1

Methodological Rules

As noted in Chapter 15, Karl Popper espoused the idea of specifying a set of methodological rules or scientific standards that researchers should adhere to in order to promote rigorous theory tests and to shield such tests from critics who seek to maintain a theory in the face of disconfirming data. The idea is to rule out, as much as possible, alternative explanations, both theoretical and methodological, for the results so that critics cannot question the theory test. Ideally, a negative result should unambiguously disconfirm a theory and a positive result should increase confidence in the theory. From the standpoint of Bayesian epistemology, this corresponds to designing a study whose Bayes factor is diagnostic rather than uninformative (see Chapter 15). In this primer, we present a list of 52 methodological and substantive artifacts/difficulties that one should seek to address when designing a study to strengthen theory tests.

The list, by necessity, is incomplete and not all issues will apply to your research. Social science is diverse in methodology and focus. Nevertheless, familiarity with all of the artifacts will serve you well no matter the type of research you conduct. We discuss some issues in more depth than others, based on our judgment of typical knowledge levels students have relative to them. Our intent is to provide a "checklist" for you to use when designing your research and that you also can use to critique research studies you read. Methodologists have evolved solutions or partial solutions for most everything we list. We do not consider the solutions because doing so would require a book length document. However, we encourage you to explore them in your readings and build a strong methodological toolbox to accompany your theoretical toolbox.

We begin with a discussion of the classic threats to internal validity described by Campbell and Stanley (1963), followed by a discussion of issues related to sampling, measurement, research design and statistical analysis. There will be overlap between some of these issues because methodologists working in different areas often have their own unique jargon and ways of thinking about similar phenomena. At the conclusion of this document, we provide a full listing of the issues in checklist form.

CAMPBELL AND STANLEY (1963) THREATS TO INTERNAL VALIDITY

Campbell and Stanley (1963) posited several factors that can undermine valid inference in experiments that involve some type of manipulation or intervention on the part of an

investigator. They use what they call a *single group pretest-posttest design* to illustrate most of them, and signify this design as O_1 X O_2, where the O represents the point in time when an observation or measure is taken and X is the administration of an intervention or experimental manipulation. Two other types of designs they discuss and that we will make reference to are the *two group pretest-posttest design*, which appears as follows:

Experimental Group:	O_1 X	O_2
Control Group:	O_1	O_2

In this design, a control group is added that is not exposed to the intervention. The other design is a *two group posttest only* design:

Experimental Group:	X	O
Control Group:		O

which is the same as the two group pretest-posttest design but omits the baseline measures. In true experiments, assignment of people to conditions is random, but scenarios occur where this is not the case. We will assume random assignment in our discussion below.

Many of the issues raised by Campbell and Stanley are discussed in the context of the single group pretest-posttest design and we will introduce many of the issues in this context. We should note that the issues also apply to longitudinal designs with no intervention but where the researcher seeks to make statements about the sources of changes in a construct over time after measuring the construct on multiple occasions. We consider eight issues described by Campbell and Stanley. As a mnemonic, taking the first letter of each label we use in the order we consider them spells the acronym THIS MESS.

Issue 1: Testing Effects

Testing effects occur when the act of completing a measure changes the construct being measured. Suppose in the single group pretest-posttest design, people with depression complete an assessment battery asking about depressive symptoms they may have and the coping strategies they use to deal with them. By completing the measures, individuals may reflect on and think more thoroughly about both the symptoms and coping strategies they use and it are these reflections, not the intervention, that cause reductions in depression and increased effective coping at the posttest. Note that these testing effects are controlled for in the two group pretest-posttest design as long as (1) individuals are randomly assigned to condition, and (2) one compares the degree of change for the

experimental group with the degree of change for the control group. Testing effects should affect both groups equally, so any differences in the posttest distributions can't be attributed to testing effects. Alternatively, with random assignment, one can use the posttest-only control group design, which eliminates testing effects altogether. Critics of research will be sensitive to whether testing effects can account for your data rather than the tested treatment/manipulation.

Issue 2: History Effects

History effects refer to events external to the study that may be responsible for changes in the outcome of interest rather than the program or intervention per se. For example, in the single group pretest-posttest design, the effects of a program to decrease stress among patients over a period of 3 months might be overestimated due to events that occur in the community or the broader geographic region, such as an improving economy leading to more family income that, in turn, reduces stress. Or, an environmental program designed to reduce energy consumption might be implemented at the same time that the price of energy spikes upward, with the latter being the cause of people using less energy rather than the program. History effects can be controlled by using either the randomized two group pretest-posttest design or the randomized two group posttest only design because such effects should occur in both the experimental and control group; any between group differences can't be due to history effects. Again, critics will be sensitive to whether history effects can account for the data rather than the tested treatment/manipulation.

Issue 3: Instrument Change

Instrument change occurs when the measuring device used to assess the outcome changes over time in ways that suggest program or treatment effects may be larger or smaller than is, in fact, the case. In Chapter 14, we discussed the idea of observer drift, in which the ways that observers use rating scales or make observations of others can change over time as they become more familiar with the judgment task or they become more sensitized to certain cues as they gain experience making observations. Such observer drift can produce changes in recorded observations from the pretest to the posttest even if a program has no effect. Or, perhaps the interpretation of the meaning of items on a self-report instrument change over time for respondents as respondents become more sensitized to constructs addressed in a program. These changes could produce pretest versus posttest differences rather than the program. Technically, such effects should operate for both the experimental and control groups in randomized two group designs, but the occurrence of instrument change is still worrisome because the construct being studied might be different than one expects as the measure of it changes. Critics of your

research will be sensitive to possible changes in instrumentation as an alternative explanation to your data interpretations.

Issue 4: Statistical Regression to the Mean

The dynamics underlying statistical regression to the mean are not well understood by many, so we develop it in some depth here. Suppose we select individuals from a population with extreme scores in a distribution (e.g., people who score high on a depression scale). If we measure their depression again at a later point in time, we often will find that their scores will "regress to the mean" of the original distribution, i.e., the mean for the extreme group will change over time towards the value of the mean of the full group. In clinical psychology, it is often the case that individuals are selected for treatment based on extreme scores on a pretest/baseline outcome distribution. At posttest, regression to the mean predicts these individuals will exhibit artifactual change that has little to do with the treatment program they participate in. There are many sources of such regression to the mean but we focus on one of them, measurement error. We will use an unrealistic example to make it easier to see the underlying dynamics.

Suppose we have a group of 9 individuals who, unbeknownst to us, have the same true levels of depression (which we will index on a 0 to 100 metric with higher scores indicating higher levels of depression). Table 1 presents their observed scores (Y) on a depression measure and their true depression scores (T) at Time 1 and Time 2, where we have ordered the observed scores from highest to lowest based on their Time 1 values (note: in practice, we do not know the true scores, but let's pretend we do for purposes of pedagogy). Each observed score is an additive function of a person's true score plus some random noise reflecting measurement error (see Chapter 13). The random error is uncorrelated with the true scores because it is random. Some of the random error pushes observed scores upward and some of the error pushes observed scores downward; again, its influence is random. Note that the error scores at time 1 are uncorrelated with the error scores at time 2, also because the error at both times is random; correlating one set of random numbers with another set of random numbers should yield a zero correlation.

Table 1: Regression to the Mean

	Time 1			Time 2		
Person `	Y1	T1	E1	Y2	T2	E2
1	93	90	+3	93	90	+3

2	93	90	+3	87	90	-3
3	93	90	+3	90	90	0
4	90	90	0	87	90	-3
5	90	90	0	93	90	+3
6	90	90	0	90	90	0
7	87	90	-3	87	90	-3
8	87	90	-3	93	90	+3
9	87	90	-3	90	90	0

Suppose we decide to focus a treatment for depression on individuals who are most depressed, so we select the three individuals with the highest observed depression scores. Unbeknownst to us, these individuals are no different in their depression levels from any of the other individuals and their elevated scores are due solely to random error. The mean observed depression score for these three individuals is 93. We expose them to our intervention designed to lower depression and then measure their scores at Time 2, after the intervention. Suppose the intervention is completely ineffective and the true scores of the three individuals remain the same. This is reflected in Table 1. Because random error at one point in time is uncorrelated with random error at another point in time, the random noise that contaminates the observed scores at Time 2 will take on different values than those that contaminated the scores at Time 1. The random errors for the highest scoring individuals at Time 1 were all positive in value (each was +3), but at Time 2, the random errors are now evenly distributed across the three individuals (one is +3, the other is 0, and the third is -3). The mean of the observed scores of the three individuals in Table 1 at Time 2 is now 90 and it looks like, at the observed score level, that the intervention had some effectiveness because it decreased depression from a mean of 93 to a mean of 90 for these individuals. However, the result is an artifact of regression to the mean due to measurement error because we biased our selection of participants towards individuals with positive random errors at Time 1 (whose errors will not necessarily be positive at Time 2).

Regression to the mean is caused not only by measurement error, but also by any variable that disrupts the correlation between variables measured at different points in time. Suppose we use a measure that is perfectly reliable and valid, Y, and we measure a construct at two points in time using it, Y1 and Y2. Suppose further that the mean and variance of Y at both times is the same and that there is a 3 month interval between measures. Suppose we select the lowest 10% of the population based on their scores on

Y1. Assuming the variable is temporally dynamic, the chances are low that on the second testing occasion the exact same individuals will again be the lowest 10% of the sample. To be sure, many of the initially selected people will indeed be in the lowest 10%, but not all of them because variables unrelated to Y1 (called *disturbance variables*) will have caused some people's true standing on Y2 to change upward and some people's true standing on Y2 to change downward, essentially mimicking the random influences on Y in our prior example. If even just a few of the originally selected people are no longer in the lowest 10% because of these random disturbances, the Y2 group mean for the lower 10% must now be closer to the original population mean (which, for purposes of this discussion, is unchanged over time). This same phenomenon operates at the upper end of the distribution as well. As such, regression to the mean is caused by the operation of any disturbance variable that causes the correlation between Y1 and Y2 to be less than 1.00. For elaboration of this point, see Kenny and Campbell (1999).

