03 to -0.10. Both Valente et al. (2016) and Leth-Steensen and Gallitto (2016) found that the JST yielded mediation Type I error rates near (or just slightly below) the nominal alpha level of 0.05, so the dependency was not consequential for Type I errors. Valente et al. (2016) found in an SEM context that the Type I error rates were inflated for a biased corrected bootstrap method and that the power of the JST and the bootstrap approach were comparable. For example, for N = 400, the comparative powers for one set of mediation scenarios were 0.90, 0.92, 0.89 for the JST versus 0.91, 0.92 versus 0.93 for the bias corrected bootstrap method. These differences are slight. In sum the relatively good performance of the JST across a wide range of simulation studies suggest it is a reasonable candidate for mediation tests of the null hypothesis of no mediation (e.g., Hayes & Scharkow, 2013; MacKinnon et al., 2002; Taylor et al., 2008; Yzerbyt et al., 2018).

Hayes (2022; Montoya & Hayes, 2016) argues for the product coefficient method over the JST because the former uses a single test of the mediational effect rather than multiple tests in the JST, one for each link in the chain of effects. By conducting multiple tests, Hayes argues, the probability of at least one Type I error increases; the fewer tests conducted, the better. This logic ignores the fact that (a) the product coefficient approach itself is not straightforward,[2] and (b) simulation studies indicate that the JST usually strikes a better balance between Type I and Type II errors than the product coefficient approach (e.g., Yzerbyt et al., 2018), i.e., Hays argument is focused only on Type I errors. In short, the choice is more complex than the sheer number of tests involved.

Most simulation-based evaluations of the JST as well as the product coefficient method have used unrealistic scenarios with a single distal variable, a single mediator, and a single outcome. The fact is we simply do not know much about how the JST or the product coefficient approaches fare under more complex scenarios typical of RETs, especially when the variables have combinations of non-normal distributions, missing data, and mixtures of continuous, binary, and/or ordinal variables. To illustrate the complexity of mediational dynamics in RETs and why a simple three variable system is a gross oversimplification, consider the sequential mediation model in Figure 9.4 with three mediators (M1, M2, M3) and an outcome (Y) each measured at a posttest and a 6-

[2] For example, there are different types of bootstrapping that can be applied with their performance varying by context.

month follow-up (I omit covariates for the sake of pedagogy and to avoid clutter, but they further complicate the model). I designate the time of assessment by subscript *t* followed by a number (2 = the posttest, 3 = the follow-up). There are reasonable first order autoregressive and contemporaneous effects in the model. The treatment condition impacts the outcome at time 2 independent of the three mediators. Also, M1 influences M2 contemporaneously.

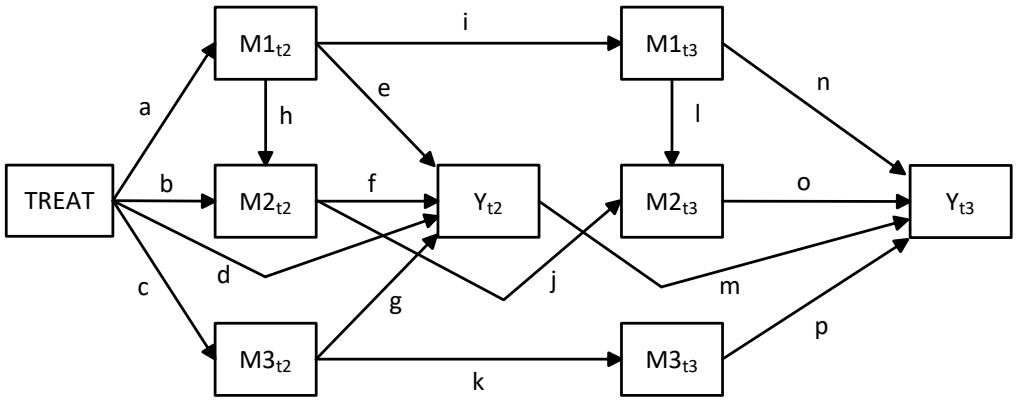


FIGURE 9.4. Example of a Treatment by Mediator Interaction

Suppose I want to map the mediation effect from the treatment (TREAT) through $M2_{t2}$ to the outcome at follow-up, Y_{t3}. It turns out there are five mediational chains from TREAT to Y_{t3} that include $M2_{t2}$ that are relevant to the mediation analysis. They are

TREAT → $M2_{t2}$ → Y_{t2} → Y_{t3}
TREAT → $M2_{t2}$ → $M2_{t3}$ → Y_{t3}
TREAT → $M1_{t2}$ → $M2_{t2}$ → Y_{t2} → Y_{t3}
TREAT → $M1_{t2}$ → $M2_{t2}$ → $M2_{t3}$ → Y_{t3}
TREAT → $M1_{t2}$ → $M1_{t3}$ → $M2_{t3}$ → Y_{t3}

If any one of these mediational chains does not have a "broken" link via a statistically nonsignificant path coefficient, then the JST would lead to an omnibus declaration that M2 mediates some of the effect of the treatment on Y_{t3}. The product coefficient method also uses the above five mediational chains to calculate the omnibus mediation effect of TREAT on Y_{t3} through M2 but using an additive combination of the product of path values. Using the path labels in Figure 9.4 the mediation effect is

mediation effect = (b)(f)(m) + (b)(j)(o) + (a)(h)(f)(m) + (a)(h)(j)(o) + (a)(i)(l)(o)

The product coefficient method must derive an estimated standard error and significance test for the above expression, all while accommodating possible non-normality, missing data, and mixtures of variable types (e.g., continuous, ordinal, binary) in the respective chains. This is no small feat, all for the purpose of making an omnibus mediation statement that, in my opinion, is not central to the objectives of RET analysis for program evaluation (see Chapters XX and XX).

In sum, the JST as a basis for evaluating the null hypothesis of no omnibus mediation versus some omnibus mediation seems to be a reasonable strategy, as good as or better than most other approaches in most contexts.

Parenthetically, if one applies both the JST and the product coefficient method to the same data, one sometimes obtains a statistically non-significant omnibus test despite the fact that the individual links within the target mediational chain all are statistically significant. Or, one might observe a statistically non-significant path within the JST tests but still obtain a statistically significant omnibus product coefficient test. Such disparities occur because the two tests are based on different statistical theories. One statistical theory is not necessarily better or "more correct" than the other in all contexts. If the tests agree in their conclusions, you might have more confidence in the conclusion. However, demanding that both tests simultaneously be "statistically significant" can lower statistical power and lower the actual Type I error rate below the a priori specified alpha level of 0.05. Such an orientation must be taken cautiously.

