2

Conceptual Fundamentals for RETs

Sometimes you can see a lot just by looking

- YOGI BERRA (NY Yankees catcher)

INTRODUCTION

Behavioral interventions range from the simple to the complex and can be diverse, to say the least. Examples include the use of monetary incentives to increase adherence to medical regimens, the introduction of school lunch programs to improve school performance, parent training programs to improve parenting, job skills training to improve employment, the use of motivational interviewing to reduce relapse among alcoholics, diet-based programs to reduce obesity, and the use of humor in advertising to impact product purchases. The spirit of an RET is to specify the mechanisms through which such interventions have their effects and to consider the boundary conditions of those effects. In this chapter, I consider core conceptual issues that should be considered when designing an RET. I begin by providing some reflections on the concept of causality and the nature of it. I then introduce an important tool for conceptual analyses, namely influence diagrams. Influence diagrams are used to specify conceptual logic models for RETs and they are used to make crucial decisions about covariate control.

After introducing this important tool, I discuss three conceptual tasks that RET designers face (1) the task of mapping an intervention onto relevant mediators and moderators, (2) the task of conducting a broader analysis of outcomes, mediators, and moderators for purposes of program evaluation, and (3) the task of defining what constitutes a meaningful effect of a program on mediators and outcomes as well as the effects of mediators on outcomes.

THE NATURE OF CAUSALITY

Causal thinking dominates much of the social and health sciences. Essential to causal thinking is the idea that one of the variables in a causal relationship, X, influences the other variable in the relationship, Y. The nature of causality has been debated extensively by philosophers of science (e.g., Bunge, 1961; Cartwright, 2007; Frank, 1961; Morgan & Winship, 2007; Pearl, 2009; Pearl & Mackenzie, 2018; Rubin, 1974, 1978; Russell, 1931; Shadish, Cook, & Campbell, 2002), most of whom agree that causality is an elusive concept that is fraught with ambiguities. The famous philosopher Bertrand Russell (1931) was so flabbergasted by difficulties with the concept that he suggested the word causality be expunged from the English language.

Scientists generally think of causality in terms of change. Variable X is said to be a cause of Y if changes made to one or more of the crucial properties of X *produce* changes in Y. Hume (1777/1975) argued that it is impossible to ever demonstrate that changes in one variable produce changes in another. At best, we can only observe changes in one variable followed at a later time by changes in another variable. Such coexistent change, he notes, does not necessarily imply causality. For example, an alarm clock going off every morning just before sunrise cannot be said to be the cause of the sun rising, even though the two events are intimately linked. Some philosophers note the tendency to anthropomorphize causality by imputing human qualities to it (Buzzoni, 2014). Pearl et al. (2016) characterize it as follows: *For our purposes, the definition of causation is simple, if a little metaphorical. A variable X is a cause of a variable Y if Y in any way relies on X for its value… X is a cause of Y if Y listens to X and decides its value in response to what it hears*.

Russell (1931) argued that causality can be established unambiguously only in a completely isolated system. If one assumes no other variables are present or operating, then changes in X that are followed by changes in Y are indeed indicative of a causal relationship. When contaminating variables are present, however, it is possible for a true causal relationship to exist, even though observations show that X and Y are completely unrelated to each other. Similarly, a causal relationship may not exist, even though X and

Y are found to be related. Having shown this using formal logic, Russell turned to the problem of how one could ever know that one is operating in a completely isolated system to demonstrate causality, such as in a highly controlled laboratory setting. The only way to be confident that the system is isolated, he argued, is if changes in X unambiguously produce changes in Y in that system. But at the same time that we want to assert the existence of an isolated system because changes in X produce changes in Y, we also want to assert that X produces a change in Y because we are operating in an isolated system. Such reasoning, Russell argued, is tautological.

As you might imagine, the issues for conceptualizing causality and how one establishes causal relationships are complex. They have been debated by very bright philosophers of science and scientists for decades (Halpern, 2015), and I certainly am not going to resolve the matter here. After reading the relevant literature carefully and giving the matter much thought, I think that, in a strict sense, causality of the type that traditional social scientists seek to infer for real world social problems is difficult to demonstrate unequivocally. Strong adherents to experimental methods take exception to this view, and I respect that. However, I personally find that the arguments of Blalock (1964), Bunge (1961), Hume (1777/1975), Russell (1931), and a host of others, taken as a whole, raise reasonable doubts that causality as pursued in the social sciences can be unambiguously demonstrated.

If causality is so difficult to demonstrate, then why is the concept dominant in social scientific theories? One answer is that the concept of causality is a type of mental model that social scientists use to help them think about our environment, organize our thoughts, predict future events, and even change future events. By thinking in causal terms, we are able to identify relationships between variables and often manipulate those variables so as to produce changes in phenomena that are socially desirable to change. Causal thinking has been used to invent lasers and transistors, to fly to the moon, and it has resulted in all kinds of remarkable human inventions. Pearl (2009) argues that "deep understanding means knowing not merely how things behaved yesterday but also how things will behave under new hypothetical circumstances, *control being one such circumstance*." (p. 415). Causal frameworks can provide such understanding.

Although we may rarely be able to unambiguously demonstrate causality between variables central to the social sciences, we certainly can have differing degrees of confidence that a causal relationship (of the form that "changes in X produce changes in Y") exists between variables. Scientific research, in my view, is conducted to establish strong, moderate, or weak levels of confidence in theoretical statements that propose causality. In particle physics, the classic five-sigma standard defines certainty as a 99.9999% chance of something being true, such as whether humans have caused climate

change. While this may rarely be attainable with social science theories, it does underscore the role that the concept of confidence plays in scientific inference.

Mental models that make use of causal statements bring with them a set of assumptions about causality or, stated another way, a theory of causality. In that sense, causal statements are always assumption laden. There are some features of causality on which most social scientists seem to agree. First, as noted, if X causes Y, then changes in X are thought to produce changes in Y (but see Sowa, 2000, and Lewis, 2000, for alternative conceptualizations). Second, a cause always must precede an effect in time. Third, the time that it takes for a change in X to produce a change in Y can vary, ranging from almost instantaneous change to weeks, months, years, decades, or centuries. Fourth, the nature and/or strength of the effect of X on Y can vary depending on context. X may influence Y in one context but not another context. Finally, cause and effect must be in some form of spatial contact or must be connected by a chain of intermediate events. I return to each of these points throughout this book.

An increasingly popular view of causality in the social sciences uses a **counterfactual framework** that grew out of the work of Lewis (2000). To illustrate the basic idea, when analyzing the causal effect of a treatment (X) on an outcome (Y), the counterfactual of interest is comparing the potential outcome that would occur if a person receives the treatment versus the potential outcome that would occur if that same person did not receive the treatment under the exact same circumstances as when the treatment was administered. If the potential outcomes are different, causality is implied. Based on this fundamental counterfactual premise, scientists and philosophers have posited a "theory of causality" as well as scientific prescriptions for establishing causality (see Menzies, 2017; Pearl, 2009; Pearl & Mackenzie, 2018). The approach, or variants of it, is also known as a **potential outcomes** theory of causality (Holland, 1986; Morgan & Winship, 2007). I discuss counterfactual theories in more depth in future chapters. However, you will often see it invoked to assert causality in social science research.

Not all scientific theories rely on the concept of causality. In fact, certain areas of physics did not progress until the notion of causality was deemphasized (see Sowa, 2000). Jaccard and Jacoby (2020) describe multiple frameworks or mental models for thinking about social science phenomena and encourage scientists to think about phenomena from the perspective of different mental models only one of which invokes causality. Having said that, causality remains a dominant system of thought in the social sciences and this book adopts the perspective of causal thinking, with all its strengths and weaknesses. However, I certainly acknowledge and respect the use of other thought systems. I will adopt a working definition of causality that emphasizes the five common features of it described above.

INFLUENCE DIAGRAMS AS A CONCEPTUAL TOOL FOR RETs

A key task when designing and conducting an RET is to specify a conceptual logic model that elucidates the operative mediation and moderation dynamics, per Chapter 1. There are many ways of expressing a conceptual logic model, but I strongly suggest you do so using an influence diagram. An **influence diagram** is a graphical depiction of the causal relationships between variables in a causal system. As you will see, influence diagrams force you to make your assumptions about variable relationships explicit and they help you think through those assumptions with clarity. They also are key to writing computer code for the statistical analysis of RET data. Many social scientists associate influence diagrams with the analytic method of structural equation modeling, but they have much wider applicability than this. They help us organize and think through our theories.

In the broader literature on causal analysis, there are two general approaches to constructing influence diagrams, an approach based on conventions introduced by the "father" of path analysis, Sewall Wright (1921, 1934), and an approach grounded in mathematical graph theory as described by Judea Pearl (Pearl, 2009; Pearl, Glymour & Jewell, 2016). In this section, I briefly review both approaches. Although I emphasize the use of influence diagrams for the construction of conceptual logic models, I also briefly consider their use for presenting statistical results, since this is a common practice.

Traditional Influence Diagrams

The key elements of a traditional influence diagram can be illustrated using a simple bivariate regression analysis where a causal relationship is assumed between the predictor and the criterion. Suppose I regress for a population of middle-aged adults their annual income onto the number of years of education they have completed. The results might be:

Annual Income = $10,000 + $2,000 Education

where the intercept is $10,000, the regression coefficient is $2,000 and the squared multiple correlation is 0.40. The intercept is the predicted mean annual income when the number of years of education equals zero. The regression coefficient indicates that for every one-year education increases, the mean annual income increases by $2,000. The squared multiple correlation is the proportion of variance of annual income that the number of years of education "explains" or accounts for. One minus this value, 1 – 0.40 = 0.60, is the proportion of "unexplained" variance due to factors other than education. It is called the **disturbance variance**. Figure 2.1 shows an influence diagram of these results:

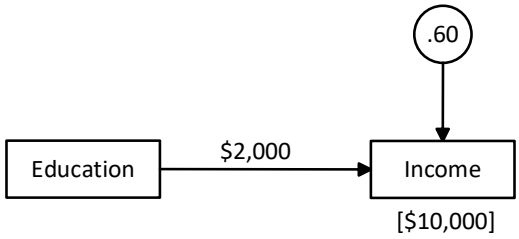


FIGURE 2.1. Influence diagram for education and income

Variables in the diagram are represented by rectangles. The straight arrow indicates a presumed causal relationship. The numerical path/regression coefficient appears above the straight arrow and the intercept is in brackets below the outcome. In influence diagrams, a variable that has a straight arrow leading into it is an **endogenous variable***;* a variable that does not have any arrow leading into it is an **exogenous variable**. Education is an exogenous variable and income is an endogenous variable. Every endogenous variable has a **disturbance term** (sometimes called an **error term**). The disturbance term corresponds to the error term in traditional regression analysis and, as noted in Chapter 1, its variance is usually of interest. Disturbance variances are typically reported in standardized form because they then reflect the proportion of unexplained variance in the outcome. In this form, they equal one minus the squared (multiple) correlation in the analysis. In influence diagrams, tradition is to represent unmeasured variables as circles. Disturbance terms are represented as circles because, technically, they are not directly measured. In some applications, intercepts are not reported because they are not of theoretical interest. When used to summarize theory rather than present results, numerical entries are omitted and the disturbance term is labeled using a label *d* or *e*. Sometimes the disturbance terms are omitted from a diagram to reduce clutter, but if this is done, their presence is implied for each endogenous variable in the system.