There are two key points to keep in mind about regression to the mean. First, regression to the mean is a group phenomenon, not an individual phenomenon. We usually have no idea if an individual is going to change upward or downward over time. But, as a group, the mean of the most extreme individuals at one or the other end of the distribution are likely to move closer to the population mean given the operation of random disturbances. Second, regression to the mean operates even if everyone changes by a systematic amount due to some intervention or event between Time 1 and Time 2. It may be the case that everyone's true score in the population increases by a constant of 10 units. Nevertheless, the presence of disturbance variables (such as measurement error) will cause the lowest scorers at Time 1 to gain more, on average, than the overall population gain of 10 units. See Kenny and Campbell (1999) for substantive examples.

In studies evaluating interventions, the best way to deal with regression to the mean is to randomly assign the individuals you select (e.g., the individuals at the lower end of the distribution) to an intervention and control group. Regression to the mean should then operate equally in both groups, so group differences in means will reflect the true intervention/treatment effect. Evaluating change within a given group (e.g., the treatment group) is problematic, however, because it will be subject to regression to the mean if people at one end of the distribution were a priori selected to be in the study to the exclusion of others. Mee and Chau (1991) present a statistical method for analyses of subgroups selected from one (or the other) end of the Y1 distribution to determine if there is an additive effect across time (e.g., because of an intervention) beyond what one would expect on the basis of regression to the mean for that subgroup.

Critics can question your conclusions based on results that are subject to regression to the mean. You want to be sensitive to its operation and control for it if it is plausible.

Issue 5: Maturation

Campbell and Stanley (1963) refer to *maturation* as "all biological or psychological processes which systematically vary with the passage of time, independent of specific external events" (pp.12-13). A single group pretest-posttest design evaluating an intervention over a six month period with middle school adolescents, for example, is contaminated by pubertal changes that naturally occur during this time period. If pubertal changes are relevant to one's outcome, then the study is flawed. An effective intervention designed to improve cognitive functioning in the elderly might appear to be ineffective in a single group pretest-posttest design because of naturally occurring, biologically based decrements in cognitive function that occur at the same time. Maturation often can be controlled by using a two group pretest-posttest design with random assignment to groups because the maturation operates in both the experimental and control groups; any differences between them cannot be attributed to maturation. Critics of your research will be sensitive to possible maturation as a contaminating factor of your conclusions.

Issue 6: Experimental Mortality

Campbell and Stanley (1963) discuss experimental mortality in different contexts but primarily in terms of problems created by research participants dropping out of a study between a baseline assessment and a posttest assessment. In the single group pretest-posttest design, if dropping out of the study is non-random, then it is possible that biased estimates of treatment effects will occur. For example, if early responders to a treatment for depression are more apt to remain in the study but non-responders are more apt to drop out (because the treatment does not seem to be working), then the estimated effects of the treatment will be overly optimistic.

Campbell and Stanley also caution about study dropouts for the randomized two group pretest-posttest design and the randomized posttest only control group design if there is differential drop-out for the experimental and control groups. Again, if the differential drop-out is systematic rather than random, then a treatment can appear to be more (or less) effective than it actually is. For example, a resource and time demanding treatment that is cognitively aversive might lead people who are less motivated to drop out of the treatment condition than in the active control condition that is less aversive and this, in turn, could bias estimates of treatment effectiveness when comparing treatment and control conditions. We return to this matter below in our discussion of intent to treat analyses, but suffice it to say that critics also will be sensitive to the operation of treatment dropouts as a contaminating factor in your research.

Issue 7: Selection Effects

Campbell and Stanley discuss selection effects in the context of two group designs without randomization. *Selection effects* refer to self-selection into the treatment or control conditions by individuals such that the treatment and control conditions are confounded by a host of individual differences now associated with the treatment variable. For example, consider an after-school program to increase reading skills in students. A researcher compares post-program reading skills for youth who completed the program with students not in the program. Without random assignment, it might be the case that students who volunteer to be in the program have different pre-program reading skills and different motivations to achieve in school than students who do not volunteer to be in the program. These pre-existing differences muddy the evaluation of program effects because one does not know if group differences after the program are due to the program or to pre-existing individual differences between the groups.

It turns out that selection effects also can undermine randomized designs if the selection variables interfere with the process of random assignment or if they exert their influence after randomization has occurred. For example, methodologists have suggested that researchers or clinic staff sometimes deviate from random assignment protocols to ensure a person who is in particular need of treatment is assigned to the intervention as opposed to control condition (Rosenberger & Lachin, 2015). Berger (2005) also suggests that researchers or clinic staff may occasionally deviate from the random assignment protocol so as to enroll patients into the intervention if they think the person is more likely to respond to treatment, thus favoring the treatment. Systematic post-randomization selection bias can occur due to treatment dropouts, non-adherence to protocols, attrition, missing data and unintended between-condition differences in the use of co-occurring treatments (e.g., medications to supplement a behavioral therapy trial). Critics will be sensitive to problems of interpretation created by possible selection dynamics in your research. You want to address them when you plan your study.

Issue 8: Selection (and Other) Interaction Dynamics

It is possible for any of the above phenomena to operate in interaction with each other or interact with treatment administration to undermine inferences about a treatment. For example, an intervention might only be effective if it is preceded by a baseline assessment because the baseline assessment sensitizes people to important issues addressed in the intervention. Note that this is not the same as a testing effect. The classic testing effect refers to the impact of completing an assessment on the outcome per se by and of itself. For a testing-by-treatment interaction, the effects of the treatment itself are

enhanced by completing the baseline assessment. Stated another way, anything that occurs after random assignment and prior to the intervention is, in some ways, part of the intervention and contaminating effects of these activities must be considered accordingly.

As an example of a selection-maturation interaction, suppose a study of the effect of a new teaching approach in elementary school on student math abilities is conducted, with students in the control group recruited into the study at the beginning of the school year and those in the experimental group recruited into the study at the end of the school year. There are selection effects present such that the students in the control group are younger than the students in the experimental group and one would expect naturally occurring increases in cognitive abilities as a result of maturation over the course of the school year.

In sum, when conducting research to test the effect of a manipulation, a program, or an intervention on outcomes, it is useful to design your study so as to rule out the types of threats to valid inference discussed by Campbell and Stanley (1963). Critics of your theory tests will use these threats to dismiss your results as uninformative. Your task is to design your study so that you can counter each of these threats. Campbell and Stanley argue that an effective strategy for addressing many of these problems is to use what is called a Solomon four group design, which is used with randomization and has the following format:

Group 1:	O_1 X O_2
Group 2:	O_1 O_2
Group 3:	X O_2
Group 4”	O_2

Note that this is a combination of the two group pretest-posttest design and the single group posttest only design described earlier. Comparing O_2 for Group 4 with O_2 for Group 2 diagnoses testing effects. Comparing O_2 for Group 3 with O_2 for Group 1 diagnoses testing by treatment interactions. Comparing O_2 for Group 2 with O_2 for Group 3 tests for intervention/treatment effects. For more details, see Campbell and Stanley (1963) and Reichardt (2019).

POPULATION/SAMPLING CONSIDERATIONS

The next set of issues we consider are those related to sampling and the specification of populations to which one seeks to make inferences. Campbell and Stanley (1963) make a distinction between what they call *internal validity* and *external validity*, with the former

referring to the supposed truth value of the conclusion within an experimental context and the latter referring to the generalizability of that conclusion to other populations, contexts, and times. Campbell and Stanley classify "threats" into those that affect internal validity and those that affect external validity. Sampling "threats" often are considered most relevant to external validity. We discuss five issues relevant to sampling threats, (1) biased sampling, (2) specification of inclusion/exclusion criteria, (3) self-selection into a study, (4) population homogeneity/heterogeneity, and (5) sampling weights.

Issue 1: Biased Sampling

One potentially important sampling issue is whether the sample in your study is representative of the population you wish to generalize to. If the sample does not represent the target population, then your conclusions about the target population can be questioned as being misplaced. An obvious and extreme example would be if you wanted to make statements about the elderly but conducted your research on young adults.

Sampling issues in social science research are more nuanced than most people realize. Population parameters (e.g., means, correlations, group coefficient differences) can be estimated with reference to either small, finite populations on the one hand to populations that are hypothetical and so large that we do not even know their size. Social science research is often conducted with the goal of explaining the behavior of large numbers of individuals, often including people who have lived previously or who have yet to be born, as well as those residing in the present. For instance, if we are studying the course of a particular brain disease, we might think of the relevant population as all people, whether currently living or not, who have ever had or will have the illness. In this case, the focus is on an extremely large hypothetical population.

The traditional model for thinking about sampling involves two steps. First, we define the population of individuals we want to make statements about. Second, we enact procedures that generate a random (or approximately random) sample from that population. A key issue in this approach is whether your sample is a reasonable approximation of a random sample from the specified population. We refer to this approach as the *population-first, sample-second* approach.[1]

Many social scientists, however, turn this logic on its head by conducting a study on a group of individuals and then declaring that this group of individuals represents a random sample from some unspecified population. The task is to specify the population that the sample represents. This essentially frames the issue of sampling in terms of generalizability rather than traditional sampling theory: From what population does our

[1] Sometimes scientists purposely oversample groups or use forms of stratified or area random sampling instead of pure random sampling, but then apply statistical adjustments to correct for such design effects.

sample data reflect a random sample? We refer to this approach as the *sample-first, population-second* approach.