Extending the Joint Significance Test to Total Effects

The logic of the JST also can be extended to the analysis of total effects in RETs, that is, the estimated effect of the treatment on the outcome. In limited information SEM with continuous outcomes, the total program effect on the outcome is usually tested using a simple ANCOVA-like regression model predicting the outcome from a dummy variable for the treatment condition plus any relevant covariates, per my discussion in Chapter 8. This is straightforward. In full information SEM, however, one identifies all the causal chains in the model that link the treatment condition to the outcome and then evaluates each of them using either the JST or the coefficient product method. For the model in Figure 9.4, there are 10 such chains:

$\text{TREAT} \rightarrow \text{M2}_{t2} \rightarrow \text{Y}_{t2} \rightarrow \text{Y}_{t3}$
$\text{TREAT} \rightarrow \text{M2}_{t2} \rightarrow \text{M2}_{t3} \rightarrow \text{Y}_{t3}$
$\text{TREAT} \rightarrow \text{M1}_{t2} \rightarrow \text{M2}_{t2} \rightarrow \text{Y}_{t2} \rightarrow \text{Y}_{t3}$

$TREAT \rightarrow M1_{t2} \rightarrow M2_{t2} \rightarrow M2_{t3} \rightarrow Y_{t3}$
$TREAT \rightarrow M1_{t2} \rightarrow M1_{t3} \rightarrow M2_{t3} \rightarrow Y_{t3}$
$TREAT \rightarrow Y_{t2} \rightarrow Y_{t3}$
$TREAT \rightarrow M1_{t2} \rightarrow Y_{t2} \rightarrow Y_{t3}$
$TREAT \rightarrow M3_{t2} \rightarrow Y_{t2} \rightarrow Y_{t3}$
$TREAT \rightarrow M1_{t2} \rightarrow M1_{t3} \rightarrow Y_{t3}$
$TREAT \rightarrow M3_{t2} \rightarrow M3_{t3} \rightarrow Y_{t3}$

If any one of these mediational chains does not have a "broken" link via a statistically nonsignificant path coefficient, then the JST would make the omnibus declaration that the treatment condition has an effect on the program outcome Y_{t3}. The product coefficient method addresses the matter also using the ten chains but calculates the total effect as follows (using the path labels in Figure 9.4):

total effect = (b)(f)(m) + (b)(j)(o) + (a)(h)(f)(m) + (a)(h)(j)(o) + (a)(i)(l)(o) (d)(m) +
(a)(e)(m) + (c)(g)(m) + (a)(i)(n) + (c)(k)(p)

Again, deriving an estimated standard error and significance test for this expression while accommodating non-normality, missing data, and mixtures of variable types can be challenging. In addition, there are "anomalies" that can occur when using the FISEM approach based on product coefficients for total effects. For example, for a correctly specified model with no direct effect of the treatment to the outcome and for which the total effect and indirect effect are identical in value, the power for the test of the total effect can be dramatically smaller than the power for the test of the indirect effect, which seems contradictory (Kenny & Judd, 2014). Concretely, suppose in a single mediator model the population path from T→M (which I will refer to as path *a*) is 0.30 , the population path from M→Y (which I will refer to as path *b*) is 0.30 and the direct effect of T→Y (which I will refer to as path *c*) is zero. In this case, the mediation effect is (a)(b) = (0.30)(0.30) =0.09 and the total effect is (a)(b) + c = (0.30)(0.30) + 0 = 0.09, i.e., they are identical. In this case, the JST approach for testing if the total effect is non-zero will achieve power of 0.80 with a sample size of 114 but the product coefficient method that uses (*a*)(*b*) + *c* requires a sample size of 966 (Kenny & Judd, 2014). This power advantage of the JST approach over the product coefficient method is often large enough that greater power results for the JST even when the value of c is greater than zero (see O'Rourke & MacKinnon, 2015). For example, if *a* = .30, *b* = 0.30 and *c* = 0.06 and N is 200, the power of the JST approach is 0.98 but for the product coefficient method where the total effect is 0.15 instead of 0.09, the power is only 0.57 (Kenny & Judd, 2014). These same type of dynamics occur when comparing the JST to a simple t test of the

overall effect of the intervention, i.e., taking into account a third variable (M) vis-a-vis the joint significance test generally will yield more statistical power when evaluating the total effect than a simple t test comparing the treatment and control groups (Shrout & Bolger, 2002). Numerous other "anomalies" have been noted for the different approaches to mediational analysis and these are identified and explained mathematically in Wang (2018; see also Loeys, Moerkerke & Vansteelandt, 2015).

HAYES CONDITIONAL PROCESS ANALYSIS

Hayes (2018, 2022) has developed a set of specialized computer programs, called PROCESS, for mediation and moderation analysis that rely primarily on SPSS and SAS software (an R version was recently released as well). The approach has become popular in marketing and some subfields of psychology. It is a form of limited information SEM (LISEM), the advantages and disadvantages of which I discussed in Chapter 8. PROCESS is easy to use but it is somewhat limited in scope compared to more general SEM software and other LISEM approaches. PROCESS uses traditional OLS regression to estimate equations and creates omnibus mediational effect estimates using the product coefficient method with either the Sobel test or bootstrapping. In my opinion, it is better to use SEM because it is more flexible than PROCESS, but PROCESS can come in handy when the sample sizes are too small to accommodate FISEM; see Chapter 27.

THE MACARTHUR NETWORK MODEL

In 2000, the MacArthur Network on Developmental Psychopathology, led by Helena Kraemer, developed the MacArthur model for mediation and moderation analysis. The motivation was to clarify what the group considered to be ambiguities in the then dominant Baron and Kenny framework. The MacArthur model defines a moderator as a variable that identifies for whom or in what contexts T affects Y. Three conditions for a moderator must be met, (1) the moderator must temporally precede Y, (2) the moderator must be uncorrelated with T, and (3) if the population is stratified on different values of the moderator, the effect of T on Y should vary across one or more of the strata. Any measured baseline variable satisfies the first two conditions, so such variables automatically represent candidates for moderation. All one needs to show is that the third condition is met for them, using methods described in the third section of this book. For example, if the effect of T on Y varies as a function of biological sex, then sex is a moderator because it meets the other two conditions as well.

In the MacArthur Network model, a mediator of T on Y explains how or why T has an effect on Y. Four conditions for a mediator must be met, (1) the mediator must

temporally precede Y, (2) T must precede the mediator, (3) T must impact the mediator, and (4) the effect of T on Y can be explained, wholly or in part, by the mediator.

The MacArthur Network model deviates from other frameworks in its insistence that the same variable cannot be both a moderator and a mediator of the effect of T on Y. This is because the framework *defines* moderators as being independent of T while it defines mediators as following from and, hence, associated with T. As an example, it might be found that social support as measured at baseline moderates the effect of a program to reduce smoking during pregnancy; the program is more effective at reducing smoking for women with higher initial levels of social support. It also is possible that the program affects the amount of social support women receive to reduce smoking and that this, in turn, leads to post-program smoking reductions. In the MacArthur Network model, social support at baseline is viewed as distinct from social support during treatment, with the former acting as a moderator and the latter acting as a mediator. It is not the case that social support is acting as a moderator and a mediator. Rather, the two measures of social support represent different variables because they are measured at different points in time.

There are scenarios in the MacArthur Network model where an interaction effect between two variables, which is traditionally associated with the concept of moderation, is said to be mediation. Consider the example in Figure 9.5 for a program to increase monetary donations for victims of a natural disaster by making the suffering of victims salient to potential donors. Suppose the program impacts the mean belief about victim suffering relative to the control condition (path *a*). However, there is a second dynamic at play; the program also makes victim suffering more salient to potential donors independent of creating mean changes in it. Stated another way, the program *increases the value of the path coefficient* for the impact of the belief on the amount donated for the treatment group but not the control group (see paths *b* and *d*). The MacArthur Network model would frame all such effects as mediation, despite the fact that other frameworks would say there is moderation in the form of moderated mediation vis-à-vis a treatment by mediator interaction. Note that the causal dynamics are not different in the MacArthur Network model than in other frameworks; the dynamics are captured by Figure 9.5. It is more a matter of labeling what those dynamics are called, mediation or moderation.

