10

Evaluating Effect Sizes in RETs

Data do not speak for themselves - they need context

- ALLEN WILCOX

INTRODUCTION

It was in 1994 when the psychologist Jacob Cohen published his now classic article “The earth is round (p < .05).” This article raised numerous questions about the merits of statistical significance testing and emphasized the need to incorporate magnitude estimation and confidence intervals into result interpretation. Just what is a reasonably sized and meaningful program effect? When can we conclude that a program affects a mediator in a way that is meaningful; or that a mediator affects an outcome in a way that is meaningful? These are difficult questions to answer. In this chapter, I describe selected statistical indices of effect size that one often encounters in the statistical literature for documenting effect sizes. None of them is perfect. All have strengths and weaknesses. The idea is that when you want to evaluate the meaningfulness of an effect, you can think about it from different perspectives and using different effect size indices. After presenting the indices, I provide additional background to help you make meaningfulness judgments in the context of RETs.

My focus is on evaluating effect size for a given link in a mediational chain rather than the effect size for the full mediational chain considered in an omnibus sense. This is consistent with my view expressed in prior chapters that the analysis of individual links of a mediational chain is more informative for purposes of program evaluation than omnibus mediational chain effects. If we know that a given link in a mediational chain is “broken” and needs repairing or that it is unacceptably weak, then our task is to figure out how to go about addressing that link to make the program effect on the distal outcome stronger. If the program fails to meaningfully change a mediator, then we need to alter the program activities so that it does, in fact, change the mediator. If the mediator, contrary to assumptions, is not relevant to the outcome, then we might streamline the program to remove its focus on the irrelevant mediator and replace the program component aimed at the mediator with one focused on a more relevant mediator; or we might figure out a way of strengthening the effect of the mediator on the outcome. Prioritizing individual link analysis is not the rule for all forms of mediation analysis, but it usually is key to program evaluation. After reviewing the ins and outs of different effect size indices commonly used for the analysis of individual links, I briefly characterize effect size indices for omnibus mediation effects through a given mediational chain. I defer discussion of effect size indices for moderators to Chapter XX. I show you how to implement the effect size analyses described here in future chapters.

INDICES OF EFFECT SIZE IN RETs

Many people associate effect size with standardized indices, such as squared correlations or Cohen's *d*. However, effect sizes also can be characterized in unstandardized form, that is, in units of the variables' raw metrics. If I tell you that a weight loss program leads to an average weight loss of 25 pounds (11.3 kilos), then this conveys, without standardization, a sense of the magnitude of the program effect. If I tell you male assistant professors earn, on average, $10,000 more per year than female assistant professors, this conveys a sense of the magnitude of sex differences in units of dollars. Raw mean differences and unstandardized regression coefficients *are* effect size indices and they can be useful and sometimes more intuitive than standardized indices. If the goal is to describe how much an outcome changes given a certain amount of change in a mediator, the unstandardized regression coefficient for that mediator provides information directly to the point: For every one unit the mediator increases, the mean of the outcome is predicted to change by the value of the coefficient. A regression/path coefficient characterizing how much, on average, viral load decreases as a function of the dose of a medication is causally informative and reflective of dose effects, yet it is expressed in unstandardized rather than standardized units.

There are many standardized effect size indices that social scientists rely on. In this section of the chapter, I first discuss the use of p values as an index of effect size and then consider indices that rely on the variances of disturbance terms in a model. I next consider Cohen's *d*, indices based on the concept of exceptions to the rule, standardized regression coefficients, and an index known as the number needed to treat. Finally, I describe risk differences, relative risks, and odds ratios. Each index evaluates effect size from a different vantage point. When describing indices of effect size in RETs, I often will reference standards that researchers use to judge the magnitude of effect size for a given index, but I do not necessarily recommend the use of those standards. Indeed, some of the recommendations for the same index are quite divergent from one another.

I have my own favorites for indices of effect size and I will identify them later in the chapter. However, I will characterize numerous effect size indices because they are popular in the social and health science literatures so you should be aware of them as well as their limitations.

Effect Size 1: p Values

In RETs, the most common use of p values is to gain perspectives on the null hypothesis of a zero mean difference between groups, a zero proportion difference between groups, and/or a zero path or regression coefficient. There are, of course, other uses of p values but a core question we want to know is if the difference between groups on some parameter is "statistically significant."

Some researchers when conducting such tests report p values to multiple decimal places. Other researchers simply note that the p value is less than the alpha level, typically set to 0.05 (e.g., $p < 0.05$). The former practice assumes that p values contain useful information about effect size and should be reported in ways that reveal that information. The latter approach instead sees the role of p values as helping one to make a binary decision; either reject or fail to reject the null hypothesis. These opposing orientations have their roots in different conceptualizations of null hypothesis testing during the early days of modern statistical theory, one by Ronald Fisher and the other by Jerzy Neyman and Egon Pearson.

Articulation of the concept of a p value in null hypothesis testing is generally attributed to Fisher. A p value represents the probability of obtaining an effect equal to or more extreme than that observed in one's study if one assumes the null hypothesis is true. The lower the p value, the more unlikely it is the null hypothesis is tenable because the data are too unlikely to have patterned themselves in the way they did if one assumes the null hypothesis is true. At some point of low probability, the null hypothesis is rejected. The p value is thus seen as a quantitative index of the strength of evidence against the null hypothesis.

Neyman and Pearson, by contrast, argued that researchers should specify *a priori* what the cutpoint for defining a "low probability" is in order to provide a clear basis for rejecting the null hypothesis. In current day research, this standard is taken to be 0.05, but it can be adjusted upward or downward depending on context. Once a researcher rejects or fails to reject the null hypothesis using the p value using the defined cutoff for a "low probability," Neyman and Pearson contend that the p value story ends. All that matters for purposes of null hypothesis testing is if $p < 0.05$ or not so that one can decide whether to reject or fail to rejected the null hypothesis. By contrast, scientists using Fisher's framework believe that knowing the observed p value to several decimals conveys quantitative information about the strength of evidence against the null hypothesis. As such, it should be reported with some degree of precision to give a sense of that strength.

The approaches of Fisher and Neyman-Pearson have been misunderstood, mischaracterized, and debated for decades. Indeed, many would object to the simplistic characterization I offer above but my intent is to give you a general sense of the contrasting viewpoints. Arguments against using p values as an index of effect size are provided as part of a formal position statement of the American Statistical Association (Wasserstein & Lazar, 2016). In one portion of the report, the ASA states one of six formal principles about p values: "*A p-value, or statistical significance, does not measure the size of an effect or the importance of a result*." The principle is accompanied by the following text:

> *Statistical significance is not equivalent to scientific, human, or economic significance. Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different p-values if the precision of the estimates differs.*

The basic idea is that because traditional uses of p values to evaluate differences are influenced by sample size, they can be poor indicators of effect size. Two studies can observe the same p value yet report very different mean differences, proportion differences, or correlations simply because of study differences in sample size. Similarly, two studies can observe different p values yet report the same mean differences, proportion differences, or correlations because of study differences in sample size. A minor effect in a study can produce a small p value if the sample size in that study is large, just as a major effect can yield a large p value if the sample size is small.

According to some, reporting exact p values represents a lack of understanding of the logic of significance testing. In null hypothesis testing frameworks, one *a priori* specifies the probability of the computed test statistic that s/he would consider rare or unlikely if the null hypothesis is true, usually 0.05. If the probability of the statistic is less than that value, then you reject the null hypothesis. If not, you fail to reject the null hypothesis. The decision is either to reject or fail to reject the null hypothesis. End of story (Lang, Rothman & Cann, 1998).

Although most agree that p values have properties that make them unsatisfactory as indices of effect size, some scientists argue they should still be reported to three or four decimals to accommodate readers who may not agree with the cutoff value you selected. By reporting a more exact p value, researchers are free to make a case for rejecting or failing to reject the null hypothesis based on a different cutoff value that they might feel is more justified than the one you selected. Others argue that the conventions for defining a "rare probability" are so widely accepted these days that such a position is specious.

Some scientists argue that knowing if a p value is between 0.05 and 0.10 is important because it signals a "trend" towards an association between variables that is then worth keeping an eye on in future research. This takes p values outside the null hypothesis testing framework in which they were derived, which critics find objectionable. It also is seen as opening the door to arbitrarily shifting decision rules by viewing an effect as viable even when the p value is larger thn 0.05, which is considered by many to be counter to good science.

The controversy about p values as indices of effect size obviously is complicated and I can't develop it fully or resolve it here. I personally gravitate towards the Neyman-Pearson position that uses p values not as indices of effect size but as a part of the null hypothesis testing framework where one seeks to make a binary decision about the null hypothesis. Stated another way, the p value as traditionally applied to null hypothesis testing can inform us whether an effect exists (i.e., is non-zero), but the p value will not provide good perspectives on the size of the effect. I almost always report p values given their prevalence in the field and to address the question of whether a non-zero difference exists, but I rely on other indices of effect size to address matters of effect size magnitude and meaningfulness in an RET. Granted, there will be exceptions to this, but in general, evaluations of the meaningfulness of effect sizes in RETs should be based on more than $p< 0.05$.

Effect Size 2: Variances of Disturbance Terms

When attempting to characterize how well one or more predictors "explains" an outcome, many researchers rely on the variance of the disturbance term associated with the outcome when predicting it from a set of target variables. The most common practice is to interpret the variance of the *standardized* disturbance term, which is the disturbance variance when the outcome has been standardized. This variance, routinely reported by SEM software, mathematically equals 1 minus the squared multiple correlation for the equation in question.[1] In RETs, this index is often used to characterize the strength of the presumed effect of one or more mediators on an outcome or the effect of a treatment condition of a mediator or an outcome. For example, if the outcome is the intention to obtain a vaccine and the mediators are (a) the perceived likelihood you will contract the virus the vaccine protects against (perceived susceptibility), (b) the perceived effectiveness of the vaccine, (c) the perceived health risks of the vaccine, and (d) the perceived hassles/costs of getting the vaccine, then the proportion of unexplained variance in the intention from these four mediators, collectively, equals the standardized disturbance variance and one minus the standardized disturbance variance reflects the the proportion of explaine variance in the intention from these four mediators.

When we evaluate effect size for each predictor in an equation individually, we typically want to adjust for the other predictors as well as covariates in order to remove their biasing effects on effect size estimates of the target predictor. In this case, we estimate the proportion of explained variance by the target predictor after controlling for the covariates and the other mediators. This statistic is called the **squared semi-part correlation** for the predictor. A squared semi-part correlation for a predictor of 0.10, for example, means that the predictor accounts for 10% of the variance in the outcome after controlling for the other predictors (covariates and mediators) in the prediction equation.

Squared semu-part correlations are *not* routinely reported by SEM software. It turns out you can approximate/estimate the squared semi-part correlation for a given predictor in a linear equation with multiple predictors in SEM knowing just the critical ratio associated with its regression coefficient for the predictor (from the t or z test for the coefficient), the sample size, the overall squared multiple correlation, and the number of predictors. All of these statistics are routinely reported by SEM software. I provide the relevant equation for calculating the squared semi-part correlation from these statistics in the Appendix and a computer program on my website to execute it. In the vaccination example, I might find the squared semi-part correlation for predicting variation in vaccination intentions from perceived susceptibility is 0.05, for perceived effectiveness it is 0.15, for perceived health risks of the vaccine it is 0.14, and for the perceived hassles/costs of getting the vaccine it is 0.04.

The squared semi-part correlation formula in the Appendix also can be used to calculate the explained variance in a mediator as a function of the treatment condition holding constant relevant baseline covariates. For example, I might find the treatment condition (treatment versus control) yields a squared semi-part correlation of 0.08 for perceived susceptibility, 0.06 for perceived effectiveness, 0.11 for perceived health risks, and 0.09 for perceived hassles/costs. All of these values reflect the "effect size" for the

[1] I discuss exceptions to this in future chapters, such as the case when there are correlated disturbances.

effect of the treatment on the respective mediator and when multiplied by 100, they index the percent of (unique) explained variance by the predictor/determinant.[2]

Many researchers want to know what value of unique explained variance represents a “small,” a “medium,” or a “large” effect with the idea that medium or large effects usually are deemed meaningful. If we generalize from Cohen’s (1988) standards, proportions of explained variance of 0.01, 0.06, and 0.14 constitute small, medium, and large effect sizes, respectively. These standards are different from what Cohen suggests as standards for squared zero order correlations also reflect the proportion of explained variance but without the introduction of covariates. For squared correlations, the standards suggested by Cohen for small, medium, and large effects are 0.01, 0.09 and 0.25. Why the standards differ is not completely clear. Fender and Ozer (2019) suggest squared correlations of 0.01, 0.04 and 0.09 to define small, medium, and large effects.