Suppose we want to characterize the attitudes of people in the United States about legalizing marijuana. It obviously would be folly for me to conduct a study on college students in a large Northeastern university, assess how favorable they feel toward legalizing marijuana, and then claim that their measured opinions can be construed as if they are a random sample of the general United States population. On the other hand, suppose we want to characterize the effect of smoking marijuana on brain physiology and We again conduct my study on college students in a large Northeastern university. We find that smoking marijuana impacts the behavior of anandamide molecules in the hippocampus. we might argue that for these particular variables and for this particular relationship, the college students essentially function as a random sample of people in the United States and the results can be generalized accordingly.

Exactly what population sampled individuals are assumed to represent depends, in part, on the characteristics of the participants in the study, the context in which the study is conducted, and the question that is being addressed. If a learning experiment is conducted on 200 college students, the investigator might want to generalize his or her results to adults in the United States. In this case, the population would be conceptualized as consisting of "all adults in the United States," and the college students are assumed to represent a random sample from this population *with respect to the variables being studied*. Obviously this is a questionable assumption. Perhaps the population should be conceptualized as "all college students." Or maybe it should be conceptualized in even more specific terms.

In practice, the approach of inferring a population from a sample is far more common in social science research than the approach of first specifying a population and then seeking random samples from it. This is especially true of randomized trials where volunteers are often recruited from flyers, radio advertisements, posters on public transportation, or patients in a particular set of clinics in specific geographic locations. Study volunteers are thought to represent a random sample from some population, and the question is just who is that population?

An important point to keep in mind when answering this question is one that many researchers do not appreciate, namely that biased samples can yield unbiased population estimates. Suppose a population has 50% males and 50% females. A researcher is interested in estimating the divorce rate in the population and, unknown to the investigator, the true overall divorce rate is 40%. Suppose further that the divorce rate for males is 40% and it also is 40% for females. Stated another way, gender is unrelated to divorce rates. Now, suppose we conduct a study where our sampling frame, for whatever

reason, purposely oversamples males relative to females by a 3 to 1 margin. If one of our goals is to estimate the overall population divorce rate, the gender bias in our sample is irrelevant; we would get the same basic result if we used a sampling frame that included equal numbers of males and females because gender is unrelated to the parameter being estimated. The sample bias in gender is moot. Bias only matters for variables that matter, not variables that are irrelevant.

Some argue that when the focus is on basic biological mechanisms or basic mental processes, it is not unreasonable to assume one's sample can be construed as a sample from people in general *for variables that matter for the particular phenomena under study*. For a study that addresses how psoriasis is impacted by a new drug, a sample of volunteers from a psoriasis clinic in Buffalo, New York, the argument goes, probably can be construed as, functionally, a random sample from a large portion of adults in the United States. Or can it? The onus is on the researcher to make a reasonable case for the population the sample represents relative to the study variables. Usually this case is made in Discussion sections of reports and usually it is framed in terms of generalizability rather than random sampling from populations.

In sum, critics may argue that your study has limited generalizability or that your conclusions are misplaced because your sample does not represent the populations you are asserting they do. These are criticisms you will want to be able to address.

Issue 2: Inclusion/Exclusion Criteria

In general, you should explicitly state if there are exclusion criteria that will be used to eliminate people from your study. If there are explicit inclusion criteria, you also should state those. Doing so allows critics to judge the appropriateness of your sample/population. Critics can raise issues about using unreasonable inclusion or exclusion criteria.

Issue 3: Self-Selection into the Study

Research in which people can self-select into study participation run the risk of producing biased samples. We are not referring here to dropping out of a study after it has commenced (per Campbell & Stanley, 1963), but rather to selection bias tied to recruitment into the study prior to it starting. With such self-selection, the generalizability of results can be compromised. For example, mail surveys can be biased if only certain types of people mail in their responses. Phone surveys can be biased if only certain types of people agree to take the survey when called. Research that recruits people through advertising is subject to self-selection because only certain types of people may respond to the ads. Studies suggest that research volunteers often tend to have higher education,

higher SES, a greater need for social approval, higher sociability, more sensation seeking, more need for conformity, they are more religious, more altruistic, and they are more likely to be female, to name a few empirically documented differentiating features.

In general, it is good scientific practice to collect data on a small set of variables for a subset of people who choose not to participate in your study to determine if they differ in meaningful ways from those who agree to participate. If non-trivial selection effects are evident, then perhaps you should revisit the recruitment protocol you are using.

In sum, critics might raise selection effects as a means of questioning disconfirming data, arguing that the disconfirming results are not generally applicable and are distorted by self-selection dynamics.

Issue 4: Population Homogeneity/Heterogeneity

Depending on one's purposes, it possible to choose a population for study that is too heterogeneous or too homogenous for purposes of adequate theory tests. When studying gender differences in anxiety in adolescents, for example, some researchers use as inclusion criteria youth who are between the ages of 10 and 18. However, this is a diverse group of youth with quite different life experiences who likely have very different sources of anxiety. Youth who are ten years old are typically in the fifth grade in elementary school, whereas youth who are 18 years old are typically seniors in high school (or they have just graduated from high school). Fifth graders typically have not experienced puberty; their brain development and cognitive abilities are much reduced relative to high school seniors; they tend not to be involved in dating or romantic relationships; they usually are not in the same school with older adolescents who may have experimented with alcohol, drugs, and sex; grading is not heavily emphasized in elementary school and there usually are no "tracking" programs for bright students; there tends to be an emphasis on cooperative learning in elementary schools and the schools are smaller; there is a single home-room teacher in elementary schools, with students spending all day in the same classroom with the same classmates; and elementary students often have relationships that are strong and close with their parents. None of this tends to be true for high school seniors. With so much heterogeneity in the population of interest, it often is difficult for a theoretical "signal" to emerge from the background "noise" of heterogeneity.

Some research questions require population heterogeneity, such as studies of longitudinal dynamics and research on moderating effects as a function of age or some other target variable. Also, some researchers are interested in characterizing general trends in diverse populations, such as national studies of the United States population. However, one also must be careful when mixing diverse populations because of the

challenges that doing so can create and because of aggregation bias dynamics, which we discuss later.

On the flip side, one also can study populations that are so homogenous that the scope of the research becomes too narrow and of limited interest, per our discussion in Chapter 3 of the main text. Researchers need to find an appropriate balance between population heterogeneity and scope relative to one's broader research goals. The bottom line is that critics of theory tests might criticize your research as focused on populations that are so homogeneous that your results have little generalizability, or they might criticize your research as working with populations that are so heterogeneous that there is too much "noise" for a viable signal to come through. You want to prepare yourself for either criticism as you plan your research.

Issue 5: Sampling Weights

In many national surveys, study designers purposely oversample subgroups in order to have a large enough sample size in those subgroups to obtain stable parameter estimates. Or, study designers might use a sampling strategy where they know certain selection biases will be present. For example, many on-line survey firms rely on panels of people who agree to be "professional survey takers" for the firms. In exchange for payment, the panel members agree to allow the firm to send them solicitations to participate in on-line surveys. The panel member can then choose to participate or not in the on-line survey. A potential "selection effect" is that such on-line panel members must have access to computers, which can bias the sample away from poor people. Survey designers often employ weights during data analysis to correct for bias as a result of oversampling, selection effects and attrition. Because the use of sampling weights is not well understood by many, we delve into the topic here in some depth. In some regards, the use of weights during data analysis is a statistical issue because it is in the context of such analyses that weights are used. However, we consider it in this section because (a) it bears directly on sampling from populations, and (b) the decision to rely on weights to correct for selection bias is a design decision.

The general idea of the weighting process can be illustrated using the formula for a sample mean. The traditional formula for the sample mean is the sum of the scores divided by the sample size. A more general formula can be written that incorporates weights. It is:

$$M_X = \Sigma w_i X_i \; / \; \Sigma w_i$$

where M_X is the estimate of the population mean for variable X, w is a weight assigned to the score of each individual, i, and the summation occurs across individuals. Each

individual's score on X is multiplied by his or her "weight value" and the sum of these weighted scores is divided by the sum of the weights. Suppose we assign everyone a weight of 1. The sum of the weights (the denominator) will equal N (the sample size) and the sum of the weighted X scores will be the sum of the X scores. The result is the traditional sample mean. In this sense, the traditional formula for a sample mean is a special case of the above formulation, namely the case where all the weights are 1.0.

Suppose we are studying a population where we know that 50% of the individuals are male and 50% are female. However, we use a sampling frame that oversamples males so we end up with 75% males and 25% females. One way to adjust for this bias is to not assign a weight of 1 to everyone when we calculate the mean. If we assign smaller weights to males and larger weights to females in proportion to the disparity between the known population distribution of gender, the weights will correct for the sample bias.

The construction of such weights can be complicated when there is more than one selection effect operating, i.e., when we have to build in corrections for multiple biases. Note also that to construct a sampling weight for a selection variable, we need to know the true population distribution for that variable. In the above example, we knew that 50% of the individuals in the population were males and 50% were females. With this knowledge and knowing the gender distribution in the sample data, we can construct weights that adjust for the disparity. When constructing weights to correct for bias in national surveys in the United States, sampling statisticians often use data from the U.S. population census as their reference. The problem with this strategy is that the census is taken only once every 10 years, so it can be outdated. The census also focuses on only a small number of variables, such as gender, ethnicity, and other core demographics. This means that sampling weights can only be calculated that take those variables into account. Also, some scientists argue that the census data are not all that accurate for some of the variables assessed. The bottom line is that the weights one constructs are only as good as the accuracy of the assumed population values of the variables that are used in the weighting process.