Figure 2.2 presents another version of an influence diagram you may encounter. In this version, the intercept is represented by a triangle with the number one inside of it (because, technically, its value derives from regressing the outcome onto a vector of 1s; sometimes the number 1 is omitted from the triangle). The value of the intercept is indicated on the arrow connecting the triangle to the outcome (sometimes a straight line without an arrowhead is used). The disturbance term is given a label and its variance is indicated by a double-headed, curved arrow directed at the disturbance term itself. This represents a more general diagramming convention: A double-headed curved arrow linking a variable to itself represents a variance.

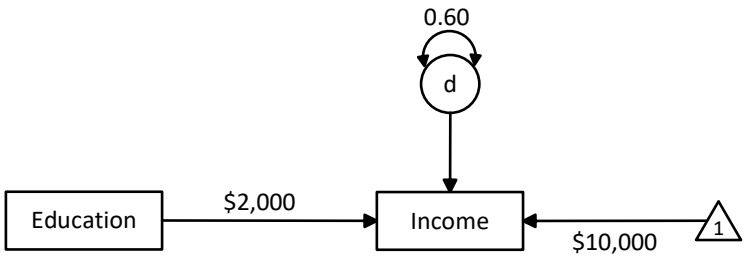


FIGURE 2.2. Alternative influence diagram for education and income

Figure 2.3 presents an influence diagram of a multiple regression model that has two predictors of annual income (a) the number of years of education and (b) biological sex assigned at birth, scored 0 = female and 1 = male. I omit the intercept in the diagram because it is not of theoretical interest. Because gender is a dummy variable, the path/regression coefficient for it is the mean outcome difference for the group scored 1 (males) minus the group scored 0 (females). In this case, males are paid, on average, $5,000 more than females holding the number of years of education constant. Each additional year of education is again worth $2,000, holding gender constant. The number of years of education and gender, taken together, account for 48% of the variation in annual income ((1-0.52)(100)), with factors outside the model accounting for 52% of the variation in annual income (0.52)(100). The double-headed curved arrow between the two exogenous variables, gender and education, represents the covariance between them. When the curved arrow is applied to a single variable, in which case it is a variance.

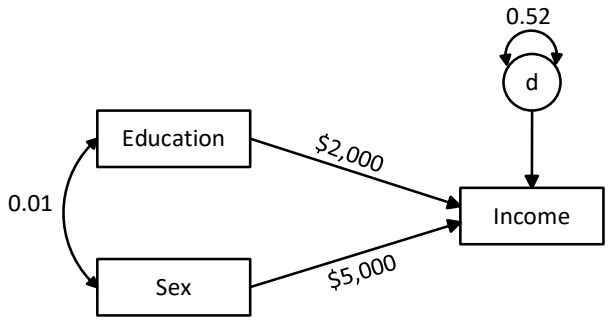


FIGURE 2.3. Multiple regression model

Sometimes path coefficients are starred with an asterisk to indicate that tests of their null hypothesis are statistically significant and sometimes their standard errors or margins of error are presented next to them. Sometimes the coefficients are reported in standardized form and sometimes in unstandardized form. In the current case, I use a mixture of the two formats, with the correlations and disturbance variances reported in standardized form and the path coefficients in unstandardized form. This would be indicated in a figure footnote. Sometimes the correlations between the exogenous variables are omitted even though they are modeled. They are omitted if they are not of theoretical interest to avoid clutter. In such cases, a footnote also indicates this is the case.

Figure 2.4 presents an influence diagram that illustrates the case of "pseudo mediation." There are two confounders, C1 and C2, both of which are common causes of variables M and Y. Note that there is no causal influence of M on Y, but M and Y will be correlated because they share common causes. M appears to be a mediator of Y because it is related to Y, but it is not a mediator because there is no causal link between them; the relationship is spurious. M is a **pseudo-mediator**.

This diagram illustrates another diagramming convention. Even though M and Y are correlated, I do not draw a double-headed curved arrow between them. This is because a correlation is implied given the common causes. As a general rule, correlations are not drawn between endogenous variables in influence diagrams because their correlational structure is defined by other elements of the model, in this case, by the confounders C1 and C2. However, the convention is to draw correlations between exogenous variables and between disturbance terms, as appropriate. That is why a correlation is indicated between the two exogenous variables, C1 and C2. Often, drawing correlations among exogenous variables introduces clutter into the diagram, in which case they are omitted from the diagram even though they are modeled. A footnote to the figure will usually indicate that this has been done.

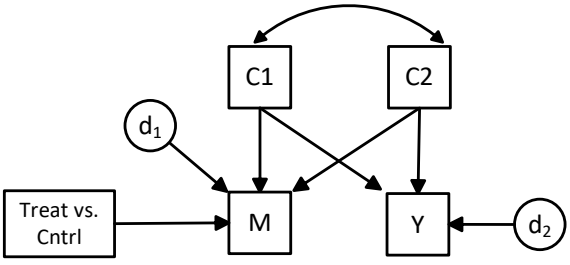


FIGURE 2.4. A confounder model

Graph Theory

Pearl, Glymour & Jewell (2016) outline a graph theory for use in causal modeling. The graphs are often called **directed acyclic graphs** (DAGs), but DAGs are only a subset of the broader graph theory Pearl et al. use. In this book, I use traditional influence diagrams instead of DAGs because they are more common in the social science literature. However, Pearl has made important contributions using graph theory, so familiarity with his approach is important for reading the broader literature on mediation and moderation.

Figure 2.5a presents a DAG for a single mediator model with one distal determinant (T, the treatment condition) and one outcome, Y. The variables are not in rectangles but are signified by letters (or names) with a dot, called a **node**, above each one. The causal arrows originate at the node and end with an arrowhead. What is called a path in a traditional influence diagram is called an **edge** and it is signified by a letter along the edge, per Figure 2.5b. These edges are said to be **directed** because the edge goes out of one node and into another node. For example, for edge A in Figure 2.5b, the edge goes out of T and into M. Disturbance terms for endogenous variables are implied rarely included in the graph. When they are included, they often are symbolized by a U.

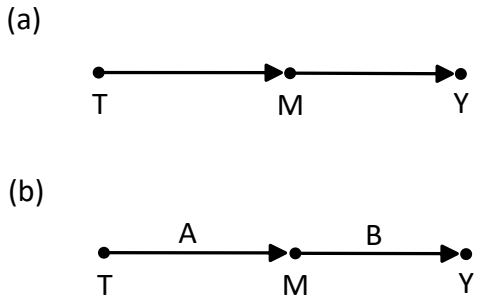


FIGURE 2.5. Graph theory diagrams

The variable that an arrow emanates from is said to be a **parent** of the variable that the arrow points to. The variable the arrow points to is said to be a **child** of the variable from which the arrow begins. In Figure 2.5a, T is a parent of M and M is a child of T. Additional terminology is that of ancestors and descendants. In Figure 2.5, M and Y are **descendants** of T (because they both derive from T), and T and M are **ancestors** of Y (because they impact Y).

Correlations between variables or disturbances are typically denoted by curved double-headed arrows, much like traditional influence diagrams, but the usual line is

dashed rather than solid (Pearl, 2010). These relationships are said to be **undirected** because they have no causal ordering. Figure 2.6 presents a graph with disturbance terms for the variables M1, M2, M3, and M4 (signified by U1, U2, U3, and U4). The disturbances for U1 and U2 are correlated. As with traditional influence diagrams, correlations between exogenous variables are sometimes excluded to avoid clutter.

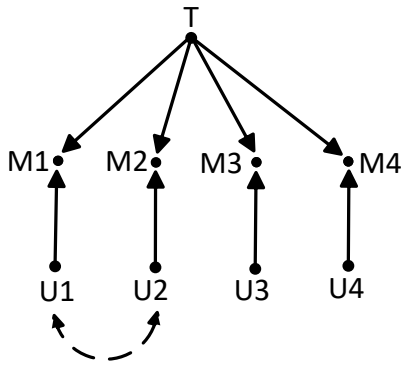


FIGURE 2.6. DAG with correlated error

One final terminology matter. The term **path** is used differently in a DAG than in traditional influence diagrams. In DAGs, a path is a sequence of nodes that have directed edges. For example, in Figure 2.5a, there is a path from T to Y consisting of the edge from T to M and the edge from M to Y. For more details about DAGs, see Pearl et al., (2016).

Influence Diagrams with Many Variables

Working with influence diagrams can become unwieldy as the number of mediators, moderators, and confounds increase. One strategy is to draw separate influence diagrams for different portions of the theory, a strategy I illustrate in Chapter 10. For example, one diagram might focus on outcomes, mediators and moderators that are central to the theory and another might focus on the nature of covariate influences. This allows me to draw attention to the primary theoretical narrative in the first diagram without the "messiness" of covariate controls (see Chapter 10). A second strategy is shown in Figure 2.7, again omitting covariates. In this case, each box contains a superordinate label naming a general category of variables with a listing of relevant variables underneath the category label. All of the variables within a category are assumed to have the same determinants and/or same consequences. Without this heuristic, the number of individual boxes and

arrows in the diagram would be overwhelming. A weakness of the strategy is that causal relationships between variables within a category are unarticulated (or presumed not to exist) and disturbance terms and their relationships also are omitted and unelaborated. This type of influence diagram often is used to represent a general conceptual framework rather than specific conceptual logic models for purposes of RET design.