Other researchers have conducted meta-analyses of effect sizes in different fields and suggest standards based on the distribution of effect sizes in those fields. For example, Schäfer and Schwarz (2019) analyzed proportion-of-variance-accounted-for effect sizes in clinical trials with psychological outcomes that were pre-registered with the federal government and found that the bottom third of the effect sizes had an average squared correlation of about 0.002, the overall average squared correlation was 0.03, and the top third had an average squared correlation of 0.17. For non-registered clinical trials, the corresponding values were 0.04, 0.13, and 0.48, which are notably different. The typical effect sizes also differed across sub-disciplines; the median effect size based on squared correlation logic in social psychology was 0.09 and in biological psychology it was 0.25. Gignac and Szodorai (2016) meta-analyzed correlations from social and personality psychology and found the average squared r was 0.04, with squared r*s* of .01 and .09 at the 25th and 75th percentiles, respectively. Gignac and Szodorai suggested researchers use these values as guidelines for declaring effects as small (0.01), medium (0.04), or large (0.09). In short, the “conventions” that have been suggested for labeling effect sizes as small, medium, or large have been varied and somewhat inconsistent.

Although standardized effect sizes in the form of squared correlations or squared semi-part correlations are popular in several disciplines, they have been criticized in other disciplines (see Greenland et al, 1991; Greenland, Schlesselman, & Criqui, 1986; Pek & Flora, 2018). One complaint is that they define effect size meaningfulness strictly in terms of unique explained variance ignoring the common explained variance among the predictors in the equation. Such common variance, the argument goes, should also be taken into account. Another criticism is that standards for declaring them as meaningful

[2] The formula in the Appendix and used in my program applies to OLS regression. However, in many cases, it can be used effectively with maximum likelihood estimation in SEM. See my discussion in Chapter 11.

are arbitrary and not well justified substantively. Examples abound where “small” effects sizes based on the above guidelines have consequential effects and where “large” effect sizes have minor effects (Kelley & Preacher, 2012). A classic example of a small effect size having a meaningful effect was reported by Rosnow and Rosenthal (2003). In the late 1980s, a study on the effects of taking low-dose aspirin once a day on heart attacks was conducted with 22,000 physicians who were randomly assigned to aspirin versus placebo conditions. The study was prematurely terminated because the initial results suggested too many deaths would occur in the control group and that it was unethical to deprive control physicians of taking daily aspirin. The effect size in r squared units was less than 0.01 and would be dismissed using traditional standards. As another example, an organization might implement a policy that leads people to arrive to work 1 minute earlier and this effect may account for 50% of the variance in arrival times because there is little variation in those times. The annual amount of savings to the company by having workers arrive 1 minute earlier might be trivial despite the large standardized effect size.

Sometimes the standards offered in the field for these effect size indices can be used as rough rules of thumb for small, medium, and large effect sizes, but I personally believe one usually must dig deeper into the substantive context of an effect to make informed judgments of effect meaningfulness. Relying on generalized standards for proportion of explained variance that ignore context is, in my opinion, a risky enterprise.[3]

Effect Size 3: Cohen’s d

When comparing two means, a popular index of effect size is that of Cohen’s *d*. Cohen’s *d* is a raw mean difference but expressed in standard deviation (SD) units. Suppose the starting salary of assistant professors at major universities is $80,000 for males, $78,000 for females, and the standard deviation, calculated separately and then pooled for males and females, is $10,000. The mean salary difference is $2,000. To express this difference in SD units, I divide it by the pooled standard deviation of the two groups, which yields a value of 2,000/10,000 = 0.20. The raw mean difference translates into 0.20 standard deviation units. Cohen (1988) has offered guidelines for judging these effect sizes: A *d* of 0.20 is a small effect, 0.50 a medium effect, and 0.80 a large effect.

The formula for *d* using population notation is

$$\delta = \frac{\mu_1 - \mu_2}{\sigma_{\text{Reference}}}$$

[3] For an interesting discussion of cases where one might prefer a non-squared correlation to a squared correlation as an indicator of effect size, see Darlington and Hayes (2017).

where $\sigma_{\text{Reference}}$ is the reference standard deviation used to standardize the raw mean difference. In most cases, $\sigma_{\text{Reference}}$ is either the standard deviation of one of the groups or the weighted average of the standard deviations of the different groups if one assumes the group σ are equal or close to one another. The sample version of the parameter is

$$d = \frac{m_1 - m_2}{SD_{\text{Reference}}}$$

where m_1 and m_2 are the sample means and $SD_{\text{Reference}}$ is the sample estimate of the population reference standard deviation.

In RCTs, some researchers define Cohen's *d* using only the control group outcome SD rather than the pooled SD (Glass, McGaw & Smith, 1981). The logic is that the control group SD better reflects the population SD independent of the intervention because the SD in the treatment group can be affected by participation in the intervention. However, if one believes the posttest population σs in the treatment and control groups are equal (or roughly so), then it is better to use the pooled SD because the population estimate of the σ is then based on a larger sample size. I provide a program on my website for calculating pooled SDs.

In some RCTs, researchers use the baseline standard deviation of the outcome for the total sample as the estimate of the reference standard deviation for Cohen's *d* because it is unaffected by the treatment and it is based on the full sample size. This practice can be problematic because outcome variability can change in both the treatment and control groups over time due to history effects, maturation, and other time-confounded dynamics during the RCT, per my discussion in Chapter 4. For example, in a therapy involving 16 weekly sessions, the variance of outcome scores might change in the control group between the baseline and the posttest because of broader environmental forces at work or because of maturation dynamics due to aging. This is particularly true in research with children or the elderly and in research with long time intervals between assessments. To counter such effects, the safest approach is to use the posttest data to estimate within-group variability for the treatment and/or control groups when calculating *d*, not the pretest SD. In the final analysis, you must decide what is most appropriate reference SD to use.

Cohen's *d* assumes that a standard deviation is a meaningful comparator against which to judge a raw mean difference. Unfortunately, this is not always the case because standard deviations can be arbitrary. An item on a depression scale might ask individuals to rate the number of days in the past week that "I was sad." Alternatively, the stem might be "I was very sad." The former phrasing can yield more variability than the latter because the more extreme phrasing of the second item biases ratings towards fewer days;

respondents use a more restricted part of the response scale based on the more extreme phrasing. In this case, variability is determined by the arbitrary word choice of the scale constructor. For variables whose metric is less arbitrary, like income, weight, or height, perhaps the SD is a meaningful comparator; or perhaps not, depending on context.

Cohen's *d* can be misleading depending on how light or heavy tailed a distribution is. An example is given by Wilcox and Tian (2011), which I show in Figure 10.1. The left panel shows two normal distributions with Cohen's *d* equal to 1.0 for the mean difference between the two distributions. The right panel shows two distributions with slightly heavy tails but with the exact same raw mean difference as the left panel. Cohen's *d* for it is 0.30. The difference between a Cohen's *d* of 1.0 and a Cohen's *d* of 0.30 is typically judged to be substantial, but the figures suggest the mean differences are comparable. Indeed, the raw mean differences *are* identical. Susceptibility to light and heavy tailed distributions is a dubious property of Cohen's *d* and it also applies to squared correlations and squared semi-part correlations as well.

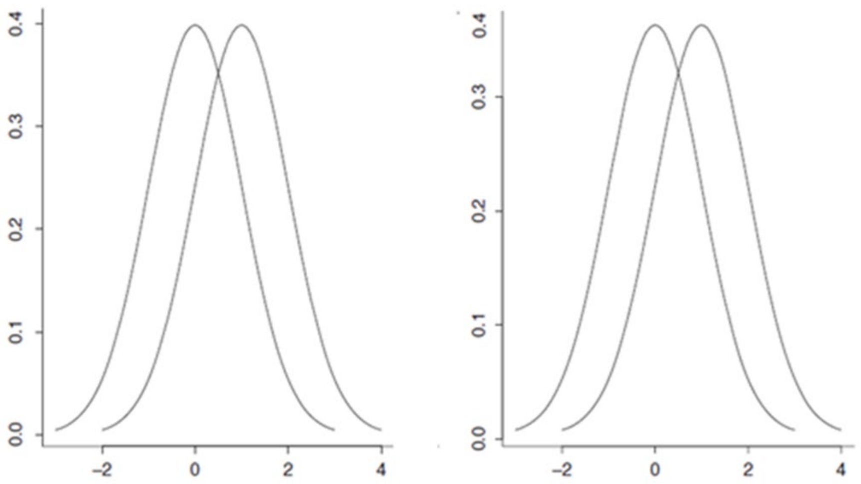

FIGURE 10.1. Illustration of variations in Cohen's d

A point of controversy when using Cohen's *d* in the presence of covariates is whether to use the unadjusted SD as the reference standard deviation or to use a covariate adjusted SD as the reference SD. In an RET, I might compare treatment and control group means on a mediator or on an outcome using biological sex, ethnicity, and social class as covariates to improve statistical power or to adjust for sample imbalance. The within-group standard deviations will be smaller if I remove the effects of these covariates on the mediator than if I do not. This, in turn, will affect the value of *d*, making it larger because I now use a smaller SD as the reference standard. Which SD is more appropriate, the covariate adjusted or covariate unadjusted one?

Olejnik and Algina (2000) argue it is important to reflect the full range of population variability when standardizing Cohen's *d* so they advocate using unadjusted SDs. Glass, McGraw and Smith (1981) also favor this practice but note possibly reporting both indices of *d*, one with and the other without covariate corrections. To me, the choice depends on one's goals but I generally lean towards trying to equate the groups as much as possible on nuisance variables (covariates) to gain a better sense of program effects. This then favors using the covariate adjusted SD. I discuss the choice in more depth below when I consider another effect size index called the probability of exceptions, where I can make the underlying logic for covariate control more explicit.

Technically, sample-based Cohen's *d* (and squared correlations and squared semi-part correlations) are positively biased estimators of their population counterparts; they tend to overestimate the true population effect size, on average. Some methodologists suggest bias corrections to adjust for this fact, such as Hedge's *g* in place of Cohen's *d* or omega squared in place of the squared correlation. Sometimes introducing the corrections creates greater sample-to-sample fluctuations in estimates, leading to less efficiency in exchange for less bias of the estimates; in other words, there can be tradeoffs to using the bias corrections. The positive bias tends to be small as the absolute value of the effect sizes become larger and as N > 80. Given this, I usually do not invoke the corrections when my N is larger than 80 and even if it is smaller, I am somewhat hesitant to do so because of other properties of the estimators.

In sum, Cohen's *d* is widely used. Although it seems straightforward, it raises numerous challenges in RETs. First, standard deviations often are arbitrary, making $\sigma_{\text{Reference}}$ arbitrary which, in turn, makes *d* arbitrary. Second, the choice of $\sigma_{\text{Reference}}$ is not straightforward; it can be defined adjusting for covariates or not adjusting for covariates and it can be estimated using only the control group SD or both the control group and treatment group SD. Third, Cohen's *d* can be impacted by heavy tailed distributions and can be misleading as a result. Fourth, as discussed earlier for squared correlations, small *d* can reflect either trivial or meaningful effects and large *d* also can reflect either trivial or meaningful effects (see Rosnow & Rosenthal, 2003; Kelley & Preacher, 2012 for examples). Cohen's *d* of 0.20 (small effect), 0.50 (medium effect), and 0.80 (large effect) can be thought of as very rough rules of thumb, but the bottom line is that they need context to be interpreted properly. I elaborate this in Chapter 11.

In the literatures I work with, researchers often report Cohen's *d* but then never say another word about effect size nor make any attempt to provide context or meaning to the reported *d* value. It is as if reporting *d* releases one from the responsibility of having to provide substantive context and meaning to it. In program evaluation settings, it is my experience that clients do not understand *d* and simply telling them the conventional standard in the scientific literature that a *d* of 0.50 is a “medium” effect size or a *d* of 0.80 is “large” is dubious and some would say even irresponsible. I revisit this matter below.

In sum, I tend not to rely on Cohen’s *d* in RETs because of the many shortcomings of it described in this section. No index of effect size is weakness free, but I personally am more comfortable using other approaches. Some methodologists disagree with me and this seems to be a case where reasonable people can disagree.