Suppose we construct sampling weights based on a careful analysis of 10 selection variables. Suppose also that the process is successful such that it does indeed correctly adjust for the bias that is operating in these variables. This does not mean that bias has been corrected on other variables. For example, we once analyzed data collected from an on-line panel that adjusted for bias on basic demographic variables for the United States population of teenagers (e.g., gender, ethnicity, class, age). The survey was on the sexual activity of American youth between the ages of 14 and 17. We knew the results of many well-conducted national surveys of this population and we knew the percentage of youth in this age range who reported having engaged in sexual intercourse was about 45%.

When we calculated the percentage in the on-line panel data using sampling weights, the estimated percentage was 25%. The sample was clearly biased on the variable of sexual intercourse. The sampling weights corrected for bias in ethnicity, gender, class, and age, but it did not correct for the bias in sexual activity. This is a criticism of the reliance on sampling weights – you simply can't anticipate and correct for all the selection effects that may be operating. The on-line panel study corrected for some selection variables but not others.

Although the use of sampling weights can correct for bias, doing so comes at a cost, namely decreased efficiency (in a strict statistical sense of the term) of estimators. Stated another way, statistical estimates based on weights tend to have larger standard errors and more sample to sample fluctuation than unweighted data. This lower efficiency is undesirable because it lowers statistical power. Making such a tradeoff is especially disconcerting if the selection effects do not really produce much bias in estimates in the first place, as we discussed earlier. To address this matter statisticians have developed indices of weight informativeness to help guide the decision about whether to use weights for a particular modeling situation. Use of these tests involve complex issues that are beyond the scope of this primer. See relevant discussions by Asparouhov (2006), DuMouchel and Duncan (1983), Feinberg (1989), Kott (1991), Little (1991), Lohr and Liu (1994), Winship and Randall (1994), Pfeffermann, (1993) and Pfeffermann and Sverchkov (2009).

In sum, critics might dismiss a theory test if you fail to use sampling weights that are provided by the study designers. Their argument is that your sample is biased and that your conclusions are therefore misplaced. However, the use of such weights is not always appropriate and you may need to justify your use (or non-use) of them based on weight informativeness tests. Indeed, if you do not, some critics may actually criticize you for using weights because it reduces the statistical power of your tests, which can undermine your conclusions.

MEASUREMENT CONSIDERATIONS

Chapters 13 and 14 in the main text described measurement issues to take into account when conducting research and designing theory tests, so we do not repeat them here. In general, critics can raise issues about the poor mapping of measures onto constructs, the unreliability of the measures, the invalidity of the measures, group or facet differences in the metric properties of measures, overgeneralizing what the measures represent, and issues of instrument change/observer drift, among others.

DESIGN CONSIDERATIONS

In this section. We briefly describe 11 methodological issues, in no particular order, that might be applicable to your research design. They are (1) improper control groups, (2) non-random assignment, (3) attrition bias, (4) carry-over effects, (5) assessment timing, (6) experimenter bias demand characteristics and reactivity, (7) fidelity and manipulation checks, (8) contamination, (9) range restriction, (10) task motivation and boredom, and (11) ecological validity.

Issue 1: Improper Control Groups

When conducting research that requires a control group, critics can raise issues about the appropriateness of the control group you choose. In clinical research, for example, distinctions are made between *active control groups*, *passive control groups*, and *treatment as usual* (TAU) control groups. For example, a study might examine the effects of cigarette smokers writing a counter-attitudinal essay about reasons not to smoke on subsequent cigarette smoking behavior. Individuals in the treatment condition come to an office, sit in a quiet room, and are provided a cover story for why they are to write the essay. One month later, participants are contacted by phone to complete a general survey on health and one of the questions assesses the frequency of cigarette smoking during the past month. Individuals in the control condition are contacted at the time of the phone assessment and administered the same survey. Indices of smoking behavior are then compared for the two conditions. This is an example of a passive control group.

Alternatively, individuals in the control condition could come to the office and write a counter-attitudinal essay, but they do so on an unrelated topic. They then complete the health survey by phone one month later. This is an example of an active control group. With an active control group, the tasks are equated as much as possible between the treatment and control groups except for the presumed “active ingredients” of the treatment program. With a passive control group, no such equating is sought; rather, nothing is done on the part of the researcher other than the outcome assessment.

As noted, some researchers use TAU as a “control” group. A TAU is an appropriate comparator to a new intervention when primary interest is improving the status quo. However, some scientists criticize the use of TAUs as control conditions when the goal of the study is to test or advance theory. For example, two researchers might evaluate the same intervention under generally comparable circumstances but the particular TAU for one researcher might be reasonably effective given standard clinic practices in his or her community whereas the TAU for the other researcher might be ineffective given standard clinic practices in the latter’s community. The two studies might find differential program

effects primarily because of quality differences in the TAU conditions despite the fact the programs themselves have comparable effects on the outcome.

Without a careful analysis of what the TAU represents, it can be difficult to know what the new program is being compared to, thereby limiting theory development. To be sure, we *do* learn that the new program is better than the status quo and that in itself might be useful for scenarios where the type of TAU studied is common. However, to build scientific theory, a reliance on the operative TAU in arbitrarily selected clinics can be limiting. Some methodologists argue that researchers conducting theory tests should instead consider creating an informative control condition in which the treatment and control conditions are equivalent in all respects except for the new "active ingredients" in the program that one seeks to evaluate.

In sum, control groups need to be appropriate for the issue being considered and critics can question results by questioning such appropriateness. You need to think carefully about your control groups and have a solid rationale for the choices you make with respect to them.

Issue 2: Non-Random Assignment

Random assignment to groups is an experimental ideal, but it is not always possible. When we are forced to assign people to different conditions in non-random ways or when there is self-selection into a treatment versus control condition, there is always the possibility that outcome differences between the groups are due to pre-existing group differences of some kind. For example, when studying the health implications of smoking, it is not possible to randomly assign people to the groups "smokers" and "non-smokers." In such cases, social scientists often use what are called quasi-experimental designs to make causal inferences or they use specialized methodological (e.g., matching) or statistical methods (e.g., analysis of covariance, propensity analysis) to try to adjust for pre-existing group differences. Consideration of these topics is well beyond the scope of this primer. For useful resources, see Holmes (2013), Reichardt (2019) and Shadish, Cook and Campbell (2001).

In designs that rely on non-random assignment, critics have recourse to attributing group differences to a confound that is associated with group membership rather than the manipulation. For example, the cigarette industry maintained for many years that the association between smoking and adverse health was attributable to SES differences in smoking; lower SES people tend to smoke more and they also have poorer access to health care, creating a confounded relationship between smoking and poor health. To address critics who object to non-random assignment and who will dismiss your results,

accordingly, you may need to use one of the quasi-experimental strategies discussed above.

Parenthetically, sometimes researchers intend to use random assignment but do not use proper assignment protocols to operationalize random assignment correctly, which also can undermine random assignment. Strict random assignment generally requires the use of tables of computer generated random numbers in conjunction with specific protocols for executing assignment (see Altaman & Bland, 1999; Cook & Payne, 2002; Kalish & Begg, 1985; Suresh, 2011).

Issue 3: Attrition Bias

We discussed attrition issues briefly in the context of Campbell and Stanley's (1963) threats to internal validity. Here, we elaborate on the matter. In clinical trials, it is not uncommon for researchers to conduct multiple follow-up assessments, such as an immediate posttest, a 6 month follow-up, and a 12 month follow-up. People can drop out of a trial during treatment such that they do not complete the full treatment protocol or they might complete the treatment but then drop-out of the study so that they do not provide data at one or more of the post-treatment assessments. We refer to the former as *treatment dropouts* and the latter as *assessment drop-outs*. Treatment drop-outs are particularly problematic because they do not receive a full "dose" of the treatment and they can cause one to underestimate the effect of the treatment because the treatment was not fully executed. In such cases, truly promising treatments might be dismissed as ineffective when, in fact, they would be effective if completed per protocol.

Dropping out of a treatment can be due to random events or it might be systematic, such as due to a failure of the treatment to produce results early on in the treatment. Or, people who are less motivated to attend treatment may drop-out of treatment prematurely. Critics of studies evaluating treatments are careful to examine treatment drop-out rates and to suggest systematic drop-out dynamics that may undermine study conclusions. When designing research to evaluate treatment effectiveness, it is important to minimize non-random attrition for treatment-dropouts.

Assessment drop-outs are not necessarily problematic as long as dropping out mimics a random process. Having said that, if there are many assessment drop-outs, this can reduce statistical power. Systematic drop outs can undermine study conclusions. As such, critics also carefully examine assessment drop-out rates and often hypothesize the presence of systematic biases that can render study conclusions ambiguous. Several effective statistical methods have been developed for adjusting for assessment drop-out bias (see Enders, 2010).

In the final analysis, you will want to adopt study procedures that minimize treatment and assessment drop-outs and that ensure dropping out of either type is not systematic in ways that bias study conclusions.

Issue 4: Carry Over Effects

Some research designs address treatment versus control comparisons using a within-subjects design rather than a between-subjects design. For example, suppose an investigator is studying the relative effects of two drugs, A and B, on learning among the elderly. Fifty people are administered Drug A and then work on a learning task. One month later, the same fifty people are administered Drug B and work on the same type of learning task. Performance is then compared as a function of the two conditions. In such designs, there is the potential for carry-over effects from one condition to the other, thereby contaminating the results. For example, the researcher needs to be certain the effects of Drug A have completely dissipated before testing Drug B. Other forms of carry-over effects are more subtle. For example, taking the learning test during the first session may sensitize individuals as to how to perform better on the test on the second occasion. For within-subject designs, critics can raise issues of carry-over effects, broadly defined. You will want to design your study so that carry over effects do not undermine study conclusions.