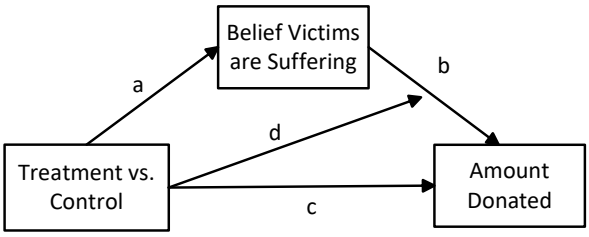


FIGURE 9.5. Example of a treatment by mediator interaction

In the MacArthur model, if two mediators interact with one another (in a statistical sense) to impact an outcome, then the interaction effect also is said to reflect mediation, not moderation. In other frameworks, the interaction effect is moderation because the effect of a mediator on the outcome depends on the value of the other mediator. Consider the model in Figure 9.6 for a program that reduces stress (path *a*) and increases positive coping skills (path *b*). The mechanism by which coping skills is thought to impact the outcome, anxiety, is by weakening the impact of stress on anxiety (paths *c* and *d*). In frameworks other than the MacArthur Network, coping skills is said to moderate the impact of stress on anxiety. However, the MacArthur Network model would not characterize coping as a moderator because it is impacted by T. Again, the causal dynamic is the same in all frameworks but the labels differ.

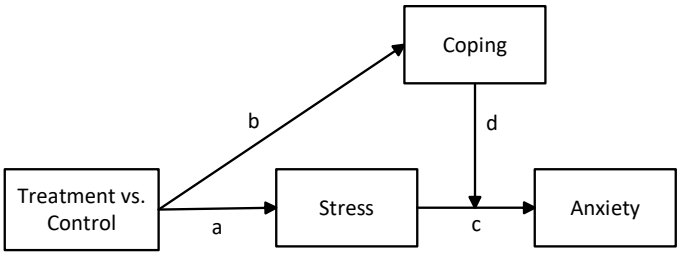


FIGURE 9.6. Example of mediator interaction

In sum, the MacArthur Network Model distinguishes itself primarily in the way it defines and conceptualizes mediation and moderation. Analytically, it is relatively agnostic as to the statistical approaches used for mediation analysis, although most

applications of it do not use SEM.[3] I find the terminology used in the framework a bit awkward but it nevertheless is firmly entrenched in several literatures. Kraemer et al. (2008) formally compare the MacArthur approach to the Baron and Kenny approach to mediation and moderation but I personally find their characterizations of the latter to be too narrow and overly strict relative to Kenny's orientations across multiple studies.

CAUSAL MEDIATION ANALYSIS

A sixth method for mediation modeling is often referred to as **causal mediation analysis**. The label is somewhat misleading because all of the approaches I consider seek to make causal statements about mediation, but the phrase nevertheless is increasingly used to refer to a particular approach to mediation analysis. Some people associate the label with Pearl's structural causal modeling (SCM) approach, while others associate it with what is known as a potential outcomes or counterfactual framework in public health and epidemiology (Imai, Keele & Tingley, 2010; Robins, 2003; VanderWeele, 2016). SCM and potential outcomes modeling have much in common but there also are differences between them (see Bollen & Pearl, 2013). I emphasize here the SCM formulation. Causal mediation analysis also is linked to structural equation modeling, but it typically uses limited information rather than full information estimation. In the following discussion, I assume you have read the material on SCM in Chapter 8.

SCM embraces somewhat different perspectives on total effects, direct effects and indirect/mediated effects than traditional mediation analyses. Most traditional forms of mediation analysis seek to decompose a total effect into (a) the indirect effect of a distal variable on an outcome through one or more mediators and (b) the direct effect of the distal variable on the outcome that is the effect of the variable on Y holding constant the mediators. SCM approaches the decomposition process differently than traditional mediation analysis does. In this section, I highlight the similarities and differences. I illustrate the SCM approach using a treatment versus control condition (T) as the distal variable, a single continuous mediator (M), and a single continuous outcome (Y). I use the notation for SCM from Chapter 8, but simplify it for you, expressing means using the more familiar symbols of $\overline{Y}$ and $\overline{M}$ rather than expectation notation.

Figure 9.7 presents the example mediation model I use. A program is designed to increase the amount of money that middle income young adults save towards retirement by educating them about the benefits of setting aside a portion of their monthly income for retirement purposes. The outcome is the amount in dollars that participants set aside

[3] Some argue that there are indeed prescribed analytic approaches in the MacArthur framework (see Kraemer et al., 2008), but in practice, studies that claim to use the framework are often quite variable in the methods they employ.

per month for retirement as measured over the course of the ensuing year after program completion. The mediator of perceived benefits is a multi-item scale where each item describes a benefit the person could receive from saving towards retirement. The items are rated on a -3 to +3 disagree-agree metric in terms of whether the individual believes s/he will obtain the benefit. The item responses are averaged so the total score ranges from -3 to +3, with higher scores indicating more perceived benefits. A (passive) control group received no education about the benefits. To keep matters simple for purposes of exposition, there are no baseline measures or covariates.

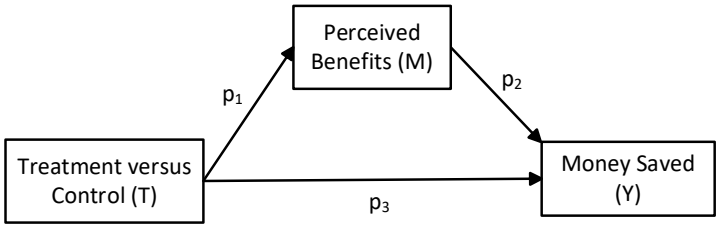


FIGURE 9.7. Simple mediation model (disturbances omitted for clarity)

Here are some summary statistics that I will use to make concepts concrete. First, the mean perceived benefits (the mediator) for the treatment and control groups are 2.0 and -1.0, respectively. This means the value of p_1 in Figure 9.7.is 3.0, which is the difference between these two means. Second, based on a linear regression of money saved (Y) onto perceived benefits (M) and the treatment condition (T), the path coefficient p_2 reflecting the effect of perceived benefits on money saved is found to be \$25; for every one unit the perceived benefit scores increase, the amount of money saved towards retirement per month increases by \$25. Third, based on this same regression analysis, the direct effect of the treatment on the outcome independent of the mediator, p_3, is \$20; holding constant perceived benefits, people in the intervention condition saved \$20 more per month than people in the control condition, probably because the program made saving towards retirement salient. The equation for the linear regression is $Y = 90 + 25M + 20T$. Finally, the Y outcome means for the treatment and control groups are \$160 and \$65 per month, respectively, so the program positively impacted retirement savings by increasing it, on average, by \$95 per month. With these statistics in mind, I now develop core mediation concepts in the SCM and causal mediation frameworks.

Three Types of Direct Effects

In the literature on causal mediation and SCM, the term *direct effects* has been used in different ways. Sometimes the term "direct effect" refers to the direct effect of a mediator on an outcome controlling for one or more confounds (p_2 in Figure 9.7); other times the term refers to the direct effect of the treatment condition on the outcome controlling for one or more mediators (p_3 in Figure 9.7). Yet other times, the term refers to the total effect of a treatment on an outcome controlling for confounds, such as those that might result from sample imbalance. I sometimes find these multiple uses somewhat confusing, but there is a coherence to them that I do not want to get sidetracked on here. My focus is on defining direct effects per p_3 in Figure 9.7. In traditional SEM, p_3 is the estimated effect of the program on the outcome *independent of the mediators in the model*; that is, if we hold the measured mediators that a program targets constant, does the treatment still have an impact on the outcome through unmeasured mediators? The causal mediation framework distinguishes three variants of p_3. Terminology varies in the literature, so I use the distinctions outlines by Mplus (Muthén, Muthén & Asparouhov, 2016).