Effect Size 4: Exceptions to the Rule

Another effect size index is a variant of what is known as the **probability of superiority** or the **common language effect size** (Vargha & Delaney, 2000; McGraw & Wong, 1992; Wilcox, 2017). When we observe a statistically significant mean difference on a variable between two groups, we often form a “rule” or principle with respect to that difference. For example, empirically we typically find that on average, males are taller than females. From this finding, we might form the general rule that “males are taller than females” or “males tend to be taller than females.” However, there are exceptions to the rule. Just how often will I encounter a case where a female is taller than a male? If I randomly select a male and a female from the general population, what percent or proportion of the time will the female be taller than the male? How many “exceptions to the rule” are there? The index I propose here to quantify this is what I call the **probability of exceptions to the rule**. It documents the pervasiveness of exceptions to the formed generalization.

I develop the logic of this index first for the case of predicting a continuous outcome from a binary predictor. Suppose I find in an RET that the children of parents who use guilt as a discipline strategy have, on average, more depressive symptoms than children of parents who do not use guilt as a discipline strategy. How many exceptions are there to this “rule”? Or, suppose I find that individuals who participated in a medication adherence program tend to adhere to their medication protocol more so than individuals in the control group. How many exceptions are there to this “rule?” I can quantify such exceptions by calculating the probability that a randomly selected individual from the “disadvantaged” group (e.g., the control group; the group with children whose parents use guilt as a discipline strategy) has a “better” outcome score than a randomly selected individual from the “advantaged” group (e.g., the treatment group; the group with children whose parents do not use guilt as a discipline strategy). This probability reflects the proportion of “exceptions to the rule” in the target population. I signify it as P_E. For example, I might find that 10% of the time (a probability of 0.10), a randomly selected child whose parents use guilt as a discipline strategy will have lower depression than a randomly selected child whose parents do not use guilt.

Statisticians have developed methods for estimating P_E (Vargha & Delaney, 2000; McGraw & Wong, 1992; Wilcox, 2017; Ruscio, 2008), although they frame it differently than I have here. There are numerous technical details (such as how to deal with ties), but I do not delve into them here. Estimation of the probability of exceptions can use non-parametric strategies (Wilcox, 2017) or strategies that assume normally distributed outcomes. The two forms of estimation often are close in value even when non-normality is present (see Ruscio, 2008; McGraw & Wong, 1992) but exceptions can occur. In the Appendix, I present a computational formulae for estimating P_E as applied to OLS regression. I also can use the formula from the Appendix that makes use of Cohen's *d* to gain a sense of P_E for published research. This is useful because because Cohen's *d* is widely reported and if I want to put the results in perspective of the probability of exceptions, I can convert the reported *d* to P_E.

As examples, Johnston et al. (2011) evaluated a randomized trial to reduce anxiety and compared an on-line CBT intervention with a waitlist control. They found CBT reduced anxiety relative to the control group with a Cohen's *d* of 1.44. This translates into a P_E of 0.15, indicating that if one repeatedly randomly selects pairs of individuals from the treatment and control conditions, about 15% of the time, the control individual will be less anxious than the treated individual. These are the exceptions to the rule. Georgia-Salivar et al. (2020), report a Cohen's *d* of 0.37 for the effects of an intervention designed to increase relationship satisfaction for couples. This translates into a P_E of 0.40, i.e., there are 40% exceptions to the rule that people in the treatment condition are more satisfied with their relationship than those in the control condition. I provide programs on my website for calculating P_E using OLS regression. The program often can be used with maximum likelihood based SEM to yield reasonable approximations of P_E in SEM contexts. I also provide non-parametric methods based in bootstrapping. See the Appendix for details.

For two continuous variables, if a mediator, M, and an outcome, Y, are reasonably positively related, then we form rules like "people who score higher than others on M also tend to score higher on Y." There are several ways of operationalizing an exception to the rule for such cases. If an individual is above average (i.e., above the mean) on M, then we would expect that individual to also be above average on Y. For what proportion of individuals is this not the case? From a different perspective, if two variables are reasonably positively related, then if I randomly select an individual with a higher score on M than some other individual, I would expect the first individual to also have a higher score on Y than the second individual. For what proportion of individuals is this not the case? It turns out that the answer to these two questions is identical when documenting probabilities of exceptions for two continuous variables. I present the formulae for calculating P_E for this scenario in the Appendix and I provide programs on my website for applying the formula. As examples, in a mediation analysis, Budge, Adelson and Howard (2013) report that avoidance coping styles (a mediator) are positively associated with the outcome of depression. When I applied the formulae from the Appendix, the P_E for the association they reported was 0.28; 28% of the individuals who were above average on avoidance coping strategies, contrary to the "rule," were *below* average on depression. Kendall et al. (2015) in an RET report that coping efficacy (the mediator) was associated with reductions in anxiety; the path coefficient they reported was -0.49, $t = 3.06$, $p < 0.05$. The "rule" one derives from this result is that people with higher levels of coping efficacy have lower levels of anxiety. The P_E was 0.46; 46% of people who were above average on coping efficacy were, contrary to the rule, also above average on anxiety.

For P_E, keep in mind that if there is no program effect or no association between variables then P_E will equal 0.50; it will be as likely to encounter individuals who are exceptions to the rule as individuals who conform to the rule. As P_E approaches 0.50, the effect in question is weaker. Also, if the correlation between M and Y is negative, then P_E reflects exceptions to the rule that high scores on M are associated with low scores on Y.

Covariate Adjustments for P_E

Like Cohen's *d*, when calculating P_E a decision must be made about whether to adjust for covariates in its calculation. Statistical control usually is desirable; otherwise the estimate of P_E as an index of effect can be biased upward or downward because of confounds. Suppose the baseline covariates in my analysis comparing a treatment to a control group on a mediator or outcome are biological sex (male or female), ethnicity (White versus non-White) and SES. If I calculate P_E with no covariate adjustments, I essentially compare a randomly selected individual from the treatment group with a randomly selected individual from the control group ignoring their sex, ethnicity, and SES. The person randomly selected from the treatment group might be a Black, low SES male and the person randomly selected from the control group might be a White, upper SES female. All that is reflected in P_E is if the former person has a higher outcome score than the person from the control group and this tabulation is used when calculating P_E.

By contrast, when I use covariates that control for income, sex, and ethnicity, the comparison changes. If the randomly selected person from the treatment group is a Black, low SES male, then his score is compared with a randomly selected person from the control group who also is a Black, low SES male when calculating exceptions to the rule. The scenario that ignores covariates contains "noise" due to sex, ethnicity and SES when calculating P_E whereas the second scenario of covariate adjustment does not. Which scenario are you more interested in? Which gives you a better sense of the magnitude of the program effect? I personally lean towards the use of the covariate control approach but scenarios might occur where you are only interested in the general probability of exceptions. A weakness of the fully non-parametric approaches to P_E estimation is that they cannot adjust for covariates. The approaches I provide on my website allow for such control.

In sum, another approach to evaluating effect size is to determine how many exceptions there are to the "rule" implied by a causal coefficient. The fewer the exceptions, the stronger the effect. The question becomes, how many exceptions to the rule is one willing to tolerate before the "rule" becomes a "non-rule." If there are only 1% exceptions to the rule, is the rule useful? How about 10%? How about 25%? How about 45%?[4] If the rule is "wearing a mask reduces the risk of becoming infected with a deadly virus," and the exception rate is 5% (P_E = 0.05), does that give you pause about the rule? What if the P_E is 0.30? Does your tolerance for more exceptions shift depending on the severity of the consequences of an exception? Is it qualified by the vulnerability or entitlement status of the individuals who experience exceptions, such as disadvantaging children or the elderly? Affirmative answers to these queries imply definitions of a reasonable P_E are context driven.

I personally like to report P_E because I find it to be both informative and relatively easy to understand. I find it keeps me honest by discouraging me from over-generalizing the trends that I see in the data. To be sure, the index has weaknesses. For example, when a pair of individuals are randomly selected for comparison, one from each group, we do not *how much* higher or lower one person's score on Y is than the other person; we only know if it is higher or not. Information about the magnitude of the group differences must be obtained from other effect size indices.

Effect Size 5: Standardized Regression Coefficients

Another popular practice for effect size analysis is referencing standardized path/regression coefficients. It turns out, there are different types of standardized regression/path coefficients that can be used. Suppose my focus is on estimating the causal effect of X on Y. One type of standardized regression coefficient is when I standardize both the X and Y variables so that each has a mean of zero and a standard

[4] If the exceptions are greater than 50%, then the "rule" is not a rule. Rather, the exceptions are the rule.

deviation of 1.0 in the regression analysis. This is the form of standardization that most statistical software reports. I use the generic term **standardized path coefficient** to refer to it. A second type of standardized coefficient is when one standardizes only the Y variable but not the X variable. I refer to it as a **partially standardized coefficient** or a **Y standardized coefficient**. The first type of standardization is typically used when both X and Y are continuous or quantitative with many values. The second type is used when X is binary and Y is continuous or quantitative with many values. I discuss each in turn.

Standardized Regression Coefficients

For the case of continuous X and continuous Y, some researchers use the magnitude of the absolute value of the standardized regression coefficient as an index of effect size. These coefficients usually (but do not have to) range from -1 to 1. For the bivariate case, the standardized regression coefficient when Y is regressed onto X equals the correlation between X and Y. With more than one predictor, this property breaks down and the standardized path coefficient represents a partial coefficient. Specifically, a standardized path coefficient indicates for every one standard deviation that the predictor increases, how many standard deviations the outcome is predicted to change, holding constant the other predictors in the equation.

The effect size concept is that the larger the absolute standardized coefficient, the larger the effect size, everything else being equal. Acock (2014) suggests that an absolute standardized regression/path coefficient less than 0.20 is weak, one between 0.20 and 0.50 is moderate, and one greater than 0.50 is strong.[5] Note that a focus on standardized regression coefficients shifts the underlying causal theory from the idea that people's scores on the predictors per se are what matter to the idea that people's outcome is impacted by their z-score position *relative to other people in the population*. Some argue that such a shift in focus is questionable (see Achen, 1987, for elaboration). I revisit this matter in Chapter 17.

Partially Standardized Regression Coefficients

A partially standardized regression coefficient is used for a binary predictor X and a continuous outcome Y. Only the Y variable is standardized to have a mean of 0 and a standard deviation of 1.0. The binary predictor usually is dummy coded so that the coefficient for it in a regression analysis equals the mean of the standardized Y for the group scored 1 minus the mean of the standardized Y for the group scored 0, i.e., the reference group. In such cases, the partially standardized regression coefficient is the

[5] These guidelines and the reliance on standardized regression coefficients are not applicable if suppression dynamics are evident for a given coefficient.

number of Y standard deviations by which the group means differ. For example, if the partially standardized coefficient equals 1.5, then the raw mean difference between the two groups maps onto a difference of 1.5 standard deviations of Y.

For the case of two groups, it turns out that the partially standardized coefficient is an analog to Cohen's *d* but instead of converting the raw mean difference to a standardized difference based on the within group SDs, the standardization is implemented relative to the full SD of Y across the two groups, i.e., $\sigma_{\text{Reference}}$ is different. Some argue that the use of the full standard deviation of Y to standardize the effect rather than the traditional pooled within group SD per Cohen's δ is not appropriate for RCTs because the full SD of Y is artificial; it reflects variability in Y where half the population has experienced an intervention and the other half of the population (the control group) has not (assuming a 1:1 random assignment allocation).

The bottom line is that all standardized regression coefficients in an RET are somewhat dubious because the standard deviation of the outcome, Y, (and for that matter, the SDs of all the mediators) reflect an artificial population in which half of the population has received an intervention and the other half has not. Does it make sense to talk about how a one unit change in X or how a one standard deviation change in X impacts the number of standard deviations that Y changes if those standard deviations are not reflective of anything in the real world? My own preference is to stay focused on unstandardized path coefficients for the core parameters of an RET where this issue is then less relevant.