Issue 5: Assessment Timing

As noted in Chapter 7 of the main text, the choice of the time interval between assessments or between a treatment and a posttest can be critical for theory tests. To take an obvious example, if one seeks to examine the effect of aspirin on headaches, it does not make sense to measure headache severity 20 seconds after an aspirin has been taken; it takes time for the aspirin to work. Similarly, it would not make sense to evaluate the effect of aspirin on headache severity one month after a person has taken the aspirin. Choice of time intervals matters. A treatment to improve parenting may show little effect on child outcomes unless parents are given sufficient time to change their behavior and those behavioral changes can filter through to the child outcomes. Critics can complain that the time interval used in a study is either too short or too long to adequately reveal the relevant causal dynamics. You need to justify your choice of time intervals.

In causal thinking, a cause is assumed to precede an effect in time. This means that causes, ideally, should be measured prior in time to effects. Sometimes, the amount of time it takes for a cause to produce an effect is extremely short, virtually instantaneous. In such cases, assessment of the cause and effect likely can take place during the same assessment session. Alternatively, because of pragmatics (e.g., cost), a researcher may

have no choice but to measure a cause and an effect during the same assessment occasion even when there is a non-trivial time interval between cause and effect. When this occurs, critics can raise objections related to timing. For example, some critics assume the causal dynamics have already played themselves out but the direction of causality is the reverse of what was hypothesized, namely perhaps the purported "cause" is really the "effect" and the purported "effect" is really the "cause." You need to be prepared to deal with critics who raise issues of inappropriate timing of assessments.

Issue 6: Experimenter Bias, Demand Characteristics and Reactivity

Experimenter bias refers to a process whereby scientists or research assistants unknowingly or unwittingly engage in behaviors that bias results of a theory test in a favored direction. In a classic example by Rosenthal and Forde (1963), graduate students were asked to train rats to learn a maze. One group of students were told the rats were "maze bright" (that the rats had been in-bred through selective breeding to be good at running mazes) while the other half of the students were told the rats were "maze dull" (the rats were in-bred to be poor at running mazes). In fact, all the rats in the study were regular laboratory rats. Results showed that the maze bright rats learned to run the maze faster than maze dull rats. Rosenthal and Forde found that the students conducting the experiment handled the maze bright and maze dull rats differently when they placed the rats into the mazes. People engaging in data collection can sometimes bias responses in desired directions and critics will inevitably raise this issue unless procedures are put in place to prevent it (e.g., using data collectors who are unaware of the theories and hypotheses being tested).

Demand characteristics refer to cues provided by researchers or the research setting that communicate the purpose of the study to respondents. Once learning the purpose of a study, participants will often act in ways to support the researcher's theory in order to try to please the experimenter. Researchers need to avoid conveying demand characteristics to study participants. Common strategies for dealing with demand characteristics include emphasizing in orienting instructions the importance of participants acting naturally and being truthful in all aspects of the study. Another strategy is to provide participants with an effective cover story that masks the true purpose of the study.

Reactivity refers to changes in study participants' behavior simply because they are aware they are being studied. Reactivity can foster either greater compliance or greater non-compliance in participants and it can reduce study validity and generalizability, accordingly. When people know they are being observed for purposes of a scientific study, the fact is their behavior can change. Critics can raise this issue. Researchers need to devise strategies to reduce or eliminate reactivity.

Issue 7: Fidelity and Manipulation Checks

When we implement a program in an applied setting for purposes of testing its effectiveness, it is important that the program be implemented faithfully. If key elements of the program are changed by program staff, unbeknownst to the researcher, or the program is only partially implemented, the estimate of treatment effectiveness can be undermined. The same is true for any type of manipulation in a laboratory or experimental study; if the manipulation is weak and does not accomplish what it is intended to accomplish, then a theory test can be undermined. Researchers routinely include in their studies fidelity and manipulation checks to counter critic arguments of program infidelity or weak manipulations. In some studies, researchers inadvertently manipulate other factors in addition to those intended, thereby creating confounds. One also wants to address this possibility when designing research, as critics may also raise issues related to manipulation confounds.

Issue 8: Contamination

In designs that use a treatment and a control group, it is possible for some individuals in the control group to inadvertently be exposed to portions of the treatment. Sometimes people in the treatment condition will talk with people in the control condition during the course of the study and tell them about their treatment experiences. Sometimes people in the treatment condition will share materials they received with people in the control condition. Sometimes individuals in the control condition will "cross-over" and fully participate in the treatment condition, unbeknownst to the researcher. Critics will want assurances that contamination has not occurred and if it has, that corrections for it have been made.

Issue 9: Range Restriction

When we examine correlations between variables, the magnitude of the correlations can be impacted by a phenomenon known as *range restriction*. Restriction of range occurs when the sample data are not available across the entire range of scores of interest for the variables being correlated. In such cases, the observed correlation in the range restricted data will be lower than if data from the entire range had been available and analyzed.

Range restriction results in a systematic underestimation of a correlation. A common scenario where range restriction occurs is when we seek to estimate the population correlation between a selection method, such as test performance on a pre-college ability test, and a criterion, such as success in college, but we only have data on college success for people who were admitted to and attend college (see Dahlke, Sackett

& Kuncel, 2019, for substantive examples and further elaboration). Range restriction also can affect conclusions from literature reviews that use meta-analyses that rely on standardized effect sizes. There are ways of statistically correcting for its effects in meta-analyses (see Dahlke et al., 2019).

In sum, if your research focuses on correlations or other standardized coefficients for purposes of theory tests, critics may raise the issue of range restriction relative to study conclusions.

Issue 10: Task Motivation and Boredom

As discussed in Chapters 13 and 14, individuals can become bored in experiments or when completing interviews; or, their motivation to take the tasks seriously may be low. As a result, they may not respond conscientiously, which can introduce random error into the study. For complex experiments or studies, critics may question results by questioning respondent motivation.

Issue 11: Ecological Validity

The *ecological validity* of a study refers to the extent to which the methods and setting of the study approximate real-world settings and thereby activate real-world processes. Ecological validity is sometimes equated with a construct known as *mundane realism*, which refers to the extent to which an experimental situation is similar to situations people encounter outside of the laboratory. For example, when studying mock juror decision making in a laboratory setting, researchers seek to make the laboratory setting as similar to real life court contexts as possible, including having juror boxes where a small group of "jurors" sit, a "judge" and a prosecuting and defense "attorney." Not all research requires ecological validity, such as research on basic biological or psychological processes, but there are research domains where it is important and where critics will raise questions about the sterility or unrepresentativeness of the study context.

DATA MANAGEMENT AND STATISTICAL CONSIDERATIONS

In this discussion, we discuss 16 data analytic issues that critics can raise with respect to a study and that you will want to address, The issues are (1) data management, (2) sample size, (3) outliers, (4) assumption violations, (5) clustering, (6) missing data, (7) intent to treat versus per protocol versus CACE analysis, (8) matters of metric, (9) endogeneity, (10) aggregation bias, (11) misspecified functional forms, (12) selective reporting and p hacking, (13) reporting of effect size, (14) reporting of indices of sampling error, (15) familywise error, and (16) accepting the null hypothesis

Issue 1: Data Management

Data management involves, among other things, ensuring data are recorded accurately and integrated accurately for purposes of data analysis. It goes without saying that this matter is fundamental to making accurate inferences from research data. You will want to use good data management protocols.

Issue 2: Sample Size

When planning a study, one must make decisions about the sample size. Most social scientists think about sample size largely in terms of statistical power, but such decisions should encompass more than just power considerations. Statistical power refers to your ability not to miss important effects by falsely declaring a test of those effects as statistically non-significant. If you declare a result as being statistically non-significant, critics will want assurances that your study is adequately powered. Statistical power varies by statistical method, so when conducting a power analysis to help determine sample size, you will want to focus on the most sample-size demanding statistical analysis you intend to conduct.

Sample sizes impact sampling error and, in turn, the width of confidence intervals about your parameter estimates. We suggest that you think about confidence intervals in terms of the concept of margins of error. All of us have encountered the concept of margins of error in everyday life in opinion polls. Reputable national polls often report a percentage accompanied by a margin of error (MOE), such as “the percent of people favoring Policy X is 65%, with a margin of error of plus or minus 5%.” The margin of error is useful because it gives us a sense of how much confidence we can have in the reported estimate. If for the above poll, the margin of error was plus or minus 30%, then we would not give the estimate yielded by the poll much credibility. By contrast, if the MOE is only 1%, we have more confidence in drawing conclusions from the reported percentage.

Margins of error are defined by statisticians as an index that conveys a sense of the amount of sampling error that is operating. It is often operationalized in terms of a confidence interval. A common strategy is to determine the absolute distance between the sample estimate and both the upper and lower limit of the 95% confidence interval. The MOE is the larger of these two absolute values. If the mean difference in a sample of the annual salary for male and female professors is $5,000 with a 95% CI of $3,000 to $7,000, the lower limit difference is

$3,000 – $5,000 = -$2,000

and the upper limit difference is

$7,000 - $5,000 = $2,000

The absolute value of both of these is 2,000, so the margin of error is plus or minus $2,000. Based on this, one would report the average salary disparity between males and females as $5,000 (±$2,000).

MOEs can be calculated for most any parameter type. The size of the margin of error one will obtain in a study is impacted, in part, by sample size. Just as one can conduct *a priori* power analyses to help make sample size decisions, there exist statistical methods that allow one to specify the margins of error one likely will obtain in a study based on a given sample size (Jaccard, 2018; Maxwell, Kelley & Rausch, 2008). We recommend you also make use of these methods when making sample size decisions so as to ensure your MOEs will not be unacceptably large and subject to criticism.