One variant of p_3 is called a **controlled direct effect** (CDE). It focuses on the case where an investigator is interested in evaluating the direct effect of T on Y when holding the targeted mediator constant at a specific value of *a priori* interest to the investigator. It is formally defined by Pearl as

$$\text{CDE}(M = m) = (\bar{Y}_{TREAT}|M=m) - (\bar{Y}_{CTRL}|M=m) \quad [9.1]$$

where CDE(M = *m*) is read as "the controlled direct effect of T on Y when the mediator has the (same) value *m* in the treatment and control groups." For example, I might want to know what the mean difference is in retirement savings for the treatment and control groups when perceived benefits takes on a score of 0, the neutral point (neither agree nor disagree) on the disagree-agree metric of the perceived benefits scale. It turns out that I can use the regression equation Y = 90 + 25M + 20T that I presented above to calculate the mean outcome for the treatment group when *m* = 0. I substitute the values of 1 for T and 0 for M in the equation, like this:

$$Y = 90 + 25(0) + 20(1) = \$110$$

which yields the predicted mean I seek. I can do the same for the control group but now using a value of 0 for T, which yields

$$90 + 25(0) + 20(0) = \$90$$

The difference between these means, per Equation 9.1 is the value of CDE(M = 0), which is $20. This suggests that T has an effect on Y independent of M when perceived benefits are "neutral" because I held M constant at a value of 0. Researchers sometimes evaluate direct effects of a treatment on an outcome at different *a priori* specified values, such as when M takes on an *a priori* specified "low" score, an *a priori* specified "moderate" score, and an *a priori* specified "high" score.

A second variant of p_3 is a **pure natural direct effect** (PNDE). It takes the same form as Equation 9.1 for the controlled direct effect but with a notable exception. Instead of the researcher choosing an *a priori* value to hold the mediator constant at, the mediator is held constant at what is thought to be the "typical" value of the mediator for the control group. A "typical" value usually is defined as the mean of the mediator, but it could be something else, such as the median or mode of the mediator. Here is the relevant formula:

$$\text{PNDE}(M = m_{CTRL}) = (\bar{Y}_{TREAT}|M=m_{CTRL}) - (\bar{Y}_{CTRL}|M=m_{CTRL}) \quad [9.2]$$

where m_{CTRL} is the mean of the mediator for the control group. The PNDE is the expected change in the outcome (reflected by $\bar{Y}_{TREAT}$ minus $\bar{Y}_{CTRL}$ in Equation 9.2) if we "freeze" the typical mediator value at the level it would take for people not exposed to the program, i.e., at the value $M=m_{CTRL}$. In the retirement savings example, the mean perceived benefits for the control group was -1. Just as I did with the CDE, I can compute the predicted retirement savings for the treatment group when the perceived benefits = -1 by substituting into the original regression equation a score of 1 for T and -1 for M:

$$Y = 90 + (25)(-1) + (20)(1) = \$85$$

and for the control group it is

$$Y = 90 + (25)(-1) + (20)(0) = \$65$$

The difference between them is $20 and this is the value of the pure natural direct effect.

The third variant of p_3 is called a **total natural direct effect** (TNDE). It is defined the same way as the pure natural direct effect but instead of conditioning on the value of the mediator mean for the control group, we condition on the mediator mean *for the treatment group*:

$$\text{TNDE}(M = m_{TREAT}) = (\bar{Y}_{TREAT}|M=m_{TREAT}) - (\bar{Y}_{CTRL}|M=m_{TREAT}) \quad [9.3]$$

The TNDE is the expected change in the outcome (reflected by $\bar{Y}_{TREAT}$ minus $\bar{Y}_{CTRL}$ in Equation 9.3) if we "freeze" the typical mediator value at the level it would be if people were to be exposed to the program. I can again calculate the value of the TNDE using the

regression equation Y = 90 + 25M + 20T. Recall that the mean perceived benefits for those in the treatment group was 2.0. I substitute this value for M into the equation and a value of 1 for T, yielding

$$Y = 90 + (25)(2) + (20)(1) = \$160$$

and for the control group it is

$$Y = 90 + (25)(2) + (20)(0) = \$140$$

The difference between them is $20 and this is the value of the total natural direct effect.

The TNDE and PNDE are conceptually distinct because they differ on the value of the mediator that is held constant. Of substantive interest is whether the results for the TNDE and PNDE are comparable; if so, this signifies generalizability of the direct effect across the different values. It is analogous to calculating the value of the direct effect of T on Y for different predictor profiles per my discussion in Chapter 5 to determine if results generalize across profiles. In the current example, the values of the TNDE and PNDE are identical, namely both were $20. It turns out that in a linear system with no interaction effects and a continuous mediator and continuous outcome, this will always be the case. Because of this property, the distinctions between the types of direct effects in such scenarios are statistically moot and we typically just refer to them as **direct effects**, per SEM tradition. When relationships are non-linear or there are moderated relationships in the model, the values of PNDE and TNDE can differ and we then need to elaborate the substantive implications of those differences, a topic I address in future chapters.

Two Types of Indirect Effects

The SCM and causal mediation literature also distinguishes two variants of the indirect effect in Figure 9.7. Traditionally, an indirect effect is captured by paths p_1 and p_2 in the figure. In the product coefficient approach, p_1 and p_2 are multiplied by one another to determine the indirect effect of T on Y through M. In our example, the indirect effect of the program on retirement savings through the perceived benefits mediator equals (3)(25) or $75. In SCM and the causal mediation approach, the first variant of an indirect effect is called a **total natural indirect effect** (TNIE). It is defined as

$$\text{TNIE} = (\bar{Y}_{\text{TREAT}}|M=m_{TREAT}) - (\bar{Y}_{\text{TREAT}}|M=m_{CTRL}) \quad [9.4]$$

where m_{TREAT} is typically defined as the mean of the mediator for the treatment group, which is 2.0 in the retirement savings example, and m_{CTRL} is the mean of the mediator for the control group, which is -1.0. TNIE evaluates the difference between the expected

posttest value of Y (i.e., $\bar{Y}_{TREAT}$) as we move the value of the mediator from the expected value of the mediator given non-exposure to the program (M=m_{CTRL}) to the expected value of the mediator given exposure to the program (M=m_{TREAT}).

I can again use the regression equation Y = 90 + 25M + 20T to calculate the relevant values in the expression. The value of $\bar{Y}_{TREAT}$ when m = 2.0 is

90 + 25(2) + 20(1) = \$160

and the value of $\bar{Y}_{TREAT}$ when m = -1.0 is

90 + 25(-1) + 20(1) = \$85

The difference between these conditional means is \$160 - \$85 = \$75 and defines the value of the TNIE. Note that this value equals the classic definition of an indirect effect by the product coefficient method, namely $(p_1)(p_2)$ = (3)(25) = \$75. Again, it reflects what the treatment mean is when we change the mediator from its typical, expected value given non-exposure to the program (M=m_{CTRL}) to what its expected value would be given program exposure (M=m_{TREAT}).

The second variant of the indirect effect is called the **pure natural indirect effect** (PNIE). It is the same expression as TNIE but now $\bar{Y}_{CTRL}$ is substituted for $\bar{Y}_{TREAT}$:

$$\text{PNIE} = (\bar{Y}_{CTRL}|M=m_{TREAT}) - (\bar{Y}_{CTRL}|M=m_{CTRL}) \quad [9.5]$$

The PNIE is the expected difference in the outcome when the mediator changes from the expected or "typical" value it would take given non-exposure to the treatment program (M=m_{CTRL}) to the expected or "typical" value the mediator given program exposure (M=m_{TREAT}) for those in the control condition.