Effect Size 6: Number Needed to Treat

Another index of effect size used for the case of binary outcomes coupled with a binary predictor reflecting the treatment versus control condition is known as the **number needed to treat** (NNT). Suppose the proportion of alcoholics in a recovery program who begin drinking heavily again within a year of program completion is 0.12 and the corresponding proportion for those in the control condition is 0.25. The NNT is the number of people we need to treat by having them participate in the program in order to have one more "success" (full abstinence from drinking) than if we just left people alone per the control condition. If the NNT is 5, this means we would need to treat 5 alcoholics with the program in order to have one "success" relative to doing nothing at all. I present the formulae for computing the number needed to treat in the Appendix and present a program on my website to apply the formula. In the alcohol example, the NNT is 7.69; for every 8 or so patients we treat with the recovery program, we will have one additional "success" of full abstinence from drinking than had we not implemented the program. Thinking on a grander scale, for every 1,000 patients we treat, about 1,000/7.69 = 130 more of them will be fully abstinent if they complete the program versus doing nothing.

NNT is an index of effect size in the sense that the smaller the NNT, the larger the effect. The question becomes, at what point does the NNT become so large that it is no longer worth pursuing the treatment/intervention? If I need to treat 2 people to have one "success," is the treatment worthwhile? If I need to treat 5 people to have one "success," is the treatment worthwhile? If I need to treat 10,000 people to have one "success," is the treatment worthwhile? If an informational campaign is low cost, easy to distribute or administer, and prevents one death per year for every 5,000 people exposed to it, is it worthwhile given its NNT is 5,000? Consider the following argument: There are 210 million adults in the United Sates, so an NNT of 5,00 would translate into saving 42,000 lives per year if all of them are exposed to the campaign. Perhaps the intervention is indeed worth it despite an NNT of 5,000 when one considers this fact. If an educational intervention is costly, resource intensive, and time-demanding but reduces high school dropouts by 1 for every 500 students exposed to it, is it worthwhile given its NNT is 500? What if its NNT is 50? What if the program is inexpensive and not resource demanding? I find that discussions with program administrators and staff about what constitutes a reasonable effect size in NNT terms often raises issues of program cost, practicality, and the impact of the program on people's quality of life (either positively or negatively). Such discussions can be quite revealing.

Some disciplines routinely dichotomize outcomes so that NNT can be applied because the NNT only applies to binary predictors and binary outcomes. For example, in clinical psychology, it is common to classify each person in a randomized trial as having shown "clinically meaningful change" or not and then to compare the treatment and control groups on the percent of people who showed meaningful change. As examples of NNT, Borkovec and Costello (1993) compared cognitive behavior therapy (CBT) for anxiety with a non-directive (control group) therapy. They found the percent of people exhibiting meaningful change in the CBT and control conditions was 58% and 27%, respectively. This yields an NNT of 3.22; for every three additional patients each therapy treats, CBT will have one more success than the control therapy. Clarke et al. (2005) evaluated a web-based treatment for depression and found that 56% of treated participants still had clinical depression whereas 76% in the control condition still had clinical depression. This translates into an NNT of 5; one needs to treat 5 patients more in the therapy to have one more "success" than doing nothing.

I find the NNT to be an interesting index of effect size in that it relates to a practical, real world concept in applied settings.

Effect Size 7: Risk Differences, Relatives Risks and Odds Ratios

When the predictor and outcome are both binary, there are other ways of characterizing effect size than the NNT. These include risk differences, relative risks, and odds ratios. Consider the case where the outcome is whether people vote in favor of policy A. The probability of doing so is 0.60 in the treatment condition and 0.40 in the control condition. These probabilities reflect the proportion of individuals in each condition who exhibit the outcome. A **risk difference** is simply the difference between the two probabilities, which in this case equals 0.20. A second index is to form the ratio of the two probabilities. If the two probabilities are identical, the ratio will equal 1.0; as the value of the ratio deviates from 1.0 in either direction, the effect size is larger. This index is called the **relative risk** and ranges from 0 to infinity. In the above example, the relative risk is 0.60/0.40 = 1.50; the probability of voting for policy A in the treatment condition is 50% larger than in the control condition using the control probability as the base for making the ratio statement. If the probability in the treatment condition was 0.80 and in the control group it was 0.40, the relative risk would be 0.80/0.40 = 2.0; the probability of the outcome is twice as large in the treatment than in the control condition. If the probability of the outcome in the treatment condition is 0.20 and 0.40 in the control condition, the relative risk is 0.20/0.40 = 0.50; the probability of the outcome in the treatment condition is half that for controls.

An **odds ratio** is like a relative risk but it first converts the two probabilities to odds and then forms their ratio. The probability of 0.60 converts to an odds of 1.50 and the probability of 0.40 converts to an odds of 0.67.[6] The ratio of these two odds is 1.50/0.67 = 2.24; the odds of the outcome occurring in the treatment condition is over twice as large as the odds of it occurring in the control condition. Note that for the relative risk, the ratio was more modest, 0.60/0.40 = 1.50. The odds ratio "sounds larger."

All three methods of characterizing group differences are valid; they are just different ways of expressing likelihood or proportion disparities. I personally find the risk difference to be the most intuitive and odds ratios to be the least intuitive, but others might disagree. Some methodologists prefer relative risks to odds ratios and have been critical of what they claim is the misleading nature of odds ratios (e.g., Davies, Crombie, & Tavakoli, 1998; Holcomb, Chaiworapongsa, Luke & Burgdorf, 2001). For all three indices, I think it is important to report the component probabilities to assist interpretation. For example, a relative risk of 2.0 occurs if the probabilities in the treatment and control groups are 0.01 and 0.005, if they are 0.50 and 0.25, or if they are 0.80 and 0.40. An odds ratio of 2.0 results if the probability in the treatment and control

[6] As discussed in Chapter 5, an odds is defined as a probability divided by one minus that probability; 0.60/(1.0-0.40) = 1.50.

groups are 0.01 and 0.005, respectively, or if the probabilities are 0.50 and 0.33, or if the probabilities are 0.80 and 0.67. The probability dynamics for these scenarios seem different to me, so knowing the component probabilities helps. It turns out that if the probabilities tend towards 0, then odds ratios tend to equal relative risks, leading some to conclude that odds ratios are not misleading for rare events.

Suppose I am told that the proportion of Black, 7th grade adolescents in New York City who got drunk in the past 6 months is 0.050 and that the corresponding proportion for Latino 7th grade adolescents is 0.045. This proportion difference, which is only 0.005, seems trivial. If I am told that the likelihood of a certain type of cancer is 6 times larger for older men than for older women, then this seems impressive. However, if the probability for men is 0.006 and for women it is 0.001, the difference in probabilities is only 0.005, which is the same as the ethnic difference in adolescent youth getting drunk. How we frame probability disparities to others matters. And, of course, a 0.005 difference in terms of cancer can be more consequential than an ethnic difference of 0.005 in drinking in youth. For example, for an adult population of 100 million males and 100 million females, a 0.005 sex difference represents 500,000 cancer cases.

There is a practice some researchers use for relative risks and odds ratios that you should be cautious of. If the probability of a successful or positive outcome in the treatment group is 0.60 and in the control group it is 0.40, the relative risk is 0.60/0.40 = 1.50. Some researchers subtract 1.00 from this value and multiply the result by 100 to conclude that people are 50% more likely to have a "success" in the treatment group than the control group. Now suppose I frame the result in terms of failure rates instead of success rates. The failure rate in the treatment condition is 0.40 and in the control condition it is 0.60. The ratio of the two failure rates is 0.40/0.60 = 0.67, so it 1 – 0.67 or 33% less likely that people in the treatment group experience a failure than people in the control group. Note that when the focus is on success rates, I make a different characterization (50% more effective) than when I focus on failure rates (33% fewer failures). Some researchers find this property of relative risks and odds ratios unsatisfactory. Again, knowledge of the component probabilities dilute such framing effects because for either success rates or failure rates, the difference is clear; it is 0.20.

I personally prefer to evaluate group disparities in probabilities or proportions from multiple vantage points while also taking into account the broader substantive context, such as the number of people affected by the occurrence of the event, the degree of severity/positivity of the event, the vulnerability of the target population, and the potential impact of the event on people's quality of life. I personally do not find odds and odds ratios to be as helpful as risk differences and, in fact, in Chapter 12 I describe additional limitations of them. I typically find that when explaining program effects to administrators/staff, they prefer percentages and percent differences (i.e., risk differences) and a variant of them, average marginal effects described in Chapter 5 and elaborated further in Chapter 12. These statistics are more intuitive.

Effect Size Indices for Omnibus Mediation Effects

In addition to the above effect size indices for a given link in a mediational chain, standardized effect size indices of omnibus mediation effects have been developed that reflect in a single number the joint effect of (a) a program on a mediator and (b) the mediator on the outcome (see Preacher & Kelley, 2011; Lachowicz, Preacher & Kelley, 2018). MacKinnon et al. (1995) use the product coefficient method described in previous chapters for a mediator and divide its result by the estimated total effect of the program on the outcome. This yields a proportion of the total effect that the omnibus mediational chain accounts for. If a weight loss program reduces weight, on average by 10 pounds and the mediator of increased exercise accounts for 4 of those pounds, then the effect size for the mediator is 4/10 or 0.40 of the total effect. The closer the value is to 1.00, the stronger the effect of the mediator. Other researchers divide the mediated effect not by the total effect but by the sum of the indirect effects across mediators. If in a study there are three mediators and they account for a total of 8 pounds of the 10 pound program effect, then exercise represents 4/8 or 0.50 of the summed indirect effects.

A problem with both of these indices is that for multiple mediators, if some of the mediated effects are positive and others are negative, the proportions can be larger than 1.0 or negative in value, which is nonsensical. MacKinnon et al. (1995) found these indices of omnibus effect size to be unstable, showing considerable sample-to-sample variability across random samples from the same population. Although popular, the approaches generally are not recommended (see Preacher & Kelly, 2011; Gellman, Hill & Vitari, 2021).

Preacher and Hayes (2008) recommend an index based on the product of coefficient method but using fully standardized coefficients throughout the mediational chain rather than unstandardized coefficients (see also Cheung, 2009). Fairchild et al. (2009), MacKinnon (2008), and Preacher and Kelly (2011) offer interesting omnibus indices but they also suffer from shortcomings and are not recommended for general use (Wen & Fan, 2015; Lachowicz et al., 2018). Lachowicz et al. (2018) recently developed indices that approximate the proportion of explained variance in the outcome due to a given omnibus mediational chain that are promising, but the indices have only been explored in models that are too simple for most RETs. Also, as I discussed for indices of squared correlations, there are no uniform standards to characterize effects based on proportions of explained variance as “small,” “medium,” or “large.” Kraemer (2014) proposed a standardized omnibus index based on Cliff's (1996) δ (not to be confused with Cohen's δ) that also is tied to the probability of exceptions. However, its properties and utility have not been explored in Monte Carlo work (see Lachowicz et al., 2018, for a discussion of this point).

The construction of meaningful standardized omnibus effect size indices for a given mediational chain is challenging when one recognizes how complex these chains can be; they can have correlated disturbances, there can be causal relationships among mediators, there can be reciprocal causation, there can be more than three links in the chain and these links can be measured contemporaneously and/or longitudinally, there can be covariates to control, some of the variables in the chain might be latent variables with multiple indicators, and the variables in the link or that need to be controlled can be nominal, ordinal, interval or ratio scaled or some combination of these psychometric properties. My own preference for RETs is to conduct careful analyses of individual links on a link-by-link basis and to do so from multiple vantage points. Again, I do not want to downplay the possible utility of omnibus chain analysis for some mediational contexts; but the fact is they can be challenging to work with in typical RET contexts and often are not all that informative. I elaborate further on these points in Chapter XX.

Concluding Comments on Effect Size Indices in RETs

In sum, in addition to raw mean differences and unstandardized regression coefficients, there are many standardized effect size indices that represent different ways to gain perspectives on the meaningfulness of an effect for a given link in an RET model. Each index I have described has strengths and weaknesses. I often describe effect sizes with clients from multiple perspectives with the idea that some indices resonate better with some clients than others. Which index seems best to you?

As will become apparent, I generally discourage applying uniform effect size standards that attach specific numerical values to "small," "medium," and "large" effect sizes independent of substantive context because context matters. Examples abound where small effect sizes based on such guidelines have consequential effects and where large effect sizes have minor effects (Kelley & Preacher, 2012). My general recommendation is that unless one index seems particularly appropriate for you in your setting, it probably is best to consider multiple indices of effect size from different vantage points. The idea is to get the best intuitive sense of how strong a link is; coming at it from different vantage points will often be helpful. As I discuss below, in addition to the quantitative indices of effect size, it also is important to take into account the broader context of the research by consulting with experts, practitioners, and clients surrounding their viewpoints on what constitutes meaningful effects. When approached in this fashion, I have found in my own work that the effect size meaningfulness for any given link in an RET usually is evident. You might be frustrated that I do not provide you with a hard and fast rule that you can apply in all contexts, like "A Cohen's d of 0.50 is a medium effect size." I am sorry but the world is not so simple. I will show you in future chapters how I make use of the effect size measures discussed above.