Another consideration when making sample size decisions is whether the statistical methods you will use rely on asymptotic statistical theory. Asymptotic theory, stated simply, refers to scenarios where the sampling distribution of a parameter behaves in mathematically tractable ways only when sample sizes are large. A mathematically tractable sampling distribution is important because we often rely on such distributions to derive p values for significance tests and to calculate confidence intervals. A key question when using statistical methods that rely on asymptotic theory is at what sample size is asymptotic theory compromised. You need to be able to justify your sample size in this context if critics raise the issue.

A final factor we mention that often needs to be considered when making sample size decisions is the stability of the covariance matrix being analyzed if you plan on conducting multivariate analyses (such as multiple regression, analysis of covariance, or structural equation modeling). The concept of sample covariance matrix stability is complex, but it has to do with how well the sample covariance matrix represents the population covariance matrix and the extent to which deviations based on sampling error can lead the analyst astray during the model evaluation process. In simple terms, a reasonably stable sample covariance matrix is one that preserves the rank ordering of correlations in the population among all possible pairs of variables and where the variances and covariances in the sample matrix are not too discrepant from the corresponding population variances and covariances. Covariance matrix instability is usually the motivating force behind the many (and often highly discrepant) rules of thumb that one encounters about the “number of subjects per variable” needed for effective analysis. Most simulation studies find that these rules of thumb have little

validity (Wolf, Harrington, Clark & Miller, 2013). The only sure way to gain perspectives on the matter is to conduct a Monte Carlo simulation relative to the statistical procedures you intend to use (Muthén & Muthén, 2002).

In sum, when making sample size decisions, you should take into account statistical power, confidence interval width, asymptotic theory, and covariance stability. Critics can dismiss study results if your sample size is too small relative to these facets of analysis.

Issue 3: Outliers

Outliers are aberrant cases whose inclusion in data analysis can distort characterizations of fundamental data trends. It is important that analysts check for outliers and, if present, deal with them. In multiple regression analyses, distinctions are made between outliers and leverage. Outliers refer to how far a predicted Y value for an individual is from the fitted regression line. Leverage refers to how unusual a person's multivariate predictor profile is. The impact of a case on regression coefficients is a multiplicative function of that case's residual and leverage. Most of the commonly used methods for analyzing outliers and leverage are problematic because the methods are impacted by the very outliers they are intended to identify. More modern and effective outlier/leverage identification methods are discussed in Wilcox (2017). Critics can question the validity of results that ignored the possibility of outliers/leverage during data analysis.

Issue 4: Assumption Violations

When we apply traditional statistical methods, we often make assumptions about how data are structured in the population. The most common assumptions concern independence of observations, normality, and homogeneous variances. These assumptions are frequently violated. For example, Micceri (1989) examined distributions of 440 large-sample secondary data bases containing a wide range of achievement and psychological measures. In every case, departures from normality were evident. It is commonly believed that many of the most frequently used statistical methods are robust to assumption violations, i.e., that they perform well even if their assumptions are violated. This is not necessarily the case. A set of more modern statistical methods have become available that provide such robustness in many contexts (see Wilcox, 2017). If your data violate assumptions of traditional methods, you should consider analytic alternatives to avoid critics questioning your conclusions relative to assumption violations.

Issue 5: Clustering

Related to the above issue is the problem of clustering. Some surveys and research designs randomly select clusters of individuals rather than individuals *per se* even though data analysis is conducted at the level of individuals. For example, we might randomly sample classes within a school and include every student in each class in our sample. In this case, each class is viewed as a "cluster." Or we might administer a therapy to each of 50 small groups of individuals, with each group consisting of 5 members. Our control group might have 50 groups of 5 individuals each, with the groups engaging in an unrelated group activity. This is a randomized cluster design where the groups are conceptualized as clusters.

When we analyze such data at the individual level (ignoring groups), a standard assumption we make is that the error/residual scores in the model are independent. However, in clustered designs, this may not be the case. For example, if a group of individuals contains a particularly disruptive group member, then the outcome scores for all members of that group might be affected, but not members of other groups, since they are not exposed to that disruptive member. Or, a group might have a particularly good group therapist/teacher/leader and all of the members of that group benefit accordingly. The presence of such cluster effects can create dependency structures among the model residuals. If the dependencies are strong enough, then adjustments need to be made in the statistical tests to accommodate them; otherwise, our p values and confidence interval are incorrect. Critics can raise objections to results derived from analyses that ignore clustering in a design where clustering is present.

Issue 6: Missing Data

Missing data are common in some types of research. How one treats missing data can affect the conclusions one makes. The traditional approaches to missing data of listwise and pairwise deletion have been found to be unsatisfactory unless missing data is relatively infrequent. More modern methods of treating missing data include multiple imputation and full information maximum likelihood (FIML), among others. Critics can raise objections to studies that use outdated or inappropriate methods for dealing with missing data. For an introduction to the more modern methods, see Enders (2010).

Issue 7: Intent to Treat vs. Per Protocol vs. CACE Analysis

For clinical trials with treatment drop-outs, an important issue surrounds the analytic strategy one uses to deal with the drop-outs. Some analysts argue for an analytic strategy called intent-to-treat (ITT) analysis. In ITT frameworks, the data from all individuals who

are randomized are analyzed regardless of whether they followed the treatment protocols. Posttest and follow-up data are collected even from treatment dropouts and these data are included in treatment evaluation. The logic is that in real world applications, people drop out of treatments, often because the treatment is too demanding or because of too many obstacles to treatment completion. If the goal is to document the likely effectiveness of the treatment in the real world, then we should allow such factors to operate because they are part of the real-world process.

By contrast, other researchers argue that it makes no sense to evaluate the efficacy of a treatment unless the treatment is fully implemented per protocol. If one wants to know the effect of a drug on an outcome, for example, what sense does it make to include people in the experimental group who have not taken the drug or have taken it haphazardly? These analysts argue for excluding treatment drop-outs from analysis and only analyzing data for people who have completed the treatment per protocol. Such analyses, of course, must deal with the possible undermining of randomization as a result of drop-outs.

When ITT and per protocol frameworks were first differentiated, the general idea was that per protocol analyses would be appropriate in early stage clinical trials when the focus is on establishing the efficacy of a treatment. Once treatment efficacy is established, it is then appropriate to evaluate treatment effectiveness via ITT analyses in later stage clinical trials, as the treatment moves closer to real world implementation. The term *efficacy* typically is associated with the spirit of per protocol analyses whereas the term *effectiveness* is associated with the spirit of ITT analyses.

Unfortunately (in our view), many researchers insist on ITT analyses even in early stage clinical trials and state that it is the gold standard that *all* clinical trials should be held to. To us, ITT and per protocol analyses address different questions. An efficacy trial focuses on a specific question, namely whether a treatment, properly executed, can affect the outcome and by how much. An effectiveness trial, by contrast, confounds two questions, (1) is a treatment efficacious and by how much, and (2) will people adhere to treatment protocols and by how much. The factors that impact treatment effectiveness are not necessarily the same as those that impact treatment efficacy. To us, both types of trials and both forms of analyses are meaningful.

Ironically, many trials that adopt ITT analyses use methodological procedures that are unrealistic in real world settings and that undermine the spirit of an effectiveness trial. For example, participants are paid to be in the study, the researchers exert extra efforts to retain patients in the trial (e.g., multiple phone calls, text messages), and the staff are monitored for implementation fidelity and intervened with if fidelity is suboptimal. These practices tend not to occur in real world settings. If one is truly interested in treatment

effectiveness in real world settings, then one should do a good job of capturing the dynamics of those settings in the context of ITT frameworks.

We also believe an exclusive focus on ITT raises ethical issues. Suppose treatment A is more efficacious than treatment B in per protocol analyses but the two treatments are equally effective in ITT analyses. If only ITT analyses are performed, then when presenting treatment options to future patients, the patients would be told that the two treatments are about equally effective. With the per protocol information in place as well, future patients might be told "Treatment A is much more efficacious than treatment B, but people have trouble staying with it. If you can manage to do so, your chances of successful recovery are much greater." Withholding such information from patients, in our view, is ethically questionable.

The choice of ITT versus per protocol analyses for treatment dropouts is complex (see Feinman, 2009). We believe both forms of analyses are informative. Recently a third type of analyses has been suggested, called complier average causal effect (CACE) analysis. CACE analyses seek to compare outcomes for individuals who completed treatment with the outcomes of control group individuals who would have completed treatment had they been offered it, i.e., the comparison is for individuals who are comparably motivated. Those advocating for CACE analyses have developed clever ways of conducting such comparisons (see Jo, 2002 and Peugh et al., 2017).

If you conduct a clinical trial, critics might criticize you for conducting per protocol analyses, they might criticize you for conducting ITT analyses, and/or they might criticize you for not conducting CACE analyses. You need to think through your choices carefully and be prepared to justify them. There is nothing wrong with pursuing all three forms of analysis if you feel each will provide useful information for your purposes.

Issue 8: Matters of Metric

Outcome measures can be nominal, ordinal, interval, or ratio in character. The type of analytic method one applies to the measures might differ depending on these metric properties. If a metric is considerably non-interval, then a critic can question your conclusions if you applied an analysis in which accurate inference requires interval level data. Note that ordinality of a measure is a matter of degree, with some measures being only slightly ordinal and others being decidedly ordinal. You can often apply analytic methods that presume intervalness to ordinal data as long as the approximation to intervalness is reasonable. Indeed, doing so is often preferable because some ordinal forms of analysis are sample size demanding and make assumptions in their own right that may be unreasonable and that get you into even more trouble than presuming the data are interval in character (Jaccard & Turrisi, 2003). You will want to think carefully

about the type of analysis you apply based on the metric of the variables you are analyzing. Critics may raise metric issues in this regard.