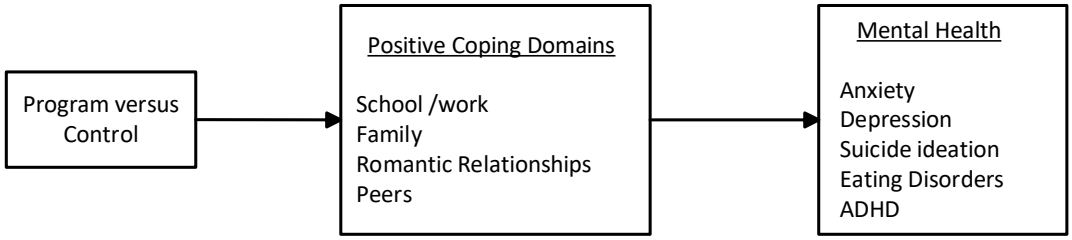


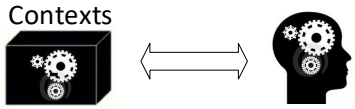
FIGURE 2.7. Multi-category influence diagram

In sum, I strongly urge you to develop conceptual logic models for the program you are evaluating using influence diagrams. Influence diagrams force you to make your assumptions about variable relationships explicit and they help you think through those assumptions with clarity. For every path that you draw between variables in the diagram, make sure the variables are clearly defined conceptually, that they strike the appropriate balance between abstractness and specificity (per my discussion in Chapter 1) and that you can articulate a compelling rationale for the existence of the path. For every path you do *not* draw, make sure you can justify why it should not be present. For every disturbance term, think about what variables those disturbance terms likely represent and justify why or why not you have included or omitted correlated disturbances between any given pair of disturbance terms. See Jaccard and Jacoby (2020) for elaboration on how to think through and construct influence diagrams at a conceptual level.

THINKING OF RETS AS OPENING THE BLACK B0X

Program evaluation and intervention analysis begins with a distal determinant (the treatment variable) on the left side of an influence diagram and a target outcome on the right side of the diagram. A central task of the RET designer is to identify the intermediary variables (mediators) through which the program impacts the outcome. Using the classic black box analogy, the task is to open the black box and discover what is in it:

If we open the box, we most certainly will find two large items that themselves require "opening," namely



These items reflect (a) the biological and mental processes that are operating in the heads of people who are the focus of the intervention and (b) the contexts these people interact with. Interventions seek to influence one, the other, or both of these elements and, of course, changes in one can influence changes in the other.

In this chapter, I consider three conceptual tasks essential to the design of RETs. First, is the task of carefully analyzing the intervention to define the determinants of the outcome it targets. This will lead to a preliminary conceptual logic model. Second, is the task of placing the targeted mediators into a broader theoretical context that specifies factors outside of the preliminary logic model that influence the RET-focused mediators and outcomes. The purpose of this task is to identify confounds and additional variables that need to be taken into account when designing the RET. Third, is the task of defining what constitutes a meaningful effect in an RET, either with reference to the effect of the program on a mediator, the effect of the program on the outcome, or the effect of the mediator on the outcome. I consider each of these tasks, in turn, and then conclude with a discussion of multiple outcome RETs.

CONCEPTUAL TASK 1: MAPPING THE INTERVENTION

The first task for designing an RET is to carefully analyze program activities to identify the causal mechanisms that each set of activities target. If the RET is to be conducted for an agency, this task can be pursued in concert with discussions with program administrators and staff. The identified targeted determinants take the role of mediators in the RET. Jaccard and Jacoby (2020) suggest that one can often turn a direct causal relationship between X and Y into a mediated relationship by asking the question "why does X influence Y?" The answer to this question contains variables that mediate the effects of X on Y. In the case of an RET, we first analytically divide the program into

discrete sets of activities and then we ask for each activity set "why does this set of activities (X) influence the outcome (Y)? What is it that this set of activities is changing that will bring about change in the outcome (Y)?" As we articulate answers to these questions, we identify possible mediators to include in the RET. For example, if individuals in a program watch a video on the health hazards of using drugs and I ask "why does this activity impact future drug use," the answer might be because it changes participants' beliefs about the health consequences of using drugs which, in turn, will affect their use of drugs; beliefs about the health consequences is the mechanism (or mediator) through which watching the video affects future drug use.

Qualitative Research and Mediator Mapping

For some types of programs, I have found mediator mapping to be facilitated by conducting qualitative research prior to the design of the RET. For example, in cognitive behavior therapy (CBT) for anxiety, one set of CBT activities addresses client negative thought patterns, another set targets coping strategies, a third set addresses stress reduction/relaxation, and a fourth set addresses emotion regulation. The targeted mediators are evident in each of these descriptions; the mediators are (a) negative thoughts, (b) coping, (c) stress/relaxation and (d) emotion regulation. Insights into the activity sets and how to frame and measure them in the RET might be gained by conducting in-depth interviews with a small sample of clients who have recently completed the CBT protocol. I might ask clients for each set of activities to paraphrase what the activity is, to state what the activity means to them, to indicate what they think the activity is meant to accomplish, to reflect on whether the activity is effective, and to elaborate why or why not. I might also ask how the activity can be modified to improve it. In my research to improve communication between parents and adolescents, I develop program activities to teach parents how to communicate more effectively with their adolescent children. I explore in qualitative research not only parent reactions to these activities, but I also solicit adolescent reactions to what I will be telling parents to do. Carefully crafted qualitative research prior to RET design often can be of use for mediator mapping.

Mapping Mediators When Program Activities are Vague

Often, mapping of activities onto presumed determinants of the outcome is straightforward, per the example in Chapter 1 where a drug prevention program targeted (a) peer resistance skills, (b) the perceived short-term negative consequences of using drugs, and (c) the perceived long-term negative consequences of using drugs. However, sometimes the mapping is less clear. For example, navigator programs that provide

cancer patients with a personal "navigator" to help patients deal with the hospital bureaucracy in order to improve patient satisfaction may be ambiguous about what exactly the navigators are to do. An analysis of the navigator training protocols might reveal the determinants that navigators are expected to address. However, it also could be that the training protocols are underdeveloped. In such cases, I might seek to specify outcome determinants derived from the broader scientific literature on patient satisfaction to help organize the RET. For example, studies suggest that overall patient satisfaction with medical services are impacted by perceptions of the quality of communication with the doctor, perceptions of the quality of communication with nurses, perceptions of the absence of logistical obstacles to obtaining services (such as obtaining test results, getting to and from the hospital), the perceived professionalism of office staff, and the perceived cleanliness of the office. I might use these documented determinants as a basis for choosing mediators/mechanisms the navigator program might affect and include them in the RET despite the fact that it is unclear what mediators the program has targeted through its use of navigators. After the RET, feedback can be given to program staff about what key mechanisms the program affects and what mechanisms it does not affect.

Mapping Mediators When Program Activities are Simple

Some programs use strategies that are so straightforward that the identification of mediators may seem unnecessary, thereby questioning the need for an RET in the first place. For example, an intervention that sends reminder text messages for medical appointments to increase appointment-keeping behavior might not seem to require an RET because one merely wants to know if the reminders work. However, the spirit of an RET is to make program assumptions explicit and to generate evidence-based suggestions for program improvement. The task of the program evaluator is to think about this simple "activity" more deeply and to identify mediators of its effect on the outcome, namely appoint keeping behavior. Why does sending a reminder increase appointment keeping behavior? How can we leverage the answer to this question to improve a reminder program via an RET?

The typical logic model for reminder-based programs is that reminders increase the salience of the behavior and that by calling attention to the behavior at time *t*, the behavior is more likely to occur at time *t+1*. Thus, behavioral salience is a hypothesized mediator of the effect of reminders on attendance behavior. We might next ask why does making the appointment salient prior to the appointment matter for appointment keeping behavior? One answer might be because it allows people to plan for or make necessary arrangements to keep the appointment (e.g., keeping one's calendar open, arranging for childcare). We also might presume that independent of such planning, making the

appointment salient, say, three days prior to the appointment, impacts the salience of the appointment on the day the appointment is scheduled. Figure 2.8 shows these dynamics.

Note that an RET that sheds light on the causal coefficients in Figure 2.8 would provide useful feedback to program designers. For example, if path *c* is weak, this means the earlier reminder did not make the appointment salient on the day of the appointment; that people still forgot about it. In this case, one might suggest either shortening the time interval between the text reminder and the day of the appointment and/or one might suggest making the reminder more memorable. If path *b* linking salience to planning behavior is weak, then this suggests that making an appointment salient does not necessarily lead people to act on it by making necessary arrangements. Perhaps adding a prompt to the message to "make your plans" would strengthen this path. If path *d* linking planning to appointment keeping behavior is weak, perhaps patients need help planning more effectively. If path *a* is weak, perhaps the reminder needs to be more vivid.

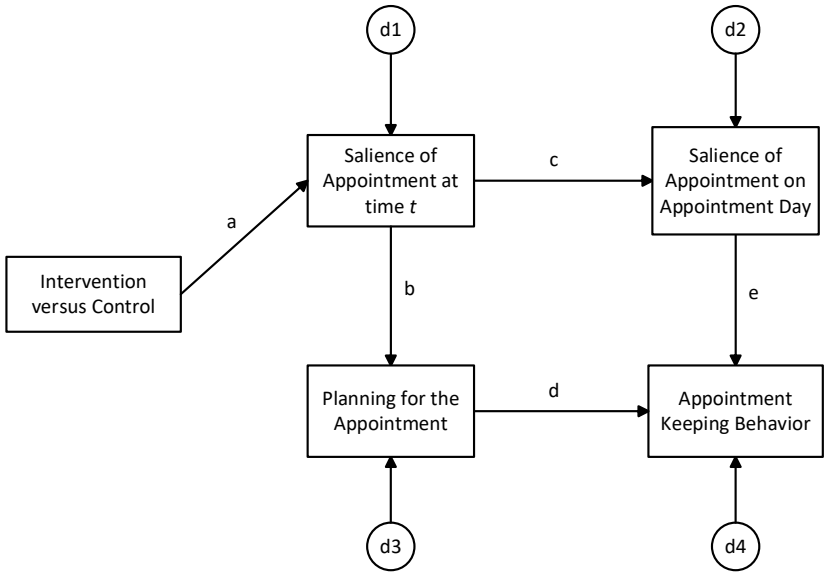


FIGURE 2.8. Text reminder example

With yet more thought, it might be evident that the program makes additional assumptions. One assumption is that everyone in the reminder condition does, in fact, receive the reminder. However, this may not be the case. Message delivery failures can be as high as 30% due to incorrect text addresses, networks being down, or people not owning smart phones (Stancombe, 2019). Low-income populations may not own smart phones or they may not keep them connected on a regular basis because some months they cannot afford to. Shelters that provide housing for the homeless often have policies

that prohibit cellphone use by residents in the shelter. Even if a message is received, it is not guaranteed the message will be opened by the recipient, read, and comprehended. To the extent that these compromising factors are operating, the effect of the program on making the appointment salient might be diminished. These variables could be added to the RET logic model as moderators of the effect of the treatment on appointment salience at time *t*.