SETTING EFFECT SIZE STANDARDS IN RETs

In this book, I consider program evaluation from two vantage points. One case is when you evaluate a program in a specific setting to provide feedback to staff and administrators about the effectiveness of their program. The second case is that of a scientist seeking to evaluate a program that is to serve as an intervention in applied settings in general. The latter scenario is not as strongly tied to a specific setting because the program is intended to be used in multiple settings and contexts. Research with the program often appeals more to discovering scientific principles. The challenges of setting effect size standards are somewhat different for the two cases.

In an ideal world, the setting of standards for effect meaningfulness should be well-informed by empirics and guided, at least in part, by science. However, too often the scientific community refrains from addressing effect size meaningfulness, relying instead on simple statistical significance (is the p value less than 0.05?) or using arbitrary standards, such as the standards of small, medium, and large effect sizes developed by Cohen (1988). Evaluation research and program development have suffered from this neglect. Although we may be able to get away with such behavior in our scientific journals, when we are in real-world contexts in which clients have hired us to evaluate and improve their programs, clients typically want concrete answers about the meaningfulness of program effects, not double speak about some concept they have never heard of (e.g., Cohen's *d*) coupled with guidelines (e.g., "a *d* of 0.50 means your program is having a 'medium" or "moderate" effect) that not only are meaningless to the client but that well-trained scientists know (or should know) are often arbitrary. In this section of the chapter, I describe approaches I have found helpful for exploring effect size meaningfulness.

Clinical psychology has sought to describe strategies for defining what they call clinically meaningful change when using outcome measures that have arbitrary metrics or metrics that are difficult to interpret. Three approaches are common, (1) expert-based approaches, (2) distribution-based approaches, and (3) anchor-based approaches. Usually, the focus is on defining the minimum effect size that would be judged as being meaningful. An example of an **expert-based approach** is the **Delphi method** (Black et a., 1999). One presents the scale or questionnaire that measures the outcome of interest to a panel of experts in the substantive area being studied. The panel is provided the context and results of the randomized trial and they are asked to provide their best estimate of what a minimal meaningful difference is between the treatment and control groups. Their responses are averaged and the averaged result and other feedback from the expert survey is sent back to the experts, inviting them to revise their estimates if they want. The process is repeated until reasonable consensus results.

Distribution based methods define minimal meaningful change based on features of the response distribution, usually the standard deviation (SD) of the outcome. Mouelhi et al. (2020) found in their review of randomized trials that distributional standards varied considerably for defining meaningful change and included minimal meaningful cutoff scores of 0.50 SDs, 0.33 SDs, 0.30 SDs, and 0.20 SDs. For example, if the SD of weight loss in response to a weight reduction program is 10 pounds, the definition of meaningful change might be 5 pounds or more (0.50 of the SD), 3.3 pounds or more (0.33 of the SD), 3 pounds or more (0.30 of the SD), or 2 pounds or more (0.20 of the SD) depending on the value of the SD and the cutoff standard applied. Sometimes the SD is calculated using the pretreatment variability in the outcome, sometimes it is calculated using the posttreatment variability in the outcome, sometimes it is calculated using the change scores from the pretreatment to posttreatment, and sometimes it is based on regional or national norms. The choice of which SD to use matters and can produce vastly disparate standards for defining minimally meaningful change. Some researchers introduce corrections for measurement error when defining the SD; others work with SD confidence intervals. Also problematic is the fact that SDs often are arbitrary, as discussed earlier, perhaps making them a poor choice for deriving a distributional standard. I am not saying that distributional methods always are flawed. They have their place. However, if you are going to rely on a SD to define a cutoff, you must ensure that the SD you use is a meaningful standard to apply.

Anchor-based methods map scale scores onto anchors and then use these anchors to define meaningful change. One popular approach in clinical trials uses as anchors the verbal descriptors of the clinical global impression-improvement (CGI-I) rating scale. This scale asks patients and/or clinicians to rate patient improvement post-treatment on a 7-point improvement metric, 1 = very much worse, 2 = much worse, 3 = minimally worse, 4 = no change, 5 = minimally improved, 6 = much improved, 7 = very much improved. Mean change scores for the target continuous measure from baseline to posttest are calculated for patients, with the mean value for category 6 then used to define the minimal meaningful change standard. For example, a randomized trial of a program to reduce irritable bowel syndrome (IBS) symptom severity used the IBS-SSS symptom severity scale. This scale yields a composite score of abdominal pain, number of days with abdominal pain, bloating/distension, satisfaction with bowel habits, and IBS-related quality of life. Scores range from 0 to 500. Mean baseline to posttest change scores are calculated for the IBS-SSS and the mean change score in each CGI-I category are estimated. In one study, for category 6 of the CGI-I ("much improved), the mean reduction on the IBS-SSS was 85. The value of 85 is therefore used as the minimal meaningful change standard, defined by the patient-reported verbal anchor of "much improved."

A variant of the verbal-based anchor method is to use external anchors to help define meaningful effects on a target outcome. For example, if a one unit decrease in a depression scale you are using for your primary outcome is associated with only a small decrease in suicide ideation, you may decide a one unit change on the depression scale is not meaningful. However, if a five unit change on the depression scale is associated with decreases in the likelihood of suicide ideation by a non-trivial amount (say by 10%), then such a change in the depression metric might be deemed meaningful (see Jaccard, 2022, for elaboration of this benchmark based approach).

Setting Effect Size Standards for Existing Programs

An important step for evaluating an existing program is to meet with relevant constituencies to discuss with them what constitutes meaningful effects for outcomes and mediators. The constituencies might be program administrators, staff, counselors, clinicians, and/or program participants. I approach such discussions with the goal of generating values for the latitude of meaningfulness, the latitude of no effect, and the latitude of effect ambiguity that I introduced in Chapter 2. Recall that the **latitude of meaningfulness** is the consensual smallest amount of change needed on the outcome to produce a meaningful effect. Any change greater than it in the desired direction is deemed meaningful. The **latitude of no effect** is defined by the maximum change value that still represents a trivial effect. Any change less than it in the desired direction is said to be non-consequential. The **latitude of effect ambiguity** is a "gray area" in which there is disagreement about whether change values within the latitude are meaningful; some members of the team might say "yes" and others might say "no." As an example, after consultation with different constituencies for an intervention to reduce heightened anxiety to normative levels for the general population, it might be agreed that if the program achieves such a result within 6 months for 20% or more patients, then this is meaningful (the cutpoint for the latitude of meaningfulness); if the program achieves the result for less than 5% of patients, then this might be considered as functionally having no effect (the cutpoint for the latitude of no effect); if the program achieves the result for between 5% and 20% of patients, then there might be disagreement about whether this is meaningful (the latitude of effect ambiguity).

To illustrate how I structure discussion, I use a case study of an RET with a continuous outcome and continuous mediators. I discuss setting standards for each facet of the RET, namely (1) the effect of the treatment on the outcome, (2) the effect of the treatment on the mediator, and (3) the effect of the mediator on the outcome. Suppose the program in question seeks to reduce depression and that clinic administrators ask me to use a measure of depression that has an extensive history in their local health system. The measure is a clinician rating of depression based on a clinician interview with the patient, with categories 0 = not depressed, 1 = mild depression, not disabling, 2 = moderate depression, somewhat disabling, 3 = clinically depressed, moderately disabling, 4 = quite depressed, quite disabling, and 5 = extremely depressed, very disabling. Clinicians can assign decimals to their ratings to make more nuanced discriminations. The program targeted three mediators, one of which was coping skills measured on a 1 to 7 disagree-agree metric using multiple items. Items were averaged, with higher scores indicating better skills. A key constituency for this RET is the clinicians who treat patients. I decide to create, among others, a clinician focus-group to discuss effect sizes for the above measures. I organize my discussion around this particular focus group.

The Effect of the Treatment on the Outcome

In the focus group, I might show a line graph for the outcome measure with an arrow demarcating the mean baseline value for patients in the clinic who are treated for depression, per Figure 10.2. I "anchor" the different scale points on the graph by giving the clinicians case summaries of anonymized prototypical patients for each category (e.g., for patients who have a score of "1," for patients who have a score of "2," and so on). After fully exploring the scale with them, I ask the clinicians to discuss how far to the left they think we need to move the arrow, on average, for reasonable clinical progress to have been made as a result of the program. I encourage them to think about change on both the individual patient level as well as the aggregate level for patients as a whole at the clinic. To help them appreciate the meaning of a mean shift of, say, 0.50, I tell them that such a shift would occur if the program shifted 50% of all patients, assuming a baseline score of 3 (clinically depressed, moderately disabling) to a posttreatment score of 2 (moderate depression, somewhat disabling). A mean shift of 0.35 would map onto a scenario of 35% of patients moving from a score of 3 to 2. And so on. The discussion is comprehensive and I make sure clinicians keep in mind the general life circumstances of patients as a whole and how treatment affects their lives more generally. I ask clinicians to identify a cutoff value that represents the smallest amount that we need to move the arrow to the left to produce what the clinicians feel is meaningful change.

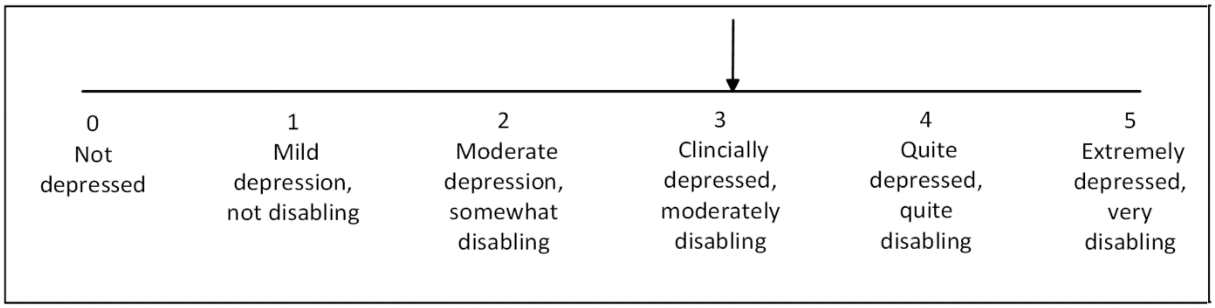
0
Not depressed
1
Mild depression, not disabling
2
Moderate depression, somewhat disabling
3
Clincially depressed, moderately disabling
4
Quite depressed, quite disabling
5
Extremely depressed, very disabling

FIGURE 10.2. Line graph of outcome

Suppose almost all clinicians felt that changes of more than -1.0 are meaningful. Suppose also that most clinicians agreed that changes less than half a unit (-0.50) would not be meaningful. Finally, suppose there was disagreement between the values of -0.50 to -1.00, with some clinicians believing such change is meaningful but others not. Once these values are identified, I can specify, (a) the lower bound for the latitude of meaningfulness (it equals -1.0), (b) the upper bound for the latitude of no effect (it is -0.50) and (c) the latitude of effect ambiguity (-0.50 to -1.00). The overall goal of the discussions is to evolve standards based on input from scientists (myself and my team) who bring past research and psychometric theory to bear, and input from the clinicians and other constituencies who bring to bear context-relevant and practical knowledge. Presenting a reference baseline on the graph is important because sometimes the amount of change deemed meaningful depends on where on the dimension the change is from. For example, a change of -1.0 may not be deemed meaningful from a baseline score of 4 but it might be meaningful from a score of 3. I explore this during my discussions.

Sometimes the measure of the outcome is a multi-item self-report. For example, the classic PHQ-9 is a measure of depression that consists of nine-items. Each item is rated relative to the past two weeks in response to the stem "Over the last 2 weeks, how often have you been bothered by the following," with each item rated on a scale of 0 = not at all, 1 = several days, 2 = more than half the days, and 3 = nearly every day. Responses to items are summed yielding a total score that ranges from 0 to 27. Scores of 5, 10, 15, and 20 on this scale are said by some to represent cut points for mild, moderate, moderately severe, and severe depression, respectively. However, the empirical bases for these cut points is not strong nor is it certain the standards apply to the particular population I might be working with. If I were to use this measure as my primary outcome, I would distribute to focus group participants a copy of the PHQ-9 items and have participants review and discuss them in depth. Instead of working with summed total scores, I might score it using the average item response a person made. I present the line graph for this scoring showing the mean baseline average score of patients in their clinic per Figure 10.3. A score of 2 means the person gave an average response of "more than half the days" to the 9 items. I ensure the focus group participants understand the scoring. I find that structuring the task in this way is more intuitive to participants than using summed total scores, unless the clinicians are already quite familiar with the PHQ-9. I again prepare prototypical summaries of anonymized patients for each major scale point. I then explore how far we need to move the arrow to the left to bring about what the clinicians think is meaningful change.