Issue 9: Endogeneity

Endogeneity usually is discussed in two contexts, (1) omitted variable bias, and (2) assuming one way causation when, in fact, reciprocal causation exists. It focuses on the case where one seeks to make a causal inference from correlational data. We first consider the case of omitted variable bias.

Omitted variable bias is also called left out variable error (LOVE). Left out variable error occurs when one leaves out of the causal equation one or more variables that directly affect the outcome over and above the predictors in the equation *and* those external variables are non-trivially correlated with the predictors. If a key determinant of Y that is correlated with X is omitted, then one commits LOVE and the fitted model is misspecified and subject to bias. As an example, suppose the true data generating process in the population is

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

but we estimate

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

The omission of X_3 can create bias in our estimates of the causal impact of X_1 and X_2 on Y. LOVE also can bias significance tests and confidence intervals.

Omitting a key determinant of the outcome over and above the predictors in an equation is not always problematic. For example, if the omitted variable (OV) is relatively uncorrelated with the predictors in the model, omitting the variable will not bias estimation, although it can reduce statistical power. If the omitted variable's impact on Y is completely mediated by the predictors already in the equation (i.e., OV is a distal determinant of Y through the X), then leaving out the variable also is of little consequence. Omitted variables are most likely to introduce problems if (a) the omitted variable is moderately to highly correlated with the predictors and (b) the omitted variable has a non-trivial, independent impact on Y.

In most social science research that seeks to make causal inferences from regression/correlational analyses, LOVE is inevitable. The question is not so much whether LOVE exists (because it almost always does), but rather whether it is of a sufficient size and nature that it will mislead us when making conclusions. Whenever you conduct a multiple regression or SEM analysis with causal inference in mind, you should

think about LOVE. You should consider measuring and including as covariates variables that are theoretically meaningful and that could create bias if omitted. For an in depth discussion of LOVE, see Mauro (1990) and Clarke (2009).

For reverse causality, endogeneity raises issues in another form. As noted in Chapter 7, a bidirectional causal relationship is when a variable, X, has a causal influence on another variable, Y, and Y, in turn, has a "simultaneous" impact on X. If we use the association/correlation between X and Y to infer the magnitude of the causal link between X and Y (as is often the case in SEM or multiple regression analyses), this can be problematic when reciprocal causality is present because that association is inflated or affected by the effect of Y on X in addition to the effect of X on Y. By ignoring this fact, our estimate of the causal impact of X on Y can be biased. Such over-estimation can be addressed through the use of instrumental regression (Bollen (2012).

In sum, you will hear reference to the problem of endogeneity. When endogeneity is invoked, a critic is usually raising the issue of omitted variable bias or reverse causality (sometimes it also is invoked for the biasing effects of measurement error). When you design studies, you need to think about endogeneity and develop strategies to counteract its biasing effects.

Issue 10: Aggregation Bias

Aggregation bias refers to distortions in conclusions that occur from the act of aggregating data. Aggregation bias takes many forms. One form is that of Simpson's paradox, which we discussed in Box 6.1 in Chapter 6. Closely related to this dynamic is the failure to include operative interactions/moderators in a statistical model that lead one to make misleading generalizations. For example, suppose that a treatment for depression only works for females but not for males. If an analyst ignores gender in the analysis, s/he might conclude that the treatment is relatively effective for both genders when this is not the case. Important and consequential subgroup dynamics are obscured. Critics can potentially criticize you for making such specification error, namely the omission of important subgroup analyses that lead to faulty generalizations and inappropriate inferences.

Another form of aggregation bias is when one inappropriately infers mechanisms are operating at the individual level based on group level data and analyses. As one example, researchers may assume that differences in population averages directly apply to individual cases, which is not necessarily the case. For example, mean math achievement scores may be higher for males than for females. This does not imply, however, that all males have higher math achievement than all females at the individual level. A randomly selected male may or may not have higher math achievement than a

randomly selected female even in the face of such mean differences. As another example, if we calculate the average wealth of each state in the United States and correlate that index with the tendency to vote Democratic, there is a positive correlation. For example, in 2004, the Republican candidate, George W. Bush, won the fifteen poorest states while John Kerry, the Democratic candidate, won 9 of the 11 wealthiest states. However, at the individual level, research has shown time and again that wealth is *negatively* correlated with the tendency to vote Democratic. The level of analysis matters!

Another form of aggregation bias is known as *dimensional aggregation*. Suppose a measure of depression has four subscales measuring different facets of depression, (1) affect, (2) negative cognitions, (3) somaticism, and (4) apathy. Suppose that only the apathy dimension is correlated with the engagement in healthy exercise behaviors by the elderly. If a researcher only works with an aggregate index of depression that sums across the four subscales, one might conclude that depression in general is related to exercise, missing the fact that it is only apathy that truly matters. One might falsely conclude in such scenarios that treatment of any of the dimensions of depression will improve exercise behaviors in the elderly, when this is not the case; only treating apathy will be effective. Or, by adding all of the unpredictability or "noise" of affect, negative cognitions, and somaticism to the predictability or "signal" of apathy, one may find a non-significant correlation between depression and exercise and falsely conclude that none of the dimensions matter.

In sum, critics can question your conclusions or theory tests if you use an analytic method that aggregates across data in ways that can mischaracterize the operative causal dynamics. You need to choose the level of analysis carefully and not generalize beyond that level of analysis to other levels of analysis.

Issue 11: Misspecified Functional Forms

When social scientists use correlation or regression based analyses, it is common for them to assume linear relationships between variables. However, the functional form between quantitative variables might be non-linear and treating it as if it is linear might yield faulty and consequential inferences. In Chapter 8 on mathematical modeling, we described a large number of different functional forms that could be operative. Chapter 11 introduced the use of smoothers to explore non-linear functions (see also the supplemental materials for Chapter 11 on our website). Critics can question your conclusions if they suspect you have use a misspecified functional form between variables.

Issue 12: Selective Reporting and p Hacking

In Chapter 15, we discussed the problem of p hacking and selective reporting of the results of statistical tests. Critics can question your conclusions if you engage in such practices.

Issue 13: Reporting of Effect Size

Traditional null hypothesis testing addresses the question of whether an association between variables exists. However, the approach has little to say about the magnitude of the association. It is becoming more common to report, in addition to p values associated with significance tests, effect size indices that capture the magnitude of an effect. Over sixty different indices of effect size have been suggested in the research literature and there is no consensus about which one is preferable. Most (but not all) of the measures can be classified into two classes, (1) unstandardized indices of effect size and (2) standardized indices of effect size. Unstandardized effect size indices are expressed in the raw metric of the outcome measure, such as the actual mean difference between two groups or an unstandardized regression coefficient. Standardized effect size indices are expressed in a transformed metric that is thought to have intuitive and interpretational appeal. The most commonly reported standardized effect size indices for ANOVA designs are Cohen's d and indices that reflect the percent of variance accounted for (e.g., omega squared, eta squared, epsilon squared). In traditional regression analysis, common standardized effect size indices are the squared multiple correlation, standardized regression coefficients, and correlation coefficients. Unstandardized indices include unstandardized regression coefficients and standard errors of estimate.

Both standardized and unstandardized indices of effect size have strengths and weaknesses. Given this, we like to report both types. Critics can question your conclusions or theory tests if you rely exclusively on p values and tests of significance and ignore effect sizes. It is best to report and directly address effect sizes in your research.

Issue 14: Reporting of Indices of Sampling Error

As discussed in the section on sample size, it is good practice to report indices of sampling error either in the form of confidence intervals or margins of error. Critics can question your conclusions or theory tests if there is evidence of large amounts of sampling error or "noise" in your parameter estimates even if the results are "statistically significant, $p < 0.05$."

Issue 15: Familywise Errors

When a researcher conducts multiple contrasts or multiple significance tests, a problem is that the Type I error rate can inflate across those tests. Critics can raise this issue. Consider a coin flipping analogy. If we flip a coin, there are two possible outcomes that can occur, one of which is a "head." For the sake of exposition, lets treat obtaining a head as an "error." The likelihood of observing a "head" or an "error" on a on a given coin toss is 1/2 = 0.50. If we flip a coin twice, there are four possible outcomes that can occur, (1) a "head" on the first flip followed by a "head" on the second flip, (2) a "head" on the first flip followed by a "tail" on the second flip, (3) a "tail" on the first flip followed by a "head" on the second flip, (4) a "tail" on the first flip followed by a "tail" on the second flip. Note that a "head" occurs on three of the four flips, so the probability of a "head" occurring on at least one of the flips is 3/4 = 0.75. Even though the probability of a "head" is 0.50 on a given flip, the probability of observing at least one "head" across two flips is 0.75. A similar process operates with multiple contrasts in statistics. The error rate on a given flip is analogous to the per comparison error rate. The error rate across flips is analogous to the familywise error rate.

Some methodologists feel that controls for inflated familywise error rates should be invoked whenever multiple contrasts are performed. However, doing so comes at a cost. Using such controls reduces statistical power for a given comparison, with the result possibly being an unacceptably high rate of Type II errors. i.e., missing an effect that is important. If Type II errors are indeed important, the reduced statistical power for familywise error rate corrections may be unacceptable. In research areas where sample sizes tend to be small, the issue is particularly germane because statistical power is low to begin with.

Decisions about the invoking familywise error rate controls are complex and you will encounter different recommendations about the need to do so (see, for example, our discussion of Bayesian epistemology in Chapter 15). Critics can criticize you for failing to adjust for familywise error and they can criticize you for adjusting for familywise error. You need to think carefully about the orientation you take with respect to this issue.