For the PNIE, the mean $\bar{Y}_{CTRL}$ is calculated under the two mediation scenarios depicted in Equation 9.5 again using the linear equation of Y = 90 + 25M + 20T. For the first expression on the right side of the equation where m = 2.0, the predicted $\bar{Y}_{CTRL}$ is

90 + 25(2) + 20(0) = \$140

and for the second expression on the right side when m = -1.0, it is

90 + 25(-1) + 20(0) = \$65

The difference between these values is \$140 - \$65 = \$75, which is the same value we obtained for the total natural indirect effect, TNIE.

Like the TNDE and PNDE, the TNIE and PNIE are conceptually distinct. In both

cases, we vary the likely value of the mediator to what it would be if people are exposed to the intervention versus what it likely would be if people are not exposed to the intervention. However, for the TNIE, we document the effect of this variation in the mediator on Y in the treatment condition whereas for the PNIE, we document the effect of this variation in the mediator on Y in the control condition. Of interest is whether these constructs take on the same or different values. If the values are discrepant, one then asks what are the substantive implications of the disparity and which of the two are we more interested in. It turns out that in a linear system with no interaction effects and a continuous mediator and continuous outcome, the PNIE always will equal the TNIE, so the distinctions between them are statistically moot. This was true for our current example. In such cases, the tradition is to refer to both as **indirect effects**, per standard SEM terminology. However, cases can be encountered where the values differ.

In sum, there are similarities between the traditional SEM language of direct and indirect effects in mediation analysis and the language of direct and indirect effects in causal mediation analysis vis-à-vis the potential outcomes framework. Often the frameworks lead to the same conclusions and the same parameterizations. But sometimes they do not and I will discuss these divergences in later chapters. Some researchers view the distinctions between the different types of effects in the potential outcomes framework as meaningful and important, while others treat them with suspicion as to their practical value (e.g., Naimi, Kaufman & MacAclehose, 2014). I tend to be neutral on the matter, seeing utility in some contexts but not others.

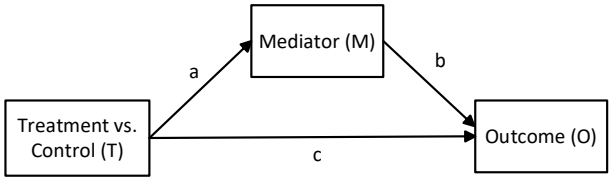
Broader Perspectives on the Causal Mediation Approach

Counterfactual conceptions of causality are core to the causal mediation and SCM frameworks. The presence of counterfactual thinking is evident in all of the above definitions of direct and indirect effects. For pure natural direct effects, for example, the conditional expectation $\bar{Y}_{CTRL}|M=m_{CTRL}$ represents a single "world" in the sense that the two components of the conditional expectation ($\bar{Y}_{CTRL}$ and $M=m_{CTRL}$) are features of the same "world," i.e., both are expressions about the control group. By contrast, the conditional expectation $\bar{Y}_{TREAT}|M=m_{CTRL}$ represents what is known as a **cross-world counterfactual** because it represents a "world" consisting of a treatment group parameter to the left of the | symbol coupled with a parameter from a different "world," namely a control group parameter. The parameters "cross worlds" to represent a counterfactual. The total natural direct effect consists of one conditional expectation ($\bar{Y}_{TREAT}|M=m_{TREAT}$) whose components are from the "same world" and a cross-world counterfactual $\bar{Y}_{CTRL}|M=m_{TREAT}$, whose components are from different "worlds." The total natural indirect effect consists of a "same world" conditional expectation ($\bar{Y}_{TREAT}|M=m_{TREAT}$) and

a cross-world counterfactual $\bar{Y}_{TREAT}|M=m_{CTRL}$. Finally, the pure natural indirect effect consists of a "same world" conditional expectation ($\bar{Y}_{CTRL}|M=m_{CTRL}$) and a cross-world counterfactual ($\bar{Y}_{CTRL}|M=m_{TREAT}$). The causal mediation framework assumes that cross-world counterfactuals are meaningful, an assumption that some researchers question (Meehl, 1970; Naimi et al, 2014). For example, some argue that perhaps it is not the case that $M=m_{CTRL}$ seamlessly generalizes to equal the mean of the mediator for the intervention group had people in that group not participated in the intervention vis-à-vis the expression $\bar{Y}_{TREAT}|M=m_{CTRL}$.

Much of the focus of the causal mediation framework in the research literature is on characterizing omnibus indirect effect coefficients through a given mediational chain, just as this is so using the classic product coefficient method in SEM. As noted in Chapter 1 and in future chapters, I tend to shy away from such omnibus characterizations of mediational chains. I argue instead that one gains greater insights into the underlying mediational dynamics by analyzing the individual links within a chain and then piecing together the omnibus implications from separate-link analyses. Knowing just the overall omnibus effect for a given mediator is insufficient because it conflates the effect of the treatment on the mediator with the effect of the mediator on the outcome. We need to dig deeper than this. Some researchers argue that documenting the omnibus mediation effect for each mediator gives insights into the relative importance of the mediators in shaping the outcome (Hayes, 2022). I question such logic in Chapter 17. Despite my inclinations, I make use of some concepts from the causal mediation framework in future chapters.

A non-trivial limitation of the causal mediation framework is that it has difficulty accommodating multiple mediators with causal relationships among one or more of the mediators as well as accommodating certain forms of correlated disturbances. The typical application of causal mediation analysis resorts to a series of single mediator analyses that consider the mediational dynamics of one mediator at a time with no correlated disturbances, like this:



In this representation, path *a* reflects the effect of T on M, path *b* reflects the effect of M on O, and path *c* reflects the effects of all excluded mediators, measured or unmeasured, from the analysis. If I have three measured mediators in my RET, then I conduct three separate analyses with each analysis using a different mediator in the M box. For a given

single mediator analysis, we can estimate the indirect effect of the mediator on the outcome (by computing path *a* times path *b* or some variant thereof) and we can estimate the total effect of T on O by summing the product of (*a*)(*b*) with *c* (or some variant thereof). The exact way we do so depends on the metric of M and O and the presence or absence of treatment by mediator interactions on the outcome.

It is well known, however, that the causal mediation framework runs into non-trivial analytic difficulties when more than two mediators are simultaneously brought into a model or when there are mediator-mediator interactions or causal relationships among the mediators. This is because the number of unique treatment effect decompositions into direct and indirect effects grows at an extremely large rate as a function of the number of mediators, especially when the mediators are causally ordered (see Daniel et al., 2015). The analyses quickly become unworkable unless one uses restrictive models with strong assumptions, thereby limiting the practical utility of the approach. The causal mediation framework also has difficulties with certain forms of latent variable modeling. Methodologists are working on overcoming these challenges and hopefully, practical solutions to them will be forthcoming. However, much work still needs to be done.

STRUCTURAL EQUATION MODELING

Structural equation modeling is yet another approach to the analysis of RETs. It is closely related to SCM but not synonymous with it. SEM predates SCM. Like SCM, SEM relies heavily on influence diagrams. The causal model represented by an SEM influence diagram makes predictions about how RET data should pattern themselves. SEM seeks to test those predictions while also providing feedback about how an intervention can be revised for purposes of improvement, i.e., it provides perspectives on (1) what is the effect of the intervention on the outcome, (2) what is the effect of the intervention on the mediators the program is intended to affect, and (3) are the mediators that an intervention targets meaningfully related to the outcome. SEM elegantly addresses models with multiple mediators, causal relationships among mediators, correlated disturbances, measurement error, longitudinal dynamics, interaction effects among mediators, interaction effects between the treatment and mediators, and it can handle both linear and non-linear relationships for variables measured with diverse metrics (e.g., ordinally scaled or binary variables), all while allowing for confound control. Given that the bulk of this book is devoted to SEM analyses of RETs, I let the remaining chapters speak to its utility and flexibility for RET analysis.