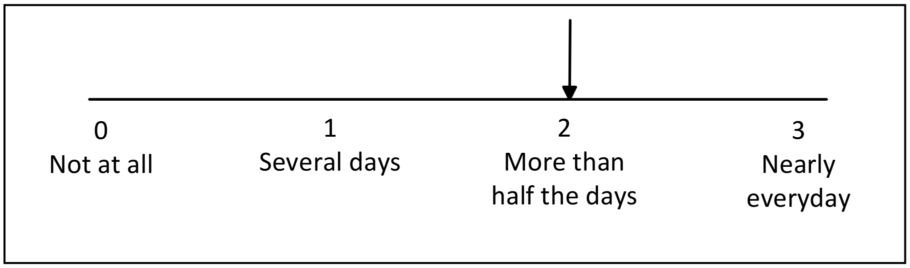

FIGURE 10.3. Line graph for PHQ-9

The above strategy is an imperfect way of seeking to define standards for effect size meaningfulness in a specific substantive context rather than in the abstract. The constituencies participating in the focus groups are reminded to take into account the populations they work with, the contexts in which the populations live, the impacts on the quality of life that accompany outcome changes, and the constraints the program works under. The three latitudes are meant to take into account the degree of consensus that emerges during discussion. I do not trivialize the challenges of using the above strategy to set meaningfulness standards when evaluating a program in applied settings. It can be difficult. However, it is far more informative and useful than simply arbitrarily declaring "a Cohen's *d* of 0.50 defines a meaningful effect."

Once meaningfulness standards for the unstandardized outcome differences between treatment and control groups have been isolated, you can use them in your evaluation study to make meaningfulness judgments for the total effect.

The Effect of the Treatment on a Mediator

For treatment effects on mediators, the focus-group approach is challenging because mediators are mechanisms rather than outcomes and usually are not as familiar to focus group participants. Given this, it can be difficult for focus group participants to make judgments about meaningful change in them. Also, meaningful change in a mediator is, in part, a function of how much the mediator impacts the outcome. In the depression example where change of -1.0 or more depression units is considered meaningful for the outcome, how much change in a mediator is necessary to bring about that much outcome change? The answer to this question helps define the standard for meaningful mediator change for the program being evaluated. The answer can be informed, in part, by examining results from the data collected during the formal RET or in pilot research. For example, suppose that the path coefficient for the effect of coping skills on depression was -0.50. This means that for every one unit that coping skills increases, the depression outcome is predicted to decrease by -0.50 units. It therefore takes a two unit change in coping skills to produce a meaningful change in the outcome of -1.0 units because (2)(-.50) = -1.0. A treatment effect that produces a 2.0 unit change in the mediator, in some ways, defines the lower bound meaningfulness standard for coping skills as impacted by the treatment condition.

One problem with the above logic is that sometimes change in a single mediator is not enough to bring about meaningful outcome change on its own, but small changes in several mediators will do so when considered collectively. This might lead me to set the meaningfulness standard to some fraction lower than a 2.0 unit change for coping skills, recognizing that the mediator only needs to accomplish 'part of the job' rather than the 'full job' of producing meaningful outcome change; the other mediators should finish the job started by the target mediator. You need to determine what you think is a reasonable fraction to use given the broader RET logic model and context. In the depression example, there are three mediators, one of which is coping skills. If I set the fraction to 1/3, I am asking each mediator to carry an equal and shared load in bringing about meaningful change of at least -1 in the outcome.

In my experience, it is easier to specify a meaningfulness standard for the effect of the treatment on the mediator once I have a reasonable sense of or I make a working assumption about the magnitude of the effect of the mediator on the outcome. Rather than make wild guesses about the latter, I find it useful to examine the RET data to gain a sense of the magnitude of the coefficient based on empirics. This post hoc approach, which I formalize shortly, is counter to philosophies advocated by some scientists who feel that such criteria should be specified a priori when clinical trials are formally registered with oversite organizations (van t'Veer & Giner-Sorolla, 2016; Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014). The idea is to protect against, among other things, p-hacking and HARKing (hypothesizing after results are known; see Jaccard & Jacoby, 2020). I am sympathetic to these goals, but the fact is that reasonable statistical practice and scientific inference sometimes require initial data exploration, such as when the choice of an analytic method depends on the nature of non-normality or nonlinearities in the data. There must be a balance between analysis pre-specification and needed analytic flexibility for optimizing scientific and statistical practice. Having said that, I do think it is important to a priori explain and justify the contingencies you will use in your analyses and what those contingencies entail.

To summarize, to make a judgment about the meaningfulness standard for the effect of the treatment on the mediator, T→M, you need to

1. Make a working assumption about what constitutes a meaningful effect of the treatment on the outcome; in the depression example, it was -1.

2. Make a working assumption about the likely true value of the path coefficient linking the mediator to the outcome, namely M→O; in the depression example, it was -0.50. This assumption can be informed either by prior research, common sense, expert opinion, or even by the results from your RET data.

3. Decide the fraction of the total program effect on the outcome that you want each mediator to account for; in the depression example, it was 0.33 but you also can assign different values to different mediators, as appropriate.

4. Then, given these working assumptions, calculate the meaningfulness standard for the effect of the treatment on the mediator using the following formula:

Standard for treatment effect on mediator = $[(F)(MCO)] / p_{M \to O}$

where F is the value of the fraction, MCO is the meaningful change for the outcome, and $p_{M \to O}$ is the path coefficient from the mediator to the outcome. For the depression example, it is

Standard for treatment effect on mediator = $[(.33)(-1)] / -0.50 = 0.66$

Thus, a true mean difference on coping skills between the treatment and control groups of 0.66 or more will contribute sufficiently to meaningful change on the outcome, *given the viability of the working assumptions and the appropriateness of the fractional assignment of the total outcome effect to the mediator*. On my website, I provide a program called *effect size standards* to apply the formula. The video associated with the program provides an example and I make use of the program in my numerical examples in future chapters. I encourage you to watch the video to help you make decisions about meaningfulness of treatment effects on mediators.

Note that this approach allows the meaningfulness standard to be tailored to the specific program evaluation context in which your RET is conducted. Some scientists might view this as a disadvantage from the standpoint of knowledge accumulation in the scientific literature, arguing that meaningfulness standards should not be nuanced by context. Your clients who hire you to do the program evaluation, however, will appreciate your sensitivity to their context and, indeed, it can be argued that meaningfulness standards more generally should be context dependent. The argument is that one size does not fit all and that scientist's insistence on common standards is naïve.

The Effect of the Mediator on the Outcome

Once a standard for judging meaningful change on the outcome has been determined, it also can be used in conjunction with the RET data to form effect size standards for mediator effects on the outcome. To do so, I use the same logic as for the T→M link, but now I apply it to the M→O link.

To outline the logic, suppose I find in the data that the program raises the coping skill mean in the treatment group by 1.2 units relative to the control group on the 1-7 coping skill metric. If I use this to represent the change I can likely bring about in the mediator, I can ask what the value of the M→O coefficient would need to be to produce a mean change of at least -1.0 on the depression measure, i.e., the outcome meaningfulness standard. The answer is -1/1.2 or -0.83. Thus, *if* my program raises coping skills, on average, by 1.2 units and *if* the M→O coefficient is -0.83, then the overall mean shift in depression between the treatment and control conditions will equal -1.0, which is meaningful. I therefore might choose -0.83 as my effect size standard for the M→O coefficient.

There are complications to this logic. First, perhaps my program did not bring about much change in coping skills but I believe that I can modify the program to do a better job. Instead of setting the meaningfulness standard for the M→O coefficient based on the change my program actually produced (which might underestimate the amount of change that can be achieved), I might instead specify the **plausible change** in coping skills that I think I can bring about. Suppose that after much thought, I decide that the plausible change in coping skills I can probably bring about is closer to 2.0 coping skill units, give or take. Now the meaningfulness standard for the M→O coefficient would be -1/2.0, which equals -0.50 instead of -0.83. Second, as before, instead of demanding that the mediator carry all the responsibility for producing meaningful change in the outcome, I can assign a fraction of the outcome change, F, that I want it to fulfill.

Here are the key action steps you need to enact to form the M→O coefficient meaningfulness standard:

1. Make a working assumption about what constitutes a meaningful effect of the treatment on the outcome; in the depression example, it was -1.

2. Make a working assumption about the true value of the path coefficient linking the treatment dummy variable to the mediator, i.e., the mean change that your program likely will produce or that it can plausibly produce. In the depression example, it was -2.0. This assumption can be informed either by prior research, common sense, or by using results from your RET data.

3. Decide the fraction of the total program effect on the outcome that you want each mediator to account for; in the depression example, it was 0.33.

4. Then, given these working assumptions, calculate the effect size standard for the effect of the mediator on the outcome using the following formula:

Standard for mediator effect on outcome = [(F)(MCO)] / PC

where PC is plausible change in the mediator. For the depression example, it is

Standard for mediator effect on outcome = [(.33)(-1.0)] / -2.0 = 0.165

You can experiment with different values of F, MCO, and PC as you seek an effect size standard to use. The program on my website called *effect size standards* also implements this formulation. The core idea is that we are able to quantify how strong the M→O link needs to be by making working assumptions about (a) the minimal meaningful program effect on the outcome, and (b) the plausible change that we can bring about in the mediator.

Commentary on Establishing Standards for Effect Size Meaningfulness

Ultimately, the effect size standards you set will be a matter of judgment that you need to defend if challenged. The framework I suggest is far more nuanced than abstractly stating a standard like "I want coping skills to account for at least 3% of the variance in depression" or "I want the program to have an effect on the mediator equal to an absolute *d* of 0.50 or greater." Preacher and Kelly (2011) note that an advantage of standardized effect sizes is that it "frees the researcher from having to prepare a new set of interpretive benchmarks for every new scale or application" (p. 95). This statement also describes a disadvantage of standardized effect sizes; they encourage not thinking about context.

One concern my colleagues express about my approach is that the resulting standards are tied to the specific population, program, and organization in which the RET is conducted. As such, the standards can lack generality. When I am hired by a client to evaluate a specific program in a specific context, I would argue that this narrow focus is entirely appropriate. It is only if I am interested in advancing science or policy more generally that the issue of meaningfulness generalizability arises.

Contrary to blindly adopting arbitrary standardized effect size benchmarks that ignore context, I encourage you to consider context and partner with relevant constituencies to determine appropriate effect size standards. The way you approach such discussions and ultimately settle on standards might differ from program to program or it may depend on the substantive application. The tools I suggest for setting effect size standards are best informed by educated guesses of true program effects on mediators and mediator effects on outcomes. If you are philosophically opposed to relying on such information post hoc within your RET, then you can obtain estimates of the parameters in pilot research. However, pilot research usually is not feasible in contracted program evaluations.

Setting Effect Size Standards for Programs in the Abstract

As noted, another scenario that occurs is when a scientist seeks to evaluate a program that serves as a model for use in multiple applied settings. In such cases, I think it makes sense to identify prototypical settings in which the program is expected to be implemented and then to conduct focus groups with different constituencies to explore effect size standards for those settings using measures that ultimately will be recommended for use with the intervention.

Scenarios can arise, of course, where explorations of meaningful change standards in prototypical settings is not possible. If so, is it possible to state in the abstract what is a meaningful amount of change to expect for a program? This is exactly the question that Cohen attempted to affirmatively answer when he suggested his effect size standards. In a similar vein, Smith and Glass (1977) conducted a meta-analysis of 400 studies from the psychological literature on the effectiveness of psychotherapy and found an average Cohen's *d* of 0.68. Smith and Glass then suggested that one might expect therapeutic effects for outcome mean differences equivalent to about 2/3 of an outcome standard deviation and that we use this as a standard for judging meaningfulness. Eysenck (1978) referred to the Smith and Glass meta-analysis as an "exercise in meta-silliness," objecting to aggregating so many diverse outcomes and diverse therapeutic methods into a single numerical summary that ignores context. The key point here is that wherever possible make an effort to contextualize standards even if you are developing an abstract program. If you are conducting an RET to evaluate a program to promote exercise by the elderly and you are unable to do the type of contextualizing research I recommend, then at least base your effect size standards on prior research on exercise by the elderly, paying particular attention to studies that map onto the contexts you envision your program applying to.