Some reminder programs mention only the desired behavior in the text message whereas other programs also provide "nudges" of encouragement that highlight an advantage of appearing at the appointment. If a program uses nudges, the RET can evaluate if a "nudge" affects the perception it is intended to affect and if that perception is relevant to appointment keeping behavior. This is illustrated in Figure 2.9 for the case where a text reminder is sent to defendants three days before they are scheduled to appear in court for a driving infraction that includes the nudge "avoid a possible license suspension – be sure to appear." Note there are two ways the effect of the nudge can be influential. Path *b* posits the nudge makes people in the reminder condition more likely to believe they could get their license suspended if they do not appear. Path *f* is a program-mediator interaction and indicates the nudge might make the perception more salient on the day of the appointment.

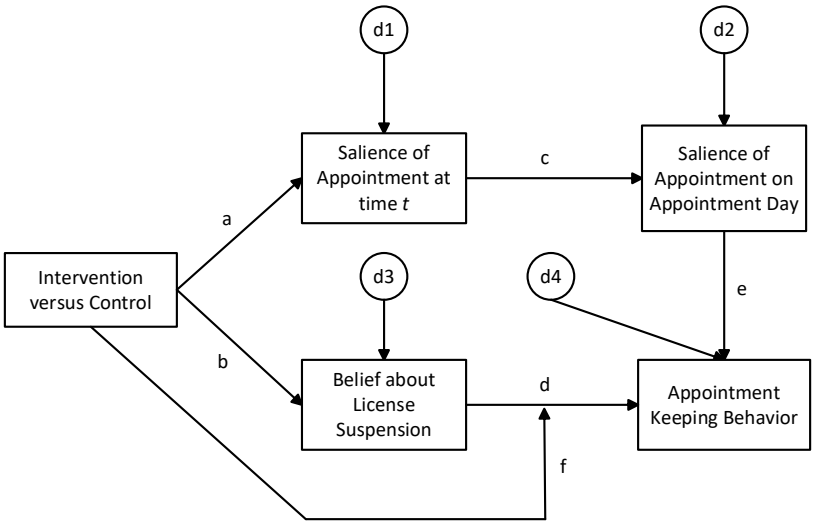


FIGURE 2.9. Text reminder with nudge

I could continue this exercise, but my general point is evident: Even seemingly simple interventions likely have an underlying logic model of mediators that can be

incorporated into an RET. Sometimes we must think more deeply about them relative to program activities. In RETs that evaluate the effects of a drug on an outcome, such as bupropion on depression, there may appear to be no formal "program activities" other than taking the medication. However, the drug invariably affects depression through biological mediators (norepinephrine, dopamine) whose effects may be moderated by individual difference variables, such as the type of depression. An RET focused on these mediators and moderators will be informative, but the kind of "program activity analysis," will be different than discussed above. There invariably are cases where a simple RCT will suffice to answer a question. However, I encourage you to adopt an RET mindset wherever feasible to identify relevant mechanisms and enrich your study.

Mapping Mediators Tied to Program Context

In some cases, mediators might be relevant that are not naturally occurring determinants of the outcome. In clinical psychology, for example, there is a large literature on therapeutic alliance and its impact on therapeutic outcomes. Therapeutic alliance refers to patient perceptions of the therapist as supportive and as fostering a collaborative relationship with the patient (Luborsky, 1976). Research suggests that perceptions of therapeutic alliance predict clinical outcomes independent of the type of psychotherapy used and across a wide range of clinical contexts (Ardito & Rabellino, 2011). If the outcome for an intervention is depression, therapeutic alliance is not a determinant of depression in the real world. However, it *is* a determinant in the context of change efforts used by therapists. Given this, one needs to take them into account when explaining and understanding program effects on outcomes. In clinical psychology, Garfield (1999) has referred to them as "non-specific" or "common" factors associated with therapeutic effects and are contrasted with "specific" factors that are more directly tied to the outcome per se rather than the context in which change is brought about (see also Cuijpers, Reijnders & Huibers, 2019).

As another example, research suggests that participant satisfaction with a program can impact long term outcomes. Zhang et al. (2008) found that patient satisfaction with treatment services for drug abuse was associated with reduced drug use one year after program completion. RET designers should routinely consider participant perceptions of the program, perceptions of program staff, the relationships between the staff and participants, and other program context variables as possible mediators of outcomes.

When Program Activities Include Moderators

Some programs formally incorporate moderator variables into program activities by presenting different information or developing different tasks for different subgroups. For

example, programs on adolescent delinquency might provide different program content to adolescent females as opposed to males on the presumption that the determinants of juvenile delinquency vary as a function of sex (Fagan & Lindsey, 2014). There is a large literature on **targeted interventions** that stresses the importance of adapting interventions depending on variables like ethnicity, age, and gender (e.g., Netto, Bhopal, Lederle, Khatoon & Jackson, 2010; Donohew, Helm, Lawrence & Shatzer, 1990). Aside from different message content, programs can formally vary other program facets. For example, programs that rely on messaging and education can differ in the sources they use to deliver the messages for different population segments (e.g., peers, health professionals, parents), the communication channels through which messages are delivered (e.g., use of face-to-face interactions, use of videos, use of podcasts), structural characteristics of the messages per se (e.g., the order of presentation of arguments within a message, the use of complex versus simple argument structures, message repetition), and the context in which the message is given (e.g., in a school, in a church, in small groups). When designing an RET, if an intervention uses different program elements for different subgroups, you will want to test for subgroup differences in program effects.

Summary of Program Analysis

In sum, identification of mediators targeted by different program activities can be identified by answering the question "*why* does this program activity (or set of activities) influence the outcome?" As one answers this question, mediators suggest themselves for inclusion in the RET. Sometimes program activities are vague, in which case potential mediators can be identified using extant theory or past research with respect to the outcome. Sometimes a program may seem so straightforward that an RET may seem unnecessary. However, usually a careful examination of program assumptions reveals ways in which an RET can be informative. Mediators also should be considered that may not be naturally occurring determinants of the outcome, such as therapeutic alliance and program satisfaction. Finally, moderators of program effects are evidenced when programs tailor program activities to different subgroups.

CONCEPTUAL TASK 2: BROADER ANALYSIS OF OUTCOMES, MEDIATORS AND MODERATORS

With the initial conceptual logic model in hand, the next step is to bring broader scientific knowledge and theory to bear to specify determinants of each mediator and additional determinants of the outcome beyond those targeted by the program. The purpose of doing so is (a) to identify confounders and omitted variables that need to be taken into account

to yield proper causal inferences, (b) to possibly add variables to the RET that might be of exploratory interest, and (c) to identify potential moderators to include to protect against falsely implied generalizability. I consider each topic, in turn.

Identifying Confounders and Omitted Variables

In an RET, evaluation of the effect of the program on the outcome is tied to random assignment to the treatment condition, as is the evaluation of the effect of the program on mediators. If random assignment is successfully implemented, causal analysis between the program and a mediator as well as the program and the outcome is reasonably straightforward. However, causal coefficients between mediators and outcomes, namely M→Y, typically are estimated from associational data rather than data tied to random assignment. This also is true when estimating causal relationships between mediators. Because confounds and omitted variables can bias causal estimates in such cases, it is important to render confounds irrelevant, either by design or statistically. To do so, we often first need to identity them so that we can measure them and bring them under control. Unfortunately, some researchers treat confounder control atheoretically; they routinely control for a "standard" set of covariates, such as SES, ethnicity, biological sex, and age, without thinking through the implications of doing so. Meehl (1971) has referred to this practice as **atheoretical partialling**. Achen (2005) calls it **garbage can regression**. I now describe how such atheoretical partialling can impede causal inference. After doing so, I provide heuristics you can use to specify meaningful confounds in RETs that need to be controlled. I emphasize here confounder analysis for the mediator-outcome link, but the principles generalize to other scenarios, such as causal relationships between mediators.

Parenthetically, in traditional regression modeling, a confounder, C, typically is controlled by adding it to the regression equation that predicts Y from M. For example, to estimate the causal effect of M on Y controlling for C, I estimate the equation $Y = \alpha + \beta_1 M + \beta_2 C$. The coefficient b_1 in the sample data is the estimated effect of M on Y holding C constant. To use this strategy, we must obtain a measure of C. Alternatively, we can control C by design. If C is socioeconomic status (SES), I might control for it by only including low SES individuals in my study.

Good and Bad Control of Potential Confounders

Cinelli, Forney and Pearl (2022) discuss what they call "good and bad" control of potential confounders. "Good control" is when control reduces bias in estimates of a causal coefficient; "bad control" is when control increases such bias. The Cinelli et al. (2019) discussion is important because it makes clear that the choice of confounders to

include in a modeling effort in order to reduce bias is a theoretical matter not a methodological side note.

Figure 2.10 presents a model representing classic thinking about confounds (to reduce clutter, I omit disturbance terms from this figure as well as subsequent ones). C is a confounder that influences both the mediator and the outcome, such as the effect of biological sex on homework completion (the mediator) and school performance (the outcome) in middle school youth. It turns out that middle school girls tend to be more apt to complete homework and they also tend to perform better in school than boys. In this case, C biases the estimate of the causal coefficient *b* because C contributes to the association between M and Y over and above the causal effect of M on Y. Thus, failure to control for biological sex will lead us to overestimate the causal impact of homework completion on school performance. Controlling for C is a form of "good control" because it removes the inflating effects of C on path *b*. Note that if either path *c* or *d* is weak, then control for C might be lower in priority because the bias C introduces will be minimal.

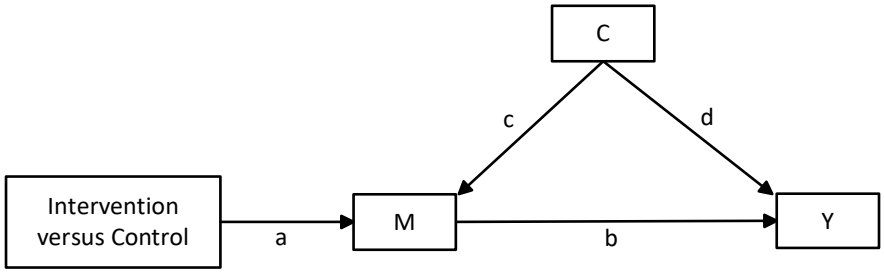


FIGURE 2.10. Common determinant confounder

Figure 2.11 presents a case where there are four potential confounders of the M-Y link, C1, C2, C3 and C4. However, C1, C2, and C3 all impact M and Y through the more proximal variable, C4. In this case, if I statistically control just C4, I have blocked the pathways through which C1, C2, and C3 impact the outcome and thereby rendered them harmless in terms of biasing the estimate of path *b*. This example illustrates the case where strategic control of a more proximal variable through which other confounders operate can adjust for multiple distal confounders. For example, suppose the outcome variable Y is a posttest measure of the consumption of nutritious foods and that C1, C2, and C3 in Figure 2.11 are biological sex, SES, and ethnicity. C4 might be a baseline measure of nutritious food consumption and I might assume that any effects of C1, C2 and C3 on Y are controlled by introducing C4 as a covariate in my statistical model. Note that not only do I indirectly control for C1, C2 and C3, but I also control for all other

confounders, measured or unmeasured, that influence M and Y through C4. Control for C4 is a case of "good control."