Some researchers who embrace the causal mediation framework mischaracterize SEM by stating SEM cannot do things that SEM readily accommodates. For example,

some claim that SEM cannot deal with non-linearity nor treatment by mediator interactions, which is not the case. Other researchers claim that SEM does not address or cannot address confounds in mediation and moderation analysis, which also is incorrect. The latter claim is often followed by a discussion of sequential ignorability in causal mediation contexts, with the implication that SEM does not concern itself with assumptions of sequential ignorability. This is false. To be sure, SEM, like all of the major mediation analytic frameworks, assumes sequential ignorability which raises statistical challenges for it. I address these challenges in future chapters. One certainly can find applications in the literature where SEM is poorly applied to RCT or RET data without sufficient concern for covariate control. However, misguided applications occur for most complex statistical methods and should not be the basis for making claims about the utility of a statistical method per se.

Another questionable criticism directed at SEM is that it does not embrace counterfactual conceptions of causality. Although counterfactual thinking is not central to SEM, such logic can be incorporated into it. Let me illustrate this by using the numerical example from Chapter 12 in conjunction with profile analysis. For this RET, I evaluated an intervention to increase parental communication with their middle school children about issues surrounding sex and birth control. The intervention was designed to increase such communication and targets three mediators, parental perceived advantages of engaging in such talks (PA), parental perceptions about how knowledgeable they are about sex and birth control (PK), and parental perceptions of how embarrassing it would be to engage in such talks (PE). Each mediator was measured on a -3 to +3 metric for a multi-item scale anchored by the average of item responses on a strongly disagree (-3) to strongly agree (+3) metric. Post-intervention communication with the child (COM) is binary (1 = talked with child, 0 = did not talk with child) as is the treatment condition (TREAT; 0 = parent participates in the control condition, 1 = parent participates in the intervention condition).

In SEM, one equation I work with is the logistic (or probit) equation that regresses the outcome onto the three mediators and the treatment condition.

$$\text{log odds}(Y) = a + b_1\,PA + b_2\,PK + b_3\,PE + b_4\,TREAT \qquad [9.6]$$

Suppose I want to calculate the total natural indirect effect for perceived advantages using counterfactual logic but now in an SEM context. Recall that the general equation for the TNIE (Equation 9.4) is

$$TNIE = (\bar{Y}_{TREAT}|M=m_{TREAT}) - (\bar{Y}_{TREAT}|M=m_{CTRL})$$

To calculate TNIE for perceived advantages (PA), I first need to calculate the probability

of communication for individuals in the intervention group when PA equals its mean in the intervention condition, the first term in the above Equation. Then, I calculate the probability of communication for individuals in the treatment group when PA equals the mean PA in the control group. The mean PA in the treatment group was 0.88 and it was 0.07 in the control group. I estimate these terms of the TNIE using profile analysis with the SEM-based Equation 9.6. To apply profile analysis using the equation, I must specify not only the values of m_{TREAT} and m_{CTRL} to use but also the values of PK and PE that I want to hold constant. I have flexibility to choose whatever values I think are substantively appropriate. Suppose I decide to use the mean values for m_{TREAT} and m_{CTRL} per traditional causal mediation logic and the respective mean scores in the intervention condition for PK and PE, which is consistent with the general logic of a TNIE. This yields the following two profile equations for the TNIE in the SEM framework:

log odds (Y) for term 1 of TNIE = $a + b_1\ 0.88 + b_2\ 0.85 + b_3\ -0.06 + b_4\ 1$

log odds (Y) for term 2 of TNIE = $a + b_1\ 0.07 + b_2\ 0.85 + b_3\ -0.06 + b_4\ 1$

I then convert these two log odds values to predicted probabilities, yielding the probability of communication for the first term of TNIE = 0.57 and for the second term, 0.45. The probability difference, or the TNIE, is thus 0.12. This is an estimate of the effect of the intervention on the proportion of parents who communicate about sex and birth control with their middle school child *through the mediational chain of perceived advantages* holding constant PK and PE at their intervention mean values. Note that these calculations use the same-world and cross-world counterfactual terms dictated by the causal mediation framework, showing that counterfactuals can indeed be incorporated into SEM through profile analyses. When I repeat the analysis using values for the PNIE instead of the TNIE, I obtained a result similar to that of the TNIE; the PNIE also was 0.12.[4]

I discuss in future chapters ways that counterfactual thinking can be brought into SEM. Here, I merely wish to question the assertion that counterfactual logic cannot be incorporated into SEM frameworks.

ADDITIONAL APPROACHES TO MEDIATION ANALYSIS

There are numerous other approaches that have been suggested for mediation analysis. As examples, Imai and colleagues (Imai, Keele & Tingley, 2010; Imai, Keele &

[4] Some implementations of causal mediation analysis use conditional regression as an estimation method but others do not. I discuss the different strategies in Chapter XX.

Yamamoto, 2010; Imai, Keele, Tingley & Yamamoto, 2010) propose different parametric and semi-parametric methods for evaluating mediation that use counterfactual concepts. VanderWeele (2009) suggests a method based on propensity-score weights in weighted regression. Linden and Yarnold (2017) describe methods based on classification tree analysis. Wodtke and Zhou (2020) articulate a regression-with-residuals approach. Vig et al. (2021) present a mediation framework based on neural network modeling. Biesanz, Falk and Savalei (2010) suggest a method in the spirit of the joint significance test that relies on partial posterior p values. Saunders and Blume (2018) describe a regression framework using a single regression model. Tofighi and Kelley (2020) develop a method they call the model-based constrained optimization (MCBO) procedure. Loh, Moerkerke, Loeys and Vansteelandt (2022) describe a form of mediation analysis called **interventional indirect effects** (see also Didelez et al., 2006; Hayes, 2018; VanderWeele et al., 2014; Vansteelandt & Daniel 2017). The approach focuses on multiple mediator models and seeks to estimate the omnibus mediation effect for a given mediator without regard to the causal structure among the multiple mediators. I describe the approach in Chapter 11. In the final analysis, mediation modeling is an evolving literature that is both diverse and complex.

CORE ASSUMPTIONS OF MEDIATION ANALYSIS

There exist critiques that question core assumptions of mediation analytics. I seek to put these critiques in perspective in this section. The assumptions made when modeling mediation vary depending on (a) the particular statistical approach used, (b) the nature of the model being evaluated, and (c) the questions being asked. Central to almost all mediation approaches is the assumption of **sequential ignorability**, which I introduced in Chapter 1. Definitions vary but sequential ignorability essentially refers to the idea that net the formal control of measured confounders, there is no meaningful unmeasured confounding of the treatment-mediator, treatment-outcome, and/or mediator-outcome relationships. This assumption typically is needed in order to obtain unbiased estimates of the causal coefficients in mediation modeling.

The presence of unmeasured confounds when estimating coefficients that document the strength of a causal link is problematic not just for mediation analyses but for many popular statistical methods in the social sciences (Montgomery et al., 2018). For example, in Chapters 2 and 6, I discuss the problem of omitted variable bias (also known as left out variable error, or LOVE) as a source of biased estimation in traditional OLS regression modeling. In RETs, the use of random assignment to treatment conditions is often thought to eliminate unmeasured confounds between the treatment condition and the

mediator (T→M) and between the treatment condition and the outcome (T→Y). Unfortunately, this is not always the case because of the occurrence of treatment dropouts, treatment non-compliance, treatment contamination, and missing data (see Chapters 26 and 27). In addition, some sample imbalance between conditions can occur even with properly implemented randomization (see Chapter 4). Despite these facts, in mediation modeling in RETs, unmeasured confounds typically are least likely to occur for the T→M and T→Y links and most likely to occur for the M→Y link and in hypothesized causal links between mediators for the case of multiple mediator models.