EFFECT SIZE INTERPRETATION AND SAMPLING ERROR

In this final section, I describe an approach for interpreting effect sizes that takes sampling error into account. Recall that the parameter of interest in the depression example was the population difference between the mean depression for the treatment and control conditions using a 0 to 5 clinician rating of patient depression. The null and alternative hypotheses in this case are:

H_0: μ_T - $\mu_C = 0$

H_1: μ_T - $\mu_C \neq 0$

I can draw a line graph to depict the possible population values of the mean difference that might operate per Figure 10.4a. Recall that the latitude of no effect was specified to be -0.5 to 0.5. The two dashed lines designate the boundaries of this latitude. If the population mean difference occurs anywhere in this latitude, the means for the two groups are deemed *functionally equivalent*. Suppose the 95% confidence interval for the sample mean difference is -0.25 to +0.25. If the lower and upper limits of this interval are completely contained within the latitude of no effect, per Figure 10.4b, then I can be 95% confident that the population means are indeed functionally equivalent.

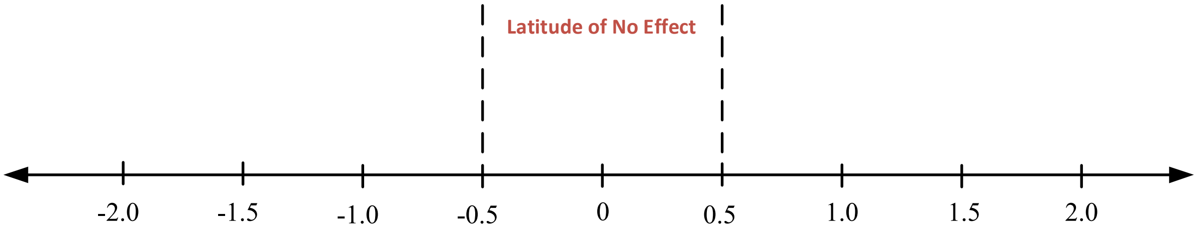

FIGURE 10.4a. Line graph showing latitude of no effect

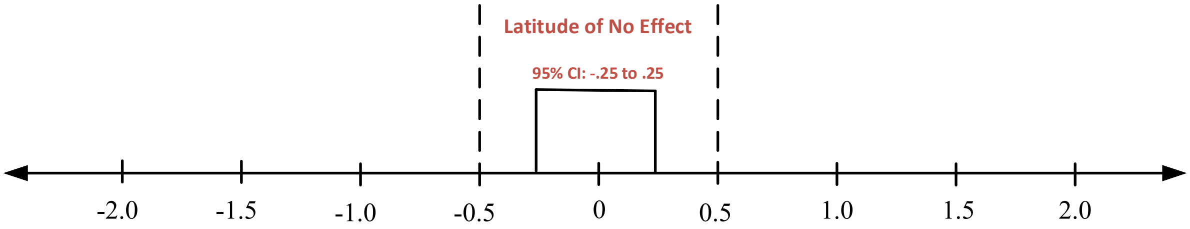

FIGURE 10.4b. Line graph showing latitude of no effect with 95% CI

Figure 10.5a shows the same line graph but now with the latitude of meaningfulness highlighted; the desired difference between the treatment and control group means is a negative number less than or equal to -1.00. Suppose the 95% confidence interval for the sample mean difference was -1.75 to -1.25, per Figure 10.5b. In this case, the sample confidence interval is fully contained within the latitude for meaningful change, so it is reasonable to conclude that a meaningful effect exists in the population.

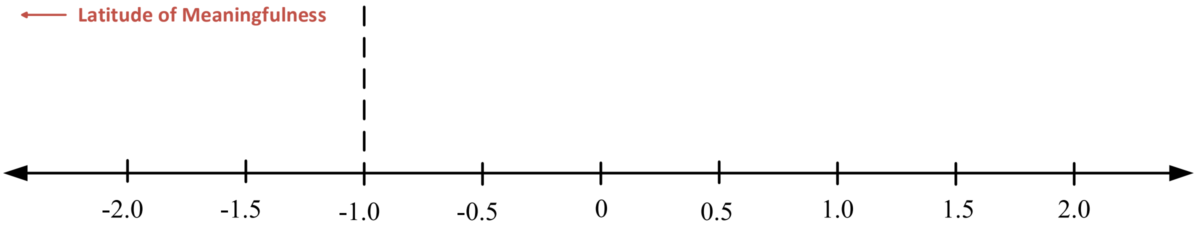

FIGURE 10.5a. Line graph showing latitude of meaningfulness

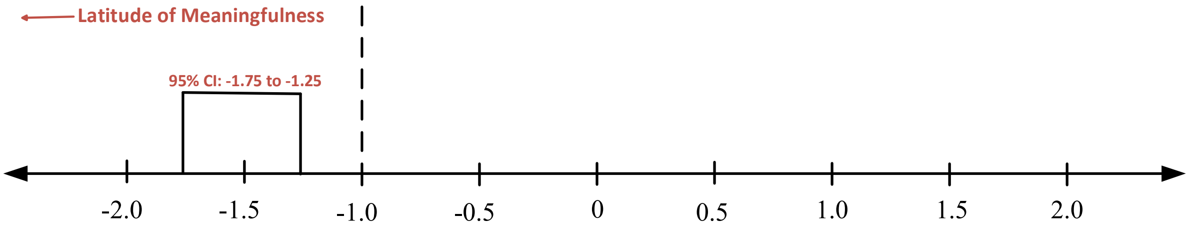

FIGURE 10.5b. Line graph showing latitude of meaningfulness with 95% CI

Finally, suppose the 95% confidence interval for the group difference overlaps two latitudes, one of which is the latitude of meaningfulness. For example, if the confidence interval for the mean difference is -1.25 to -0.55, then there is overlap between the latitude of meaningfulness (≤ -1) and the latitude of effect ambiguity (>-1.0 to <-0.5). In this case, I cannot confidently conclude there is a meaningful population difference between the means because some of the confidence interval is in the latitude of effect ambiguity. The meaningfulness of the effect is suggestive, but not conclusive.

Basically, the approach I recommend is to calculate the confidence interval for the parameter of interest in your RET and then map that confidence interval onto the latitudes of meaningfulness, effect ambiguity, and no effect. If the confidence interval is fully within the latitude of meaningfulness or fully within the latitude of no effect, the conclusion is straightforward. If the interval overlaps two latitudes, the result is less straightforward. Note that this approach does not require interpreting p values, although one can do so if one desires; conclusions are completely effect size driven. The approach takes into account sampling error in the effect size estimates. Note also that the larger the confidence interval, the more likely it is one will obtain inconclusive results because wide intervals are likely to produce overlap between latitudes. Smaller sample sizes are likely to lead to inconclusive results, which makes intuitive sense. In my approach, instead of choosing sample sizes to increase statistical power, I choose sample sizes to produce narrow confidence intervals. I discuss sample size selection for achieving narrow confidence intervals in Chapter 28.

Interestingly, if one relies exclusively on statistically significant p values to make statements of meaningfulness (e.g., $p < 0.05$ is a meaningful effect, $p > 0.05$ is not), then this is tantamount to defining a standard for a meaningful effect as any population difference that is not zero. This is because the p value tests the null hypothesis of a *zero mean difference*. Perhaps a meaningfulness standard of anything-but-zero is justifiable in some contexts. However, if an anything-but-zero value is chosen as the standard, I believe that researchers need to make a case for its choice. It does not seem very defensible to me in most applied settings.

It is instructive to contrast this approach with more typical approaches used in published research. When reviewing articles from clinical psychology for possible use as examples, I found that none of the articles addressed the topic of effect size meaningfulness. Instead, the articles reported significance tests for treatment versus control group differences and presented Cohen's *d* statistics for effect size, leaving it to readers to form their own judgments about meaningfulness. In the Discussion sections, nothing was said about the magnitude of the *d*s and discussion centered almost exclusively on effects that were statistically significant ($p < 0.05$). Thus, although effect sizes were reported, discussion revolved around the significance tests. To me, the researchers were not taking matters of effect size seriously enough.

I isolated studies of treatment programs that used the classic CES-D scale for depression as an outcome. As noted, the metric of the CES-D ranges from 0 to 60, but the vast majority of people score less than 30 on it, giving it a functional metric of about 0 to 30. Several psychometric articles discuss what constitutes meaningful clinical change on the CES-D, with minimal meaningful change typically being defined as somewhere between 5 and 10 scale points. For purposes of illustration, I will define the latitude of no effect as change less than an absolute value of 5, the latitude of effect ambiguity as absolute changes between 5 and 8, and the latitude of meaningfulness as absolute changes larger than 8.

In one study, the sample size was 19 per group and the mean in the control group was 12.89 (SD = 4.78) and in the treatment group it was 5.68 (SD = 4.74). The reduction in depression was -7.21, which was statistically significant (t(36) = 4.67, p < 0.05 *d* = -1.51). The 95% confidence interval for the mean reduction was -10.34 to -4.08. The confidence interval overlaps the latitude of meaningfulness (-8 or less) and the latitude of no effect (-5 to +5), so we cannot conclude with confidence that the treatment produced a meaningful effect, despite its statistically significant p value; there is too much sampling error in this study to draw a firm conclusion, likely because of the small N. With the sample of 19 per group and the results that were observed, I would not agree with the authors that the treatment effects were meaningful.

In another study, the sample size was 90 per group and the mean in the control group was 9.64 (SD = 8.23) and in the treatment group it was 6.80 (SD = 5.77). The reduction in depression was -2.84, which was statistically significant (t(178) = 2.68, p < 0.05 *d* = -0.40). The 95% confidence interval for the mean reduction was -4.93 to -0.75. The confidence interval is completely contained within the latitude of no effect (-5 to +5), so I would declare the population group means as functionally equivalent. By contrast, the authors interpreted the results as supporting the meaningfulness of the treatment because they relied exclusively on the p value for their conclusion.

The framework I propose shifts the focus of analysis from one of testing group differences against a null hypothesis of zero to one of testing group differences against a null hypothesis of the minimum value of meaningful change. This can be more sample size demanding, but it is appropriate if one wants to know if meaningful change has been brought about by a program after taking into account sampling error. Perhaps a case can be made for using a lower percent confidence/credibility interval than 95% given the shift in focus from a null hypothesis of a zero difference, but this will depend on the substantive context of the study. For an extension of this approach to omnibus mediation analysis rather than effect sizes for a link within a mediational chain, see Beribisky et al., (2020).

A somewhat frustrating result for program evaluators will be situations where a "mixed message" results relative to the sample mean difference, the estimated confidence interval for the mean difference, and the meaningfulness standard. Consider a study where I evaluate an intervention to increase moderate to vigorous physical activity (MVPA) in adolescents and where the minimum meaningfulness standard for the program is to increase MVPA by an average of 30 minutes per week. Suppose the treatment versus control posttest mean difference is 37 minutes and that this is statistically significant ($p < 0.05$) by traditional null hypothesis testing standards with a margin of error of ±8 minutes. The 95% confidence interval is 29 minutes to 45 minutes. Because the meaningfulness standard of 30 overlaps the lower limit of this interval, I cannot conclude with 95% confidence that the intervention produced a meaningful effect on MVPA. To be sure, I can confidently conclude that the intervention effect is non-zero because the null hypothesis of no effect was rejected. And, if you ask me my single best guess of the true mean difference between the treatment and control conditions, it would be the sample mean of 37 minutes, which also exceeds the meaningfulness standard. Despite these facts, I can't say that I am 95% confident the intervention produced a meaningful result due to the overlap of the lower limit and the meaningfulness standard.

If I instead calculate a 90% confidence interval for the above, the margin of error becomes ±6.60 and the confidence interval is 30.4 minutes to 43.6 minutes. I can now state with 90% confidence that the intervention produces a meaningful effect.[7] This suggests a possible strategy of reporting to clients the sample mean difference that represents your best guess of the program effect but then also convey the confidence level you have that the meaningfulness standard is exceeded in the population given the operative sampling error based on your exploration of different confidence intervals. In the current case, I would say that my best guess of the intervention effect is to raise MVPA on average by 37 minutes and that I am 90% confident that the program produces a meaningful effect when one takes into account the operative sampling error.