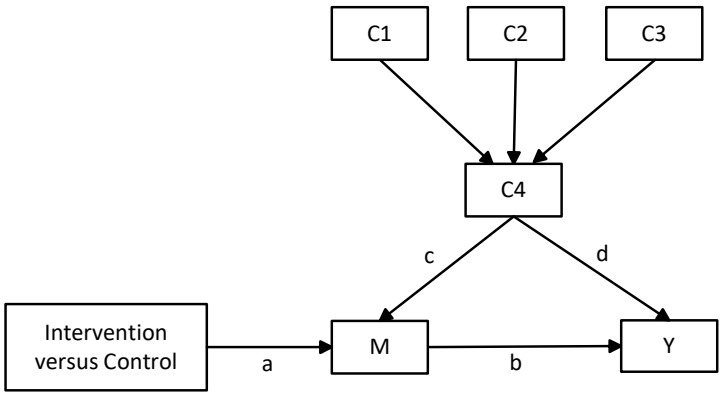


FIGURE 2.11. Using a proximal confounder to control for distal confounders

Figure 2.12 presents another example of "good control," where the confounder, C, impacts Y and is correlated with but not causally related to M. The association between M and Y is impacted not only by the causal effect of M on Y but also by the association between C and M and the effect of C on Y. For example, depression (M) and anxiety often are comorbid, i.e., depression occurs with anxiety but it is not caused by anxiety per correlation *d.* Both depression and anxiety affect suicide ideation. If anxiety is ignored, then its contribution to Y will be partially "credited to" depression (path *b*), introducing bias.

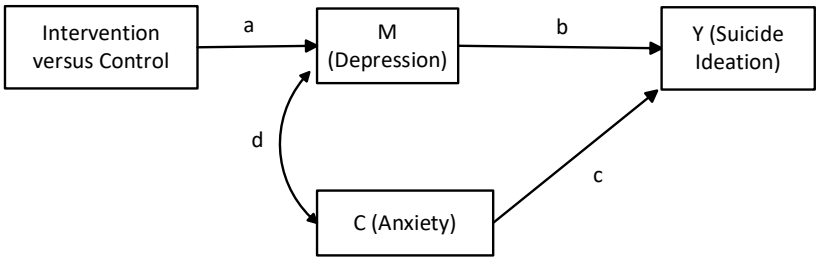


FIGURE 2.12. Correlated confounder that is not a common cause

Parenthetically, the dynamic in Figure 2.12 is a common source of estimation bias

in the social and health sciences outside of the context of RETs. As an example, some scientists argue there is a link between habitual coffee consumption and future coronary heart disease (CHD) per path *a* in Figure 2.13. Coffee consumption is correlated with smoking, but one does not "cause" the other. It is well known that smoking impacts CHD, per path *b*. If we do not control for the effect of smoking on CHD, we will overestimate the effect of coffee consumption on CHD because the effect of smoking on CHD is "passed" to path *a* as a function of the strength of the association between coffee consumption and smoking (link *c*). Controlling for smoking is a "good" control.

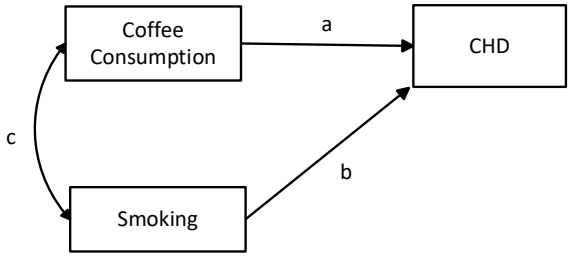


FIGURE 2.13. Coffee consumption and smoking example

Figure 2.14 presents an example of "bad control." The confounder, C, is correlated with both M and Y, but the correlations are spurious; there is no causal connection between C and M nor between C and Y. In this case, statistically controlling for C when estimating path *b* will produce a biased estimate of *b* because the association between M and Y will be adjusted for C as if C influences Y despite the fact that it does not. For example, a researcher might evaluate the effect of parent-adolescent relationship satisfaction (a mediator) on school performance (the outcome). Adolescent religiosity is correlated with both relationship satisfaction and school performance (due to the association of religiosity to variables like ethnicity, biological sex, and age of the adolescent) but not in a causal sense. As such, using religiosity as a covariate when regressing school performance onto relationship satisfaction can introduce bias into the mediator-outcome causal estimate (path *b*). This type of "bad control" shows that just because a variable is correlated with both M and Y, this does not mean it should be controlled. The broader model context needs to be considered. This form of bad control is sometimes referred to in the literature as **M-bias** (see Cinelli, et al., 2022).

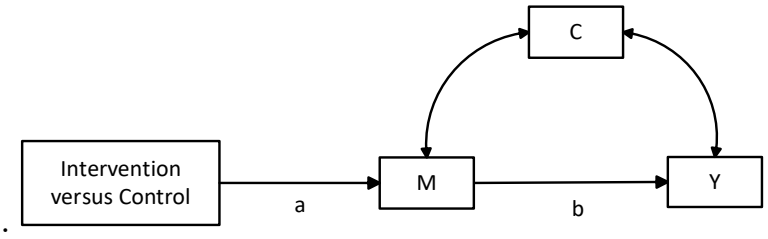


FIGURE 2.14. Treating a causally irrelevant confounder as causally relevant

Figure 2.15 presents another case of "bad control." C mediates the effect of M on Y. By statistically controlling for C when estimating the effect of M on Y, we block the very effect we want to estimate! In this case, we treat C as a "confounder" when its true function is that of an intermediate mediator. Controlling for C will bias the estimate of the overall causal effect of M on Y. This dynamic also holds if I seek to estimate the overall effect of M on Y but C only partially mediates some of that effect. By controlling for C, I remove some of the total effect I seek to estimate. As an example, suppose M is obesity and Y is the occurrence of heart attacks. A researcher might treat cholesterol levels as a confound when, in fact, it is a mechanism by which obesity impacts heart attacks. By controlling for cholesterol, I will underestimate obesity effects on heart attacks.

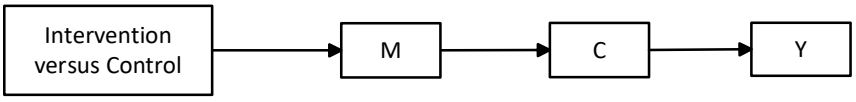


FIGURE 2.15. Treating a mediator as a confounder

Figure 2.16 presents a third example of "bad control." Suppose M influences an external variable C (by virtue of the solid arrow from M to C) but that C has no effect on Y. M and C are correlated and if we fit the model as specified in Figure 2.16, the coefficient estimates will be unbiased. However, if we mindlessly control for C when predicting M from T (path *a*) by misspecifying the causal dynamic underlying the association between M and C to be C→M (the dashed arrow) rather than M→C, this introduces bias into the estimate of path *a*. The bottom line is that just because M and C are correlated does not mean we need to control for C when evaluating path *a*. We need

instead to correctly specify the causal dynamic and let that specification guide our covariate strategy rather than covarying anything that is associated with, in this case, M.

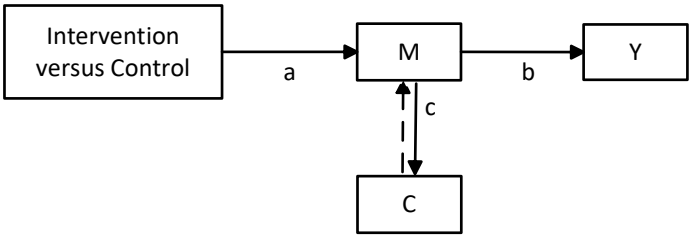


FIGURE 2.16. Misspecifying the effect of the mediator as a cause of the mediator

Figure 2.17 presents a similar example of a "bad control" but now applied to path *b*. C is an outcome of Y but we might control for it if we mistakenly think it is a cause of Y. If we control for the misspecified effect of C on Y, the estimate of the path linking M to Y will be biased because we adjust for a correlate of M that does not exert a causal influence on Y.

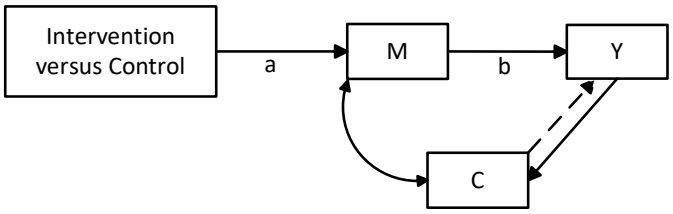


FIGURE 2.17. Misspecifying causal directions

Another form of bad control is when inclusion of a covariate inadvertently changes the essence or meaning of the construct you want to study. For example, several studies report a *tall bias effect* by correlating people's height with their earnings (Judge & Cable, 2004). These studies often find that taller people tend to have higher earnings than shorter people. When conducting their analyses, it is not uncommon for the researchers to statistically control for a host of potential confounds, one of which is weight. However, weight is, to some extent, influenced by height and controlling for weight shifts the

analysis from height effects to a study of the combination of height and body composition effects, which is not the focus of the tall bias effect.

Yet another form of "bad control" involves what is known as collider variables, per Figure 2.18. A **collider variable** is a variable that is influenced by both M and Y, hence it is correlated with each of them. The term collider is used because in an influence diagram, the arrows from variables that lead into the collider appear to "collide" on the variable that is the collider, per Figure 2.18. If we inadvertently statistically control for C when assessing the relationship between M and Y in this case, the association between M and Y can change and the coefficient for path *a* can be biased (see Pearl, 2009, for a mathematical proof). Collider effects also are known as **Berkson's paradox** (Berkson, 1946). It is another reason to be cautious about mindlessly controlling for any variable correlated with both M and Y.