Unmeasured confounds are a fact of life in much of the social and health sciences. They almost always operate in mediation modeling. The question is not so much whether they are present but rather whether the degree to which they are present meaningfully misleads us in term so of the conclusions we make. Unmeasured confounds can cause errors but some errors are tolerable and inconsequential. If I want a sense of how long it takes to drive from New York City to Boston and inquire about the distance between the two cities, if what I am told is off by a few feet or inches, the error will not matter practically. If I report male versus female differences in annual income for high level executives in the United States and my estimate of the sex difference is off by, say, a few dollars, this is not going to be of consequence. By the same token, the strength of confounding by an unmeasured confounder might be strong, it might be moderate, it might be weak, or it might be negligible. If confounding is so weak that it produces trivial error in our estimates, then we can effectively ignore it because it is *functionally* zero.

As discussed in Chapter 2, Clarke (2005, 2009; Clarke et al., 2018) objects to traditional textbook discussions of omitted variable bias and unmeasured confounds because he claims the discussions oversimplify omitted variable dynamics. In reality, Clarke argues, there typically are so many omitted variables or unmeasured confounders that it is almost impossible to know how they bias coefficients when considered multivariately. Some plausible unmeasured confounders may induce positive bias while others induce negative bias. Their effects ultimately may cancel each other or, alternatively, the unmeasured confounders may magnify each other synergistically. We simply do not know. The likelihood that unmeasured confounders undermine conclusions depends on the nature and complexity of our models and the amount of error that we can tolerate, among other things. In my view, it is irresponsible to object to mediation modeling by waving the general wand of "unmeasured confounders" without making a thoughtful case for such dismissal by advancing a compelling narrative about which unmeasured confounds are at work and building a case for their biasing effects. We, of course, should always acknowledge the risks of confounds in our modeling efforts and then adopt an attitude of risk mitigation and risk management of those confounds when

designing, executing, and analyzing our RETs. This is good modeling practice in general. However blindly dismissing mediation analysis on the general grounds of unmeasured confounders is naïve at worst and somewhat simplistic at best.

A strategy I discuss in Chapter 2 for dealing with sequential ignorability is to identify during the design phase of an RET the most likely and important potential confounds, measure them, and then control for them analytically, as appropriate. In other words, convert the most critical unmeasured confounds to measured confounds, leaving only the weak or very weak unmeasured confounders to do their dirty work, perhaps in offsetting or trivial ways. We may not be able to completely remove bias, but perhaps we can reduce it to the point that its effects are negligible. In essence, we seek to achieve reasonable approximations to sequential ignorability so that we can make viable inferences about mediational links. I discuss strategies for identifying and prioritizing plausible confounders in Chapter 2 (see also Mändli & Rönkkö, 2023, and Vanderweele, 2019).

Another strategy for addressing the problem of unmeasured confounders is to use instrumental variables in conjunction with correlated disturbances to take their effects into account (see Chapter 6 and the document titled *Dealing with Correlated Disturbances* in Chapter 11 on the Resources tab of my web page). A third strategy is to use specialized experimental designs that render the unmeasured confounds moot (see Chapter 29). One must be cautious with the latter approach because these methods often make their own set of assumptions that may be problematic.

A common adage one hears in social science research is “do not control for posttreatment variables.” The idea is that if a measured variable refers to a person’s status after the intervention, the variable should not be included as a covariate or control variable in regression or SEM analyses that link the treatment condition to the outcome assuming one’s goal is to estimate the overall effect of the intervention on the treatment. This recommendation usually derives from the observation that the posttreatment variable might constitute a mechanism through which the intervention impacts the outcome. By controlling for it, you will underestimate the intervention effect on the outcome by not allowing a mechanism or “active ingredient” through which it operates to be “active.” If the magnitude of the overall intervention effect on the outcome is the question you seek to answer, then posttreatment controls must be invoked judiciously so as not to negate meaningful active ingredients of the intervention.

Having said that, if one blindly adopts the adage “never control for posttreatment variables” (which is sometimes asserted as such in the research literature) then mediation modeling as commonly practiced grinds to a halt because mediators typically are measured posttreatment and are simultaneously included in regression-like models in

order to assess the strengths of all or parts of mediational chains. Sometimes we seek to answer substantively important questions that require control for measured mediators at the posttreatment. For example, if the estimand of interest is whether the effect of a mediator on an outcome holds independent of other mediators, then controlling for those other mediators may be necessary in a regression or SEM model, albeit in a statistically principled way.

One, of course, still must be careful about one's choice of covariates because it is possible for subtle biases to enter into some forms of mediational modeling through posttreatment controls. One example is collider bias in complex multivariate mediational chains. I discuss collider dynamics in Chapter 2 and do not repeat that discussion here. I delve into collider bias in mediational modeling in the document called *Collider Bias and Mediation Analysis* on the Resources tab of my website for Chapter 2 and refer you to that document for elaboration.

As I discuss in Chapter 10, traditional mediation analysis often seeks to decompose a total effect into (a) indirect effects through *a priori* specified mediators and (b) direct effects of the treatment condition on the outcome independent of those mediators. It is in the context of such foci that unwanted collider bias from unmeasured confounds can arise. By contrast, my approach to RET analysis for purposes of program evaluation is different than decomposing omnibus total effects into direct and indirect effects. In my framework, I examine each mediational chain on a link by link basis to determine the strength of each link and to isolate where in a mediational chain a link might be "broken" or be so weak that it needs to be addressed by program designers. The direct effect of T→Y independent of the measured mediators is more of an incidental parameter because it only tells me that the intervention may be affecting the outcome through other unspecified and unmeasured mechanisms that I am left to speculate about. The real "meat" of program evaluation is the analysis of the separate links in the different measured mediational chains and the estimated total effect of the intervention on the outcome per se. Because I emphasize somewhat different questions for program evaluation than traditional mediation modeling, the modeling challenges I encounter do not overlap completely with those of traditional mediation modeling. Again, the challenges you face depend on the questions you seek to answer and the modeling you use to answer those questions.

Like most statistical methods, both traditional regression and SEM based mediation analysis make assumptions that can and should affect the confidence we have in our conclusions. As Montgomery et al. (2018) nicely put it, unpacking the "black box" of interventions through mediation modeling typically must be paid for in the form of assumptions and potential bias in our estimates (more or less). Montgomery et al. (2018)

emphasize that researchers should make explicit their assumptions in their research reports, approach their conclusions humbly in light of those assumptions, and, where possible, report margins of error and perform sensitivity tests to determine the consequences and boundaries of assumption violation.

At its heart, SEM guided mediation analysis is designed to provide perspectives on the viability of causal models of mediational links. SEM as applied to program evaluation (or most any other substantive area for that matter) *cannot* prove the existence of causal links between variables in the social and health sciences. Rather, causal inference evolves from a logic model that goes something like this: Based on a detailed and careful analysis from the perspective of multiple constituencies of a program or intervention, we formulate what we think is a viable causal model about why a program affects the target outcome of interest. This model invariably is based not only on a detailed logical analysis of the program, but also past theory and scientific knowledge, past research, and viable hunches about the underlying causal dynamics. The causal model typically includes both meaningful mediation links as well as confounds for those links that need to be controlled lest we be led astray when evaluating the model. Based on this causal model, we then collect data to gain perspectives on the formulated model. Importantly, the model makes predictions about how the data should pattern themselves. In full information estimation SEM, this often takes the form of predictions about the variances and covariances of the variables in the model and the statistical significance of model path coefficients. In limited information estimation, it often takes the form of conditional independencies and the statistical significance of path coefficients. An interesting facet of SEM is that the model we posit sometimes includes a measurement theory in addition to a structural theory so that we can take into account measurement error when evaluating model predictions. These measurement submodels also make predictions about how data should pattern themselves, as you will see in Chapter 11.