CONCLUDING COMMENTS

A central task of RETs is to determine if (1) the program effect on the outcome is meaningful, (2) if the program effects on mediators are meaningful, and (3) if the mediator effects on the outcome are meaningful. Many social scientists rely on standardized effect size indices coupled with guidelines for labeling effects as "small," "medium," and "large" to make meaningfulness judgments. The guidelines vary and the rationales for them are either underdeveloped or fairly arbitrary. Importantly, effect size does not equate with effect meaningfulness. Small effects can have major consequences and large effects can have trivial consequences. Judgments of meaningfulness require we broaden our focus to account for context. Meaningfulness judgments ultimately are derived from a collaborative process that involves exchanges between scientists,

[7] As discussed in Chapter 6, some methodologists prefer the use of Bayesian credible intervals when MOEs are used in this fashion, but a case also can be made for using traditional confidence intervals, especially when the Bayesian analysis relies on uninformative priors.

practitioners, and program participants, among others.

My own preference is to use unstandardized indices when documenting effect size. Such indices generally avoid many of the artifacts that afflict standardized indices, such as range restriction and distortion due to heavy tailed distributions. Researchers complain that unstandardized metrics can be arbitrary and lack inherent meaning. My response is that we should make our metrics non-arbitrary rather than mindlessly shift to standardized metrics with arbitrary standards (see Jaccard, 2022, for how to do so). Among the more intuitive standardized indices, in my opinion, are the probability of exception to the rules, the number needed to treat, and risk (percentage) differences and relative risks.

When reporting and interpreting effect sizes, it is useful to include margins of error or confidence intervals for those effect sizes so that readers can appreciate the sample-to-sample fluctuations of them. I make formal use of such information by integrating effect size estimates and their confidence intervals into an interpretational framework that uses a latitude of meaningfulness, a latitude of effect ambiguity, and a latitude of no effect. The framework leads to conclusions that do not rely on p values, although p values can be used to assist conclusions, if desired. Sample size decisions in this framework are driven more by the minimization of confidence interval width rather than statistical power.

Some readers may be disappointed that I have not told them which effect size indices to use and what standards to apply to those indices. Unfortunately, the task is not that simple. I have made known my own preferences, namely (a) to work with unstandardized metrics and to try to make them non-arbitrary if they are arbitrary vis-à-vis discussions with relevant constituencies, and (b) for standardized indices, to rely on multiple indices to provide different vantage points on effect size, with my own preferences leaning towards the probability of exceptions to the rule, the number needed to treat, and risk differences or relative risks coupled with reports of their component parts. I have shown that once you settle upon a definition of a meaningful effect for the outcome, it is possible to derive from this, coupled with RET linear equations, standards for meaningful effects for mediator influences on outcomes and for treatment effects on mediators. I also highlighted issues that you should consider as you form meaningfulness standards in a given RET, taking into account such factors as the number of people affected, the impact on the quality of their lives, the severity and reversibility (or positiveness and sustainability) of the outcome for people, the vulnerability of the affected population, and the costs and organizational readiness to bring about change, among others. I also showed you how to take into account sampling error using the latitude of meaningful effects, latitude of no effects, and latitude of effect ambiguity.

Finally, I provided you with strategies that might help you form effect size standards through focus groups and qualitative work with relevant constituencies. The above mindset is a far richer approach than reporting, say, a Cohen's *d* and letting it speak for itself, without commentary (see the opening quote to this chapter). I apply my approach throughout the remaining chapters of this book.

APPENDIX: CALCULATION OF EFFECT SIZES

This appendix provides formulae for the calculation of different effect size indices. The formulae apply to OLS regression, but often can be used with output from Mplus for RET analysis (see Chapter 11 for details).

Squared Semi-part Correlation

To calculate a squared semi-part correlation for a given predictor in a regression equation with multiple predictors use the following equation:

$$\mathrm{sr}^2 = (\mathrm{CR}^2\ (1-\mathrm{R}^2))\,/\,(\mathrm{N}\text{-}\mathrm{k}\text{-}1) \qquad [10.A.1]$$

where sr^2 is the squared semi-part correlation for the predictor, CR is the critical ratio for the predictor coefficient, R^2 is the overall squared multiple correlation, N is the sample size, and k is the number of predictors (for a z test, some statisticians use N in place of N-k-1; with N > 100, the difference in results usually is trivial). As an example, suppose the outcome is a continuous measure, Y, and the target predictor, X1, is a continuous measure. There are three covariates, so the number of predictors is 4. Suppose the squared correlation for the four predictors is 0.23 and the critical ratio (t ratio) for X1 is 2.31, with N = 200. Using Equation 10.A.1, the squared semi-part correlation for X1 is $((2.31^2)(1\text{-}.23))/(200\text{-}4\text{-}1) = 0.02$, indicating X1 uniquely explains 2% of the variation in smoking behavior.

Partial Correlation

To calculate the partial correlation for a given predictor in an equation with multiple predictors, use the following formula:

$$\mathrm{pr} = \mathrm{sr}\,/\,\sqrt{1-(\mathrm{R}^2-\mathrm{sr}^2)} \qquad [10.A.2]$$

where sr is the square root of sr^2 from Equation 10.A.1, but signed (positive or negative) in the same direction as the coefficient for the predictor. As an example, suppose that X1 in a multi-predictor equation has a semi-part correlation of 0.30, a squared semi-part correlation of 0.09 and that $\mathrm{R}^2 = 0.40$. The partial correlation:

$$\mathrm{pr} = \mathrm{sr}\,/\,\sqrt{1-(\mathrm{R}^2-\mathrm{sr}^2)} = 0.30\,/\,\sqrt{1 - (0.40 - 0.09)} = 0.36$$

Cohen's d for Independent Groups

To calculate Cohen's d, calculate the difference between the two means in question and then divide this difference by either (a) the pooled standard deviation for the two groups, (b) the standard deviation for the control group, or (c) a covariate adjusted standard deviation.

A covariate adjusted Cohen's *d* index can be calculated from Mplus output if one predicts a continuous outcome/mediator from a binary predictor plus the covariates by dividing the path coefficient for the binary predictor (assuming the use of 0-1 dummy coding) by the square root of the unstandardized residual for the outcome/mediator under the output label `Residual Variances` in the section on `Model Results`.

Probability of Exception for Two Groups Based on Cohen's d

To calculate the probability of exception for a binary predictor of a continuous outcome from Cohen's *d* use the equation

$$P_E = 1 - \Phi\,(d_A / \sqrt{2}) \qquad [10.A.3]$$

where d_A is the absolute value of the *d* statistic and Φ is the cumulative standard normal distribution function. In words, calculate an intermediate value, z, as d_A divided by the square root of 2. Then translate z into a probability value, p, in the cumulative standard normal distribution. This is accomplished using tables in statistical texts or on a web calculator. As examples, a z of 0 translates into a probability of 0.50, a z of 0.33 translates into a probability of 0.63, a z of 1.00 translates into a probability of 0.84, and a z of 1.65 translates into a probability of 0.95. P_E is one minus this p value. For example, if Cohen's $d = 2.10$, then P_E is

$$P_E = 1 - \Phi\,(2.10/\sqrt{2}) = 1 - \Phi\,(1.484) = 1 - 0.931 = 0.07$$

Formula 10.A.3 assumes there are equal or roughly equal sample sizes in the two groups.

Probability of Exception for Predictor in Multiple Regression

To calculate P_E for two continuous variables with no covariates (i.e., for bivariate regression), use a Pearson correlation coefficient in the following equation:

$$P_E = 1 - [(\arcsin(r)/\,\pi) + 0.5] \qquad [10.A.4]$$

where r is the absolute value of the observed correlation, arcsin is an arcsine function and π is the mathematical constant pi (which is 3.14159). For example, if X and Y are correlated 0.35, then P_E is

$$P_E = 1 - [(\arcsin(.35)/ 3.14159) + 0.5] = 1 - (.11164 + 0.5) = 0.39$$

When the predictor of concern, X1, is one of many predictors in a linear equation, use the program on my website to calculate the partial correlation for the quantitative predictor of interest and then use the absolute value of it in Equation 10.A.4 to calculate P_E.

If the predictor in the multiple regression is a dummy variable that reflects a contrast between two groups, use

$$P_E = 1 - \Phi\,(b / SD_e) \qquad [10.A.5]$$

where b is the absolute value of the regression coefficient for the predictor and SDe is

$$SD_e = \sqrt{2(1 - R^2)(var(Y)}$$

where R^2 is the squared multiple correlation for the equation and var(y) is the variance (or squared standard deviation) of the outcome variable, Y. The derivation of Equations 10.A.4 and 10.A.5 and their robustness to violations of normality are described in Krasikova, Le and Bachura (2018).

Semi-Parametric Probability of Exception

On my website, I provide two programs for calculating the probability of exception using approaches based on summary statistics typically reported in journal articles, one program for a binary predictor and one program for a continuous predictor, each with the ability to control for covariates. These methods make parametric assumptions and yield point estimates of P_E but not confidence intervals or margins of error. They also can be applied to output from SEM software like Mplus.

I also provide two semi-parametric methods for when you have access to raw data, both of which allow for covariate control and both of which use percentile bootstrapping to generate confidence intervals. One strategy relies on the work of Vargha and Delaney (2000) using what they call an A statistic for the case of a binary predictor. It accounts for discrete data and ties, whereas the other methods assume continuous variables. A is defined as

$$A = [\#(X > Y) + .5\#(X=y)]/n_X n_Y$$

where # is the count function, X and Y are vectors of scores for the two groups, respectively, and n is the respective group sample sizes. For the case of covariates, the program generates a set of score residuals for each group after predicting the outcome from the covariates using OLS regression. It then applies the above equation to these residuals. Bootstrap replicates are generated from the original sample and includes bootstrap of the covariate regression equation. The program for continuous predictors uses the same principle but defines the probability of superiority based on Li (2018) by summing the following across individuals:

$$\#[\text{sign}(X_i - M_X)\cdot\text{sign}(Y_i - M_Y) > 0] + 0.5\#[\text{sign}(X_i - M_X\ ) = \text{sign}\ (Y_i - M_Y) = 0])/N$$

where N is the number of paired X-Y observations, # is the count function, MX is the mean of X, MY is the mean of Y, sign is a sign function. Conceptually, the equation counts the number of times when an X score is above (or below) the mean of X is paired with a Y score that is also above (or below) the mean of Y. If both the X and Y scores are identical to their corresponding means, then a count of 0.50 is used.

Number Needed to Treat

The NNT is the number of people we need to treat by having them participate in the program in order to have one more "success" than if we just left people alone per the control condition. If the formula for the NNT when "successes" are represented by a smaller proportion (e.g., the proportion of alcoholics who relapsed)

$$NNT = 1 / (P_C - P_T) \qquad [10.A.6]$$

and when "successes" are represented by a larger proportion, it is:

$$NNT = 1 / (P_T - P_C) \qquad [10.A.7]$$

where P_C is the proportion of people in the control condition exhibiting "success" and where P_T is the proportion of people in the treatment condition exhibiting "success."

Strategies have been developed to extend the concept of NNT to continuous outcomes but these ultimately involve some form of dichotomization of Y, which I find questionable unless one has a strong theoretical rationale for it. One can compute the NNT directly from data by applying the threshold to the continuous variable to dichotomize it and then using Equation 10.A.6 or 10.A.7. If one does not have access to the raw data, one can estimate the NNT using an approach by Furukawa and Leucht (2011), not the one by Kraemer and Gibbons (2009). The formula is

$$\mathrm{NNT} = 1/(\Phi\,(\mathrm{d} - \Psi(1 - \mathrm{CER})) - \mathrm{CER}) \qquad [10.A.8]$$

where Φ is the cumulative standard normal distribution function as described for formula 10.A.4, Ψ is the inverse function of the cumulative standard normal distribution, d is Cohen's d for the outcome, and CER is control group event rate, i.e., the proportion of cases in the control condition that have an event score of "1" for the dichotomized outcome measure. The dichotomization occurs at an a priori determined cutoff value by the investigator. For example, if Cohen's d is 0.34 and CER is 0.42, the NNT is

$$1/(\Phi\,(0.34 - \Psi(0.58)) - 0.42) = 1/(\Phi\,(0.34 - 0.202) - 0.42) = 1/(0.555 - 0.420) = 7.41$$