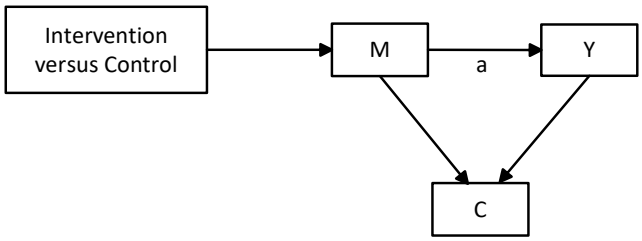
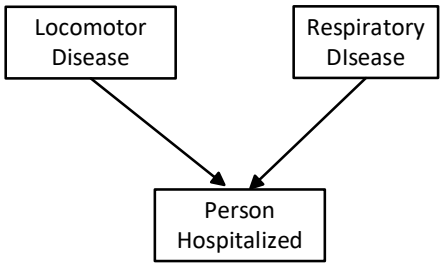


FIGURE 2.18. Control of colliders

A classic example of collider bias was reported by Sackett (1979) who studied the relationship between locomotor disease and respiratory disease. In the general population, the occurrence of these two diseases is statistically independent. Sackett examined the relationship between the diseases in a group of hospitalized patients and to his surprise he found a positive association between them. Without knowledge of the result in the general population, one might infer from Sacket's results that a causal link between the diseases exists because locomotor disease can lead to inactivity and inactivity, in turn, can cause respiratory disease. It turns out that both locomotor disease and respiratory disease impact to some extent whether someone is hospitalized but people who have both diseases simultaneously are especially likely to be hospitalized. I might frame the influence diagram for hospitalization as a function of these two variables as follows:



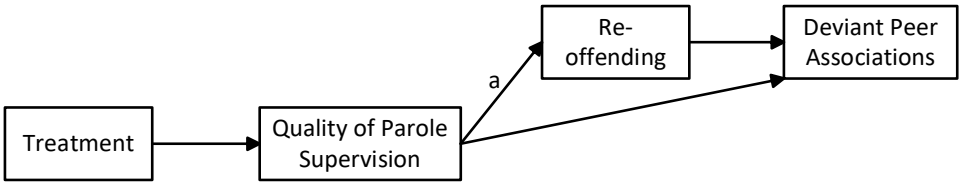
In this model, hospitalization is a collider that is influenced by locomotor disease and respiratory disease. If for the sake of pedagogy I score each disease dichotomously, then people with a score of 1 on locomotor disease and a score of 1 on respiratory disease should have a higher likelihood of being hospitalized than people with a score of 1 on one of the diseases but a score of 0 on the other disease, everything else being equal. Similarly, people with a score of 0 on both diseases will have the lowest probability of being hospitalized. Because Sackett studied only hospitalized patients, the variable of hospitalization was "controlled," i.e., it was held constant by design. Controlling for the collider in this case produced a bias in the estimated relationship between the two diseases because people for whom both diseases co-occur were more likely to be hospitalized and hence included in the sample. By restricting the sample to hospitalized patients, we essentially oversample people with both diseases relative to the general population. This, in turn, created an association between the diseases when in the general population, none exists.

Pearl and Mackenzie (2018) characterize collider dynamics using a coin flipping example. Suppose we flip a coin twice for each of a hundred trials and write down the results for each trial but only when at least one of the two flips in the trial shows heads. This will occur for about 75 of the trials. What we will find for the recorded trials is that every time the coin landed tails on the first toss, the coin will have landed heads on the second toss. This is because we conditioned on a collider by censoring all the tail-tail trials. If I score a head = 1 and a tail = 0, the calculated correlation between the first flip and the second flip across the 75 recorded trials will be -0.50 even though the tosses without the censoring are independent.

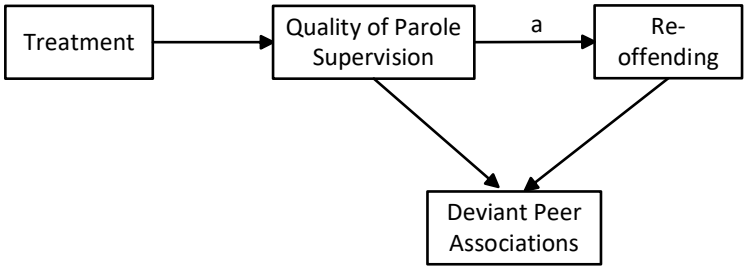
Sackett's analysis of collider effects is interesting because the control of the collider was inadvertently applied by virtue of study inclusion criteria, namely the participants were restricted to individuals who were in a hospital rather than by statistical means. This makes evident that a researcher's choice of inclusion and exclusion criteria for a study also needs to be considered relative to the causal conclusions one makes.

Collider bias can insert itself into mediational analyses in subtle ways. Suppose a program, T, aimed at recently released incarcerated persons seeks to influence the quality

of parole supervision these individuals receive (the mediator) which, in turn, is presumed to influence reoffending, i.e., engaging in criminal activity post release (the outcome). If the analyst statistically controls for other post-release outcomes that are causally impacted by both quality of parole supervision (M) *and* re-offending (Y), collider bias can result. Examples of such "bad" covariates might include post-release substance use and post-release associating with deviant peers. Here is a casual diagram that shows these causal relationships for post-release associating with deviant peers in the context of a model where I seek to estimate path *a*, the effect of the mediator on the outcome (note: deviant peer associations are measured after reoffending):



In this diagram, quality of the parole supervision influences re-offending which, in turn, impacts subsequent associations with deviant peers. If I retain these same causal paths in the model but re-arrange the spatial positions of the variable boxes, the role of deviant peer associations as a collider is obvious:



The above influence diagram, of course, does not dictate that statistical control of the collider be undertaken. The traditional SEM estimation algorithm (elaborated in Chapter 7) is to regress each endogenous variable in the model onto all variables with straight arrows pointing directly to them. Following this algorithm, the variable deviant peer associations would not be a predictor/covariate in any of the model equations. However, a researcher bent on atheoretical partialling might create collider dynamics by unwittingly and blindly including deviant peer associations as a covariate when examining the effects of quality of parole supervision on re-offending.

An often overlooked potential source of collider bias in RCTs and RETs is treatment dropout or attrition. If attrition from the study is non-random and attributable to both the intervention (e.g., some individuals drop out of the intervention because they find participation too time consuming) and to outcomes (only individuals for whom the treatment is working stay in treatment), collider bias can result when one statistically corrects for treatment dropouts; see Elwert & Winship, 2014). Collider bias also can manifest itself in mediation analyses in multiple mediator models with complex causal structures among the mediators.

Like confounders, collider bias varies in magnitude and in many cases it can be so weak as to be ignorable. If one or both of the implicated causal paths that creates a collider dynamic is weak (e.g., the M→C path or the Y→C path), collider bias often will be inconsequential. Greenland (2003) evaluated the consequences of varying magnitudes of collider links and found that *both* collider paths must be quite large for the bias to be non-trivial. In some cases, one can avoid collider bias through study design by weakening the causal paths involved in the collision. In the example on attrition, I might implement strategies that discourage people from dropping out of the study because it is too time consuming by providing more flexible scheduling of intervention activities. This, in turn, could eliminate or weaken the influence of treatment participation on dropping out of treatment and eliminate the collider bias from attrition.

Some methodologists make blanket recommendations against adjustments for any post-treatment measured covariates, in part, out of concern for collider bias. Other methodologists recognize that such a position is unnecessary for addressing certain substantive questions of interest (Hernán & Robins, 2020). Sometimes a collider serves a dual role as a collider and a confounder, meaning you are damned if you control for it and damned if you do not. Greenland (2003) notes that collider bias sometimes is considerably smaller than confounder bias, which favors controlling for the offending covariate but living with the small amount of collider bias that might result. I include a document on my webpage for Chapter 2 titled *Collider Bias and Mediation Analysis* that describes additional ways that colliders can covertly intrude and undermine mediation analysis and strategies for dealing with it in mediation analyses. I disagree with the assertion that making covariate-like adjustments to a post-treatment variable is never justified. To be sure, when we do so, we often must make certain assumptions for our inferences to be reasonable and it is incumbent on us to make those assumptions explicit and defend their viability. I say more about these matters in future chapters.

Guidelines for Confounder Identification

I could continue with more examples of bad controls (e.g., inappropriately treating an instrumental variable as a confound; see Ding, Vanderweele and Robins, 2017, and Pearl, 2011), but the general idea should be clear. Rather than atheoretically controlling for a multitude of covariates, you need to think carefully about the broader causal theory that is operating in your RET and how potential confounders "fit" into that theory. Mändli & Rönkkö (2023) discuss two perspectives that have emerged with respect to covariate control. One, known as the **frugal perspective** argues a control variable should be included as a covariate only if a compelling case can be made for its confounding properties. Carlson and Wu (2012) summarize this viewpoint as "when in doubt, leave it out." A second perspective is the **prolific perspective** that argues that more controls are better than too few because covariate inclusion reduces the chances of omitted variable bias (Antonakis et al., 2010). Mändli & Rönkkö (2023) consider arguments both for and against each perspective and conclude their appropriateness depends on the research context. Sometimes one approach is best, other times the other approach is best. I encourage you to read the Mändli & Rönkkö (2023) article for elaboration.

As will become evident in future chapters, explicit covariate control using reasonably valid measures of covariates is central to the analysis of RETs, especially when trying to rule out the biasing effects of confounds. We obviously cannot measure and control for every possible confound, so we seek to measure and control for those that we think are most important (i.e., that will create the most bias) and hope the ones we ignore are not inferentially consequential. One approach to confounder identification that I find useful when predicting Y from one or more mediators in a linear equation when I plan an RET is to focus attention on Y and then, either through literature reviews or consultation with experts, make a list of what I think are its core determinants. The list will invariably be long and will identify explanatory variables at many levels of analysis. At the most proximal level, the explanatory variables contain the name of the outcome in the variable definition. For example, if the outcome is obtaining a test for HIV, then the determinants might include such variables as emotions associated with *taking an HIV test*, perceived advantages and disadvantages of *taking an HIV test*, felt normative pressures to *take an HIV test*, and stigma associated with *taking an HIV test*. A second level of explanation identifies more general variables that do not reference the outcome name per se. For HIV test taking, they might include constructs like (a) personality, (b) values, goals, aspirations, and general attitudes, (c) mental health related constructs (e.g., depression, generalized anxiety), and (d) alcohol and drug use. A third level of explanation identifies variables from broader contexts, such as families, peers. work, schools, neighborhoods, providers, religion, media, government/policy, and cultural

contexts.

Next, I divide the compiled list of identified determinants into those that I think have moderate to strong associations with the dependent variable and those that have only weak relationships with it. In most cases, the exclusion of the former plausible covariates will be more consequential in terms of creating bias, so I lean towards prioritizing them in my analyses. From that list, I next prioritize covariates that I think also are moderately to highly correlated with the mediators in the equation. In general, covariates that are moderately to highly correlated with M *and* Y will be more likely to bias M→Y coefficients if omitted. Next, I review each of the surviving covariates relative to the different scenarios described above for "good" and "bad" covariates and eliminate any covariates that I judge to be "bad." For example, if one of the covariates reflects part of the mechanism by which my mediators impact Y, I might eliminate the mediator. Or if including the covariate in the equation changes the meaning of a mediator in ways that are theoretically counterproductive, I might exclude the covariate.