With the data in hand, we then compare how the data are predicted to pattern themselves by the model with how the data actually pattern themselves. If there is close correspondence between the predicted and observed data patterns, then this increases our confidence in the posited model. If there is not close correspondence, then we question the model and perhaps reject it. In this sense, the SEM approach allows us to gain perspectives on the viability of causal models but it does not *prove* unambiguously that causal relationships exist. We feel more confident in our hypothesized model when the data pattern themselves in accord with model predictions and feel more confident that the causal dynamics within the model may be operating. However, even then, we also must recognize that there may be one or more alternative models that account for the data equally well. Sometimes we can competitively test the two models against one another

with the data at hand but other times, we need to collect data in a newly designed study that will allow us to choose between the competing models. If we cannot do so because of lack of resources or practical constraints, we accept the limitations of the collected data and adjust our confidence in the tested model accordingly.

If we have what we think is a viable model that is data consistent, then we often take data analysis a step further and interpret the values of the parameter estimates of the model, i.e., the estimated path coefficients, variable correlations, disturbance terms, and the parameters of the measurement model. Strictly speaking, the path coefficients are not causal coefficients. They are *estimates* of the causal coefficients tied to the assumption that the model in which they are embedded is a reasonable approximation to the true underlying causal dynamics at play (if there is such a thing). I carefully reflect on the signs and magnitudes of the various coefficients to ensure they make conceptual sense and then consider the practical implications of them for program design and implementation. If I see a coefficient value that makes no sense and contradicts logic and past research, it may lead me to re-think the model.

The process of specifying a causal model, deriving predictions about how the collected data should pattern themselves, and then comparing model predictions with the collected data is fundamental. The process characterizes what we do in social science research for both experimental and non-experimental research, cross-sectional and longitudinal research, and for multi-level research more generally. When we conduct an experiment to test the efficacy of a vaccine, for example, we formulate a theory or model about the biological mechanisms the vaccine should affect and the ultimate outcome that should occur, such as disease prevention. The model makes predictions about how the data we collect in an efficacy trial that focuses on these variables should pattern themselves. All we can then do is determine if the data do, in fact, pattern themselves that way. If the predictions are close to the observed data patterns, then we have increased confidence in the conclusions we make about the efficacy of the vaccine. However, we recognize that there may be limitations to our study design that do not permit us to unambiguously assert our conclusions are accurate. There almost always will be some error involved. Given a good fitting model, we then interpret the parameter estimates of the model to make statements about how efficacious the vaccine is.

In the above sense, our ability to evaluate causal models and the parameter estimates they yield is not only a function of how we analyze the data but also how well we design a study to take into account the noise and confounds that can mislead us. Study *design* is crucial. Some scientists believe that some study designs are inherently better than other study designs for evaluating causal models, but their choice of examples is not always on target. For example, some believe that longitudinal designs are better than

cross sectional designs for evaluating causal dynamics. But consider the case of evaluating whether changes in A impact B where the length of time it takes for changes in A to translate into changes in B is extremely short, say a matter of seconds. Would we have better sensitivity for detecting this causal dynamic using a cross sectional design than a longitudinal design that, say, measures A and B several weeks or months apart? Granted there are challenges that would arise in the cross sectional study but perhaps addressing the challenges of making causal inferences in cross sectional data that more closely approximates the operative time interval between cause and effect would be better than using a blatantly wrong longitudinal design that seriously misspecifies the relevant causal lags. Instantaneous change or near instantaneous change often is impossible to study "longitudinally." I discuss the issue in more depth in Chapter 16.

In the final analysis, study "noise," confounds, measurement error, and model specification error are facts of life in most social science research. This is not a reason to abandon such research just as we do not want to abandon efforts to evaluate programs or interventions because of them. Rather, we should seek to articulate the operative sources of bias involved in a study, manage them as best we can when designing our study, and then keep them salient when we draw conclusions and make decisions about how to proceed. Such is the nature of mediation modeling.

WRITING REPORTS OF MEDIATION ANALYSES

In the scientific literature, several expert panels have made recommendations about the core facets to include when writing reports about mediation. On the Resource page of my website, I provide a link to an influential consensus panel's recommendations for reporting mediational analyses in RETs (Lee et l., 2022). In the interest of space, I do not repeat their recommendations here but they are comprehensive. They also are somewhat unrealistic given the page and word limits that operate in most journals. Fortunately, most reputable journals have on-line supplement webpages where authors can provide more information about their study, designs and modeling efforts. If you make use of these supplemental pages, you usually will be able to address the recommendations of expert panels.

Writing for journals is one thing. Writing for administrators, executives, and program staff is quite another. Such reports often begin with a short "Executive Summary" that highlights the most important findings of the evaluation in non-technical and practical terms. Graphics and visual aids might be used liberally. Jaccard and Jacoby (2020) provide useful tips to keep in mind when writing reports and structuring presentations. I do not shy away from providing technical information that is written for

other scientists in my final reports, but much of it I put in appendices. Unfortunately, the contexts and audiences one encounters for program evaluations are so varied that it is difficult to provide specific guidelines other than to keep it simple, keep it practical, and make it interesting. I use a heuristic of three words: communicate, motivate, and facilitate (CMF). I seek to **communicate** what I found, I seek to **motivate** evaluators to act on the results I present, and I seek to **facilitate** ways they can act on the results by making implementation recommendations. I design my evaluation strategies from the outset with all three of these tasks in mind, conducting interviews with staff to help me strategize each of them.

CONCLUDING COMMENTS

In sum, a variety of methods have been suggested for analyzing mediation in RETs, including the Baron and Kenny method, the coefficient product method, the joint significance test, the Hayes PROCESS approach, the MacArthur network model, SCM and its associated causal mediation analysis, and SEM, among others. Each method has strengths and weaknesses and I address many of them as I develop the fundamentals of SEM-based analyses of RET data throughout this book. I define the SEM umbrella more generally than many researchers and, as such, I see it as encompassing a broad range of analytic tools that all can be brought to bear when evaluating programs using RETs. I do not hesitate to augment traditional SEM with analytic methods that enhance its flexibility and applicability. However, SEM is the core framework I use to organize my thinking about analyzing RET data.

Of the many methods for testing the null hypothesis of mediation versus no mediation for a given mediator, I lean to use of the joint significance test, but I recognize its shortcomings. As I have noted, my primary focus is usually on the evaluation of the strength and statistical significance of each individual link in a mediational chain using results from FISEM and often one or more LISEM methods for sensitivity purposes. To document effect size for a given link in a mediational chain, I use one of the effect size methods discussed in Chapter 10. I personally do not find the omnibus test of a given mediator to be all that informative for program evaluation purposes because it confounds the effect of the treatment on the mediator with the effect of the mediator on the outcome. Nevertheless, some researchers like to work with it. They argue that mediation is fundamentally multivariate in character so that the test of mediation should be multivariate in ways that reflect that character. It is like arguing in ANOVA, the omnibus F test is the proper way to evaluate the null hypothesis of no group mean differences. Some methodologists argue for this perspective while other methodologists instead argue

that (a) the null hypothesis of no group differences is not of much interest in its own right, and (b) knowing what is happening at a more fine grained level in terms of specific group differences for the groups being evaluated provides more useful information as long as it is handled properly statistically. I return to this issue in more depth in later chapters.