The number of plausible confounders at this point may still be intimidating and it may not be possible to measure and control for all of them. I rank order the remaining covariates as best I can from "most important to control" to "least important to control" and then measure and control for as many of the most important ones as feasible, but only after careful consideration of the broader context of the model in which they are embedded and the questions I seek to answer. As well, strategic control of proximal confounders (per Figure 2.3) can reduce the number of distal confounders to control. For example, an effective strategy in RETs often is to control for the baseline of the target endogenous variable because it serves the function of a proximal determinant that then controls for a multiple distal determinants (per Figure 2.11). I elaborate this strategy in future chapters.

The formal evaluation of confounders that are best to control also can be facilitated by expressing models in the form of traditional influence diagrams or DAGs. A useful tool for identifying confounders based on graphical analysis is provided on the website www.dagitty.net; see the link on the programs tab of my website for the program labeled graph theory. I provide a video for how to use the tool and some tips.

Although covariate control is generally seen as good research practice, Clarke (2005, 2009; Clarke, Kenkel & Rueda, 2018) has objected to traditional textbook treatments of omitted variable bias because they oversimplify the operative dynamics. Discussions of covariate control, Clark argues, usually deal with the case of a single omitted confounder but, in reality, there typically are so many potential omitted confounders that it is almost impossible to know how they are going to bias coefficients when considered multivariately. Some plausible confounders may be positively

associated with the outcome while others are negatively associated with the outcome. If we use a heuristic for selection that leads us to only choose a subset of the positively associated confounders to measure and control to the exclusion of influential negative confounders, then their inclusion might actually increase bias because they constitute a systematically incomplete picture of the multivariate causal dynamics. The ultimate effects of including confounders and colliders in model equations depends on the correlations between them and the excluded variables, the strength of the effects of the included and excluded variables, and their respective variances. Clarke is pessimistic about the use of control variables and, in all honesty, I am sympathetic to many (but not all) of his arguments (see also Ritter & Vance, 2011). I personally think we need to do the best we can and that common sense, theory, and past research often will take us a long way towards reasonable causal inference (Pearl, 2009; Pearl & Mackenzie, 2018). However, we also must keep in mind the tentativeness of our conclusions in some contexts, with the amount of tentativeness being a function of the likelihood that additional consequential confounders exist that were not taken into account.

The Secretary of Defense, Donald Rumsfeld, at a press conference in 2002 made the following infamous remarks when discussing the U.S.-Iraq war:

> *...as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tends to be the difficult ones.*

Despite the doublespeak, there are lessons to be taken from this missive. When we design RETs, we need to think long and hard about (a) what we know (the known knowns), (b) influences that we know likely operate but that we cannot measure or control (known unknowns), and (c) protecting ourselves from being misled by unknown factors that we are not even aware of but that could lead us astray (unknown unknowns). I return to the matter of the statistical control of confounders in later chapters once I develop additional basics in statistical modeling.

In conclusion, confounder control in RETs is an important *theoretical* enterprise in RETs. You must think long and hard about how confounders and omitted variables can bias the causal inferences you want to make and how you can go about reducing those biases. It is not acceptable to have a standard set of covariates that you always use nor to approach the matter atheoretically. You need to think through confounds for every core

endogenous variable in your model. For more details on confounder control, see Pearl (2009, 2011), Elwert and Winship (2014), Greenland and Pearce (2015), Rohrer (2018), Vanderweele (2019), Myers et al. (2011a, b), Mändli and Rönkkö (2023), and Cinelli et al. (2019). I revisit this topic in future chapters.

Adding Mediators for Exploratory Purposes

Once program-based mediators and confounders have been identified for use in an RET, you might also consider adding theoretically meaningful mediators of the outcome that the program does not formally address but that could be of interest to program designers and administrators. The idea is that the program might be improved by expanding it to address a malleable mediator that shows promise in the RET. By including exploratory mediators in the RET, their contributions to the outcome relative to other mediators can be empirically documented. For example, suppose a program to reduce adolescent drug use focused on two mediators, (1) the perceived negative short- term consequences of using drugs, and (2) the perceived negative long-term consequences of using drugs. Suppose based on a literature review we decide to include in the RET for exploratory reasons a determinant external to the program, namely peer resistance skills. The RET might find this variable is a solid predictor of future drug use, thereby suggesting to program designers they expand their program to focus on it.

Adding Moderators to Evaluate Generalizability

In addition to mediation, RETs address moderators of program effects. Whereas mediation addresses the question of *why* a variable, X, impacts another variable, Y, moderation asks the question "for whom, where, and when does X impact Y?" It seeks to identify boundary conditions of effects. For example, we might ask, "for whom does X influence Y, and for whom does it not?" We then seek to identify the characteristics that distinguish these two groups, and in doing so, we identify a moderator variable(s). Or we might ask, "in what contexts does X influence Y, and in what contexts does it not?" We then identify the characteristics that distinguish these different contexts, and in doing so, identify a moderator variable(s). Or, we might ask, "when (in terms of time or timing) does X influence Y, and when does it not?" We then identify the characteristics that distinguish these time periods (e.g., during middle school but not during high school identify a moderator variable(s).

If one adds moderator analysis to an RET, a central task is to identify what moderators to explore. In program evaluation, one mindset for doing so is to think about moderation as evaluating the generalizability of (a) program effects on mediators, (b) program effects on outcomes, and (c) mediator effects on outcomes. In the intervention

research that I conduct with adolescents, I am interested in whether my programs are equally effective for different gender groups, different ethnic groups, for youth of different ages (middle school versus high school) and across socio-economic strata. To the extent my programs are equally effective across these groups, I feel more confident in the generalizability of the program. I therefore include these variables as moderators in my RETs. This is probably the most common use of moderation analysis in an RET.

A second mindset for identifying moderators is to draw on extant theory. For example, in psychology, a personality variable called the *need for cognition* reflects the extent to which people prefer to hear relevant arguments, supporting information, and the logic underlying propositions as opposed to preferring to "skip the details" and be told by a credible expert the position or orientation s/he should take (Cacioppo, Petty & Kao, 1984). Suppose you are evaluating a parenting program designed to improve parenting that does so by making five parenting recommendations to program participants. The program relies on having multiple credible sources and celebrities endorse each recommendation but does not explore the reasons why parents should do so in much depth. This program likely will be more effective for people who are low in need for cognition than for people who are high in need for cognition. As such, need for cognition might be chosen as a moderator in the RET. If the moderator is affirmed, then discussion with program staff and administrators about how to deal with the dynamic might ensue.

A third mindset for identifying moderators is to adopt a purely exploratory approach using mixture modeling to empirically identify *post hoc* subgroups of individuals who exhibit different causal coefficients for program effects on mediators and for mediator effects on outcomes. This exploratory approach is discussed in future chapters.

Summary of Broader Conceptual Analyses

In sum, another task for designing an RET is to perform a broader conceptual analysis of the mediators and outcomes for purposes of identifying confounders and omitted variables to address. Confounder analysis requires consideration of "good" and "bad" controls, which means avoiding atheoretical partialling. Instead, confounder analysis is integrated into the theory that guides the RET. As well, broader conceptual analyses should seek to identify exploratory mediators that can yield information for program improvement. Finally, broader conceptual analyses should consider identifying moderators to address issues of generalizability and the need for tailoring programs.

CONCEPTUAL TASK 3: DEFINING MEANINGFUL EFFECTS IN RETs

A third task researchers face when designing an RET is defining what constitutes a

meaningful program effect. Evaluation of effect sizes is considered by some to be a statistical enterprise, but I contend it is very much a substantive matter that requires careful consideration of non-statistical issues. I focus my discussion on program effects on the targeted outcome but the ideas also are relevant for program effects on mediators.

In the social sciences, many researchers rely on the statistical significance of an effect in an RCT to decide if an effect is meaningful: If the difference in outcome values for treatment versus control conditions is statistically significant ($p < 0.05$), then the difference is said to be meaningful. With statistical significance, however, the primary focus is on a null hypothesis that the difference between two population parameters, such as the difference between means or the difference between percentages, is *exactly* zero. If a result is statistically significant ($p < 0.05$), all we can conclude is that the difference is not "exactly zero," a conclusion that is not very compelling. For example, if I find a statistically significant difference between two weight reduction programs for obese individuals, all I can conclude from the p value is that the amount of reduction is not the identical in the two programs; p values say little about the magnitude of the program difference. Rather than seeking to determine if a difference is "not exactly zero," Cohen (1994) argues that analysts instead should seek to make statements about the magnitude and meaningfulness of program or group differences. Magnitude estimation moves us away from framing questions in terms of *whether* two groups differ to framing questions in terms of *the extent to which* two groups differ.

A common strategy for characterizing effect size meaningfulness is to use Cohen's (1988) criteria for defining small, medium, and large effects, with effect sizes that are medium or large being declared as meaningful. For the analysis of a mean difference between two groups, medium effect sizes correspond to a Cohen's d of 0.50. This maps onto an effect size of about 6% explained variance when expressed as a percent of variance accounted for. By this logic, if a mean difference between two group accounts for less than 6% of the variation in the outcome at the population level, then the difference is deemed non-meaningful.

The problem with this logic is that Cohen's criteria are arbitrary. Why 6%? On what basis did Cohen choose 6%? It turns out that Cohen himself expressed reservations about his criteria, noting that characterizing effects as small, medium, or large "is an operation fraught with many dangers" because such "definitions are arbitrary" (Cohen, 1988, p. 12). Cohen stated that the terms are "relative…to the specific content and research method being employed in an investigation" (p.25), arguing that his criteria are "recommended for use only when no better basis for estimating the effect size index is available" (p. 25). Cohen also said that "these proposed conventions were set forth… with much diffidence, qualifications, and invitations not to employ them if possible,"

noting that “the values [have]… no more reliable basis than my own intuition” (p. 532). Cohen laments that the criteria “were needed in a research climate characterized by a neglect of attention to issues of magnitude.”

These statements by Cohen do not indicate an enthusiastic endorsement by him of the use of his criteria for characterizing effect sizes. Funder and Ozer (2019) report Cohen as later telling friends that he regretted ever suggesting the standards. Other methodologists also have expressed skepticism. Glass, McGaw, and Smith (1981, p. 104) state “there is no wisdom whatsoever in attempting to associate regions of the effect-size metric with descriptive adjectives such as ‘small,’ ‘moderate,’ ‘large,' and the like.” Lenth (2008) refers to Cohen’s effect size criteria of "small", "medium", and "large" as T-shirt effect sizes that lead power analyses to arrive at the same required sample size no matter what the characteristics of the outcome or the setting. For a medium effect size, for example, the scientist will choose the same sample size regardless of the accuracy of one’s measure, the homogeneity or diversity of the individuals in the population, the severity or positiveness of the outcome, and other features of the phenomena under study.

Making judgments about the trivialness versus meaningfulness of a program effect is necessarily tied to specific criteria within the substantive area of interest. As such, the matter is as much a conceptual/substantive issue as it is a statistical issue. In applied domains, the meaningfulness of a (parameter) difference, be it for a positive or negative event, should be judged in terms of (a) how likely it is to affect the overall quality of life of individuals, (b) how many individuals are affected, (c) the sustainability or reversibility of the effect over time, (d) the vulnerability (ability to “defend oneself” against the negative event) or entitlement (helping the rich get richer at the expense of the poor) of the affected individuals, and (e) the costs (broadly defined) of implementations addressing the event, among other criteria. A 5% difference in mortality rates is more consequential than a 5% difference in transient, minor head colds. About 600,000 people die of heart disease in the United States every year. If a new drug could be found that reduces this by only 1%, it translates into a savings of 6,000 lives per year, which over a five-year period is 30,000 lives. However, an effect size of 1% is well below what Cohen characterized as a small effect size. As another example of contextualization, a program to promote medication cost savings that results in, on average, $800 savings per year is meaningful to individuals who are poor but less so to those who are rich.

When making judgments of trivialness or meaningfulness, there inevitably will be different constituencies (e.g., scientists, practitioners, politicians, managers, consumers, workers) who bring different values and different criteria to bear. If a study estimates a 10% difference in preferences for chocolate ice cream as a result of an advertisement, this might be seen as a trivial difference by some but a truly important difference to a

marketing firm specializing in national advertising for ice cream.

There are many examples of attempts to define effect meaningfulness beyond a reliance on statistical significance or Cohen's standards. In program evaluations focused on disease burden, an index known as the quality-adjusted life year (QALY) is sometimes used to create benchmarks for meaningful effects. One QALY equals one year in good health. QALY values are assigned to individuals based on weighted utility values associated with a given state of health. For example, a year of life lived in a situation with a utility of 0.5 (such as being bedridden) has a QALY value of 1 year × 0.5 or 0.5 QALYs. If you are dead, your QALY score is 0. Controversy surrounding the index exists about the assignment of utilities and the definition of benchmarks. However, standards and methods for determining utilities and benchmarks have been proposed and often are used by different governments for program evaluation (e.g., Torrance, 1986; Beresniaket al., 2015; Holmes, 2013; Brazier & Tsuchiya, 2015).

Benchmarks for meaningfulness have been suggested for psychological constructs, but these often have a limited evidence base. For example, the classic Center for Epidemiologic Studies Depression Scale (CES-D) is scored from 0 to 60, with higher numbers indicating greater levels of depressive symptoms. A score of 16 on the scale has been said to be the cut-off for "significant depression." It turns out the score of 16 was informally suggested in one of the first studies published on the CES-D by Weissman, Sholomskas & Pottenger (1977) and it gained traction thereafter. Weissman et al. specified the value based on the ability of the cutoff to predict a score of 7 or greater on the Raskin Depression Scale (Raskin, Schulterbrandt & Reatig, 1967). Ironically, the Raskin scale itself has an arbitrary metric with little scientific justification for the meaningfulness of a score of 7. More recent studies have explored the utility of a cutoff of 16 using signal detection theory, with mixed results (Vilagut et al., 2016).

The above discussion makes evident that evaluating the meaningfulness of a result is a complex process. Social scientists have tended to avoid thinking about the matter by blindly focusing on $p < 0.05$ criteria or adopting the arbitrary cutoffs reluctantly suggested by Cohen (1988). The task is even more challenging when the metric of the outcome (or mediator) is arbitrary. An arbitrary metric is one whose specific numerical values are not concretely linked to specific locations on the underlying dimension being measured nor to external behavioral benchmarks that give the scores meaning. For most Americans, for example, if told that a person weighs 90.7 kilos, people will have no sense of how heavy the individual is. If told that another weighs individual is 68.0 kilos, one knows the latter individual weighs less that than the former individual, but by how much? Most Americans would have no idea. This is because, for them, the kilogram metric is arbitrary (even though it is completely valid and reliable). Many measures used in clinical

trials, such as self-report measures of depression, anxiety, and self-esteem, have this arbitrary quality, making it difficult to judge meaningfulness about group differences.

In this book, I directly tackle issues of evaluating the meaningfulness of effect sizes. In doing so, I make use of the concept of latitudes (or numerical regions), and distinguish three types, (a) a latitude of no effect, (b) a latitude of meaningfulness and (c) a latitude of effect ambiguity. Suppose a social scientist evaluates a program designed to enrich income by teaching job skills to low-income individuals. Suppose that the population mean annual income increases in the treatment condition by $3,055 and in the control condition it increases by $3,054. Technically, the program had an effect by raising income relative to the control condition; the null hypothesis of no group difference in the populations is false. However, all would agree that this difference is trivial and inconsequential. The two groups are, for all intents and purposes, *functionally equivalent*. Suppose the population level difference between treatment and control individuals was $500 instead of $1 in favor of the program. Would we conclude the groups are functionally equivalent? What if the difference was $1,000? What is the range of values on the outcome that most people would agree if treatment minus control mean differences fall within that range, the groups can be declared as functionally equivalent. This range of scores is called the **latitude of no effect**.

The **latitude of meaningfulness** is defined as the range of values that most people would agree represent a meaningful difference between the groups on the outcome. For example, almost all would agree that a $5,000 mean difference is meaningful. When specifying the latitude of meaningfulness, we identify the smallest income difference that defines a meaningful effect and this becomes the lower bound of the latitude of meaningfulness. Values greater than it in the desired direction are deemed meaningful.

Finally, the **latitude of effect ambiguity** is the range of values between the lower bound of the latitude of meaningfulness and the upper bound of the latitude of equivalence. It is a "gray area" where people disagree about effect size meaningfulness.

The approach I use is to specify the three latitudes for each facet of an RET, namely (a) for the effect of the program on an outcome, (b) for the effect of the program on a mediator, and (c) for the effect of a mediator on an outcome. I then locate a given study result into a latitude and interpret the result accordingly. I elaborate my approach in Chapters 10 and 11. The point I make here is that making such judgments is not a methodological nor statistical detail to be left to statisticians. Rather, it is a substantive matter that requires careful thought at the level of theory and practice.

MULTIPLE OUTCOME RESEARCH

Many RCTs address multiple outcomes. The motivation for using multiple outcome measures varies. In some cases, there is only a single outcome variable but multiple measures of the same outcome are obtained; the different measures are seen as **interchangeable indicators** of the same construct. This approach allows researchers to adjust parameter estimates for measurement error and it provides multiple pieces of information about the parameter of interest (see my discussion of SEM in Chapter 3).

Another reason for using multiple outcomes is because a single measure may not adequately capture treatment effects in multiple important domains. In this case, the multiple measures are not interchangeable indicators of the same construct but rather reflect conceptually distinct aspects of a broader outcome domain each of which are of interest in their own right. For example, apathy, somaticism, and negative affect are often seen as correlated but conceptually independent facets of depression. Each of these facets might have different mediators and moderators so one might construct an RET logic model for each outcome independently and test the models separately. Alternatively, there may be overlap in the logic models for separate outcomes and one might find it informative to integrate them into a single logic model, per Figure 2.19. In this model, mediator 1 impacts two facets of the outcome (facet 1 and facet 2), mediator 2 impacts all three facets, and mediator 3 impacts only outcome facet 3. Mapping these differential causal relationships might be useful for purposes of program evaluation. As well, it might be the case that causal relationships exist among the multiple outcomes. For example, negative affect may impact both apathy and somatization. In such cases, one likely would want to model and take into account the causal relations among the outcome facets.

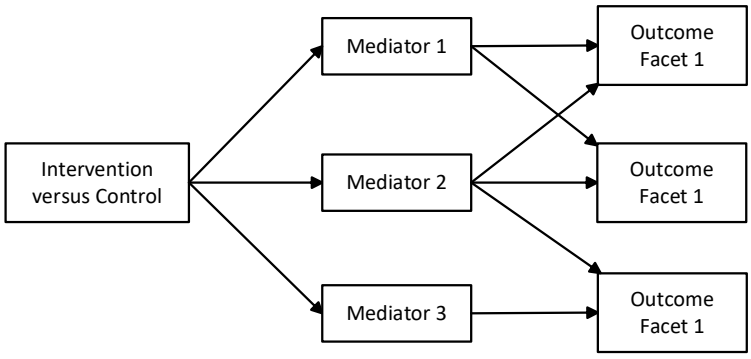


FIGURE 2.19. Multiple outcome RET (disturbance terms are omitted for clarity)

Requiring trialists to specify a single primary outcome to answer a single primary question (Friedman et al., 2015).is counter to RET philosophy. RETs often have more than a single question to answer and when theory dictates multiple variables/outcomes should be focused on, one should do so.

CONCLUDING COMMENTS

Before conducting an RET, you should specify the relevant mediators and moderators to consider. These variables are identified through careful analysis of the program activities as well as qualitative research and literature reviews about each mediator and outcome. Based on this, it probably is best to formalize your conceptual logic model using an influence diagram. To make reasonable causal inferences, you must be able to identify confounders and omitted variables that can bias estimates of causal coefficients so you can adequately address them in the RET. The choice of confounders and omitted variables to focus on can only be resolved by careful consideration of their place in the broader theoretical context that guides the RET. Confounder control is as much a theoretical matter as it is a methodological matter. RET design also requires that we think about effect generalizability or boundary conditions. Are there meaningful subgroups for whom the treatment or program is not working well? Who are they and why is the treatment not working for them? A final critical task for RET design is to specify what constitutes a meaningful effect for mediators and outcomes. These definitions ultimately are used to assist decision making regarding how to improve or streamline a program. Addressing this matter is far more complex than using a generic rule like “a Cohen’s d of 0.50, or its equivalent, is meaningful” for every mediator and every outcome in the logic model. More nuanced consideration is needed, as I illustrate in future chapters. Finally, it sometimes is recommended that a randomized trial specify a priori a single outcome that will be used to answer the primary study question. This practice is suspect in RETs, both on methodological and theoretical grounds.