Issue 16: Accepting the Null Hypothesis

Situations occur where researchers are interested in declaring the equivalence of groups in terms of their mean scores on variables or in terms of percentages. It turns out that the popular null hypothesis testing approach to statistics that most of us use is not well suited to making such declarations. Consider the case where a researcher wants to test if there are mean differences in starting salaries for new Assistant Professors at major universities

in the United Sates as a function of gender. The traditional null hypothesis is that the difference between the two population means is zero:

H_0: $\mu_M - \mu_F = 0$

The alternative hypothesis is that the difference between the two populations is not zero:

H_1: $\mu_M - \mu_F \neq 0$

If we collect sample data, analyze it, and reject the null hypothesis ($p < 0.05$), then we confidently conclude that the gender difference in salary is not exactly zero. We favor H_1. If the statistical test yields a statistically non-significant result, then it is not the case that we can accept the null hypothesis (H_0) and conclude that there is no gender difference in average salaries. The null hypothesis specifies that the difference in salaries for males and females is exactly zero and there simply is no way to know with any reasonable degree of certainty that the difference in salaries equals a single, exact value, i.e., zero, because of the inevitable presence of sampling error. Even if the observed mean difference is zero, we can't conclude that the true population difference is zero, again, because of the presence of sampling error. We find ourselves in an uncomfortable position of suspended judgment: We can't say that the gender difference in salaries is not zero (because $p > 0.05$), but we also can't say that there is not a gender difference.

Critics can raise objections to your conclusions or to theory tests if you have inappropriately accepted the null hypothesis by concluding no group difference or no association based on p values. To be sure, you can state that the data are not sufficient for you to confidently conclude there is a group difference or association. But, technically, you cannot accept the null hypothesis of no difference. Specialized statistical approaches for declaring equivalence between groups have emerged outside of traditional null hypothesis testing, in a branch of statistics known as equivalence testing (Lakens, Scheel & Isager, 2018; Wellek, 2010). If you are interested in asserting group equivalence, consider using these approaches.

ADDITIONAL CONSIDERATIONS (FOR METHODS ABD THEORY)

In this section, we briefly mention five additional issues you will want to take into account when trying to design effective theory tests.

Issue 1: Weak Conceptual Logic Models

In Chapter 4, we discussed the importance of building strong conceptual logic models for

one's theory. Critics also can question your conclusions and theory tests by attacking your conceptual logic models. See Chapter 4 and the supplemental materials for Chapters 3 and 4 for strategies to address this issue.

Issue 2: Poorly Defined Constructs and Relationships

Critics can question the clarity and appropriateness of your conceptual definitions. Use the strategies described in Chapter 5 to be clear about your concepts. They also can question ambiguity about variable relationships. See Chapter 6.

Issue 3: Generalizing Beyond One's Data

Critics can complain about researchers generalizing beyond their data by extending their conclusions to other populations or contexts that the critic questions are applicable. Editors and reviewers often encourage you to speculate about the implications of your work, but you also must be cautious about not taking this matter too far. This issue is related to the one we identified earlier for the sample-then-population sampling strategy.

Issue 4: Inadequate Literature Review and Theory Interpretation

A high quality and comprehensive literature review is important to any theory test. Critics can raise issues about prior literature that you have failed to consider or your treatment of that literature in superficial ways. They also can argue that you have misinterpreted past research and relevant theories. You need to be thorough and accurate in your consideration of prior work and theories.

Issue 5: Alternative Explanations

Before engaging in any research to evaluate a theory, we always make a point of stepping back and thinking about every possible alternative explanation to our theory and theory tests. We encourage you to engage in a final "thought session" in which you take a big picture view and try to identify any final alternative explanations that may undermine your theory test and then attempt to deal with them.

CONCLUDING COMMENTS AND A SUMMARY

We have described a long list of factors that can compromise research design and theory tests. Most of these factors are as applicable to qualitative and mixed methods research as they are to quantitative research. When we design studies, we use a "checklist" of the above factors that we systematically work through to ensure our research addresses every

one of them. We stress that there may be other issues that invariably come into play depending on the topic you are studying. Nevertheless, here is the checklist of factors for you to address that summarizes the contents of this primer:

1. In the spirit of Campbell and Stanley, address possible testing effects

2. In the spirit of Campbell and Stanley, address possible history effects

3. In the spirit of Campbell and Stanley, address possible instrument change

4. Address possible regression to the mean

5. In the spirit of Campbell and Stanley, address possible maturation effects

6. In the spirit of Campbell and Stanley, address possible experimental mortality effects

7. In the spirit of Campbell and Stanley, address possible selection effects

8. In the spirit of Campbell and Stanley, address possible interactions of the above (e.g., testing by treatment interactions)

9. If using a population-first, sample-second approach, address any factors that undermine obtaining a reasonable approximation to a random sample from the population

10. If using a sample-first, population second approach, address any factors that limit too much the population one can generalize to

11. Specify inclusion/exclusion criteria

12. Address possible self-selection into the study

13. Address possibly using too heterogeneous a population/sample

14. Address possibly using too homogenous a population/sample

15. If sampling weights are available or derived, address the appropriateness of their use

16. Address the reliability of your measures

17. Address any systematic error that may bias your measures

18. Address the validity of your measures

19. Ensure your measures provide an appropriate mapping of your constructs

20. Address whether your control group is appropriate

21. Address how to deal with possible non-random assignment to conditions

22. Address possible attrition bias, including treatment dropouts and assessment dropouts

23. Address possible carry-over effects

24. Address possible issues of assessment timing

25. Address possible experimenter bias

26. Address possible demand characteristics

26. Address treatment fidelity and manipulation effectiveness

27. Address possible contamination issues

28. Address possible range restriction issues

29. Address possible low task motivation and boredom effects

30. Address issues of ecological validity, as appropriate

31. Ensure you have good data management in place

32. Address issues of power, confidence interval width, asymptotic theory, and covariance stability for choosing a sample size

33. Address possible self-selection into the study

34. Address outliers

35. Address assumption violations of statistical tests

36. Address clustering, as applicable

37. Address missing data using modern methods

38. Address treatment dropouts using ITT, per protocol, or CACE analysis

39. Ensure analysis is metric appropriate

40. Address endogeneity issues

41. Address the many forms of aggregation bias, as appropriate

42. Address misspecification of functional forms

43. Avoid selective reporting and p hacking

44. Report effect sizes

45. Report indices of sampling error

46. Address familywise errors

47. Do not accept the null hypothesis (without equivalence testing)

48. Address the quality of your conceptual logic model

49. Ensure your constructs are clearly defined

50. Address issues of generalizing beyond one's data

51. Address the comprehensiveness and quality of the literature review and your interpretation of that literature and your theories

52. Address all alternative explanations

When evaluating research, you also can use this checklist as a framework for critiquing the quality of a study. If you invoke a given criticism, you should have a well-developed rationale for it, be able to make a convincing case that it is operative, and be able to state how it could potentially change study conclusions. It is not fair to "wave the wand" of a given criticism and dismiss a study without such elaboration. Your critique should be thoughtful, compelling, incisive, and meaningful. When critiquing studies, we also seek to be constructive and suggest ways to overcome shortcomings. We approach our reading of research by trying to take away the positives of it rather than focusing on the negatives, but also being sure to keep high scientific ideals in mind.

REFERENCES

Altaman, D. & Bland, J.M. (1999). How to use randomize. *British Medical Journal, 319*, 703–704.

Asparouhov, T. (2006). General multilevel modeling with sampling weights. *Communications in Statistics: Theory and Methods, 35*, 439-460.

Bentler, P.M. (1989). *EQS, structural equations, program manual, program version 3.0*. Los Angeles: BMDP Statistical Software, Inc.

Berger, V. (2005). Quantifying the magnitude of baseline covariate imbalances resulting from selection bias in randomized clinical trials. *Biomedical Journal, 47*, 119-127.

Bollen, K. (2012). Instrumental variables in sociology and the social sciences. *Annual Review of Sociology*, 38, 37–72.

Campbell, D. T., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand-McNally.

Clarke, K. (2009). Return of the phantom menace: Omitted variable bias in political research. *Conflict Management and Peace Science*, 26, 46–66.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey: Erlbaum.

Cook, T & Payne, M. (2002). Objecting to the objections to using random assignment in educational research. In F. Mosteller and R. Boruch (Eds.), *Evidence matters: Randomized trials in education research*. Washington, D.C.: Brookings.

Dahlke, J., Sackett, P. & Kuncel, N. (2019). Effects of range restriction and criterion contamination on differential validity of the SAT by race/ethnicity and sex. *Journal of Applied Psychology*, 104, 814-831.

DuMouchel, W.H. & Duncan, G. J. (1983). Using sampling survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association, 78*, 535-543.

Enders, C.K. (2010). *Applied missing data analysis*. New York: Guilford.

Feinberg, S. E. (1989). Modeling considerations: Discussion from a modeling perspective. In Kaspryzk, D., Duncan, G., Kalton, G., & Singh, M.P. (Eds.) *Panel surveys*. New York: Wiley.

Feinman, R. (2009). Intention-to-treat. What is the question? *Nutrition and Metabolism*, 6, 1-7.

Holmes, W.M. (2013). *Using propensity scores in quasi-experimental designs*. Newbury Park, CA: Sage.

Finney J. (2000). Limitations in using existing alcohol treatment trials to develop practice guidelines. *Addiction*, 95, 1491–1500.

Jaccard, J. (2018). Complex statistics made accessible: Confidence intervals, margins of error and precision analysis. Miami, Florida: Applied Scientific Analysis.

Jaccard, J. & Bo, A. (2018). Prevention science and child/youth development: Randomized explanatory trials for integrating theory, method, and analysis in program evaluation. Journal of the Society for Social Work and Research, 9, 651-687.

Jaccard, J., & Turrisi, R., (2003). *Interaction effects in multiple regression*. Newbury Park: Sage.

Jo B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics*, 27. 385–409.

Kalish L.A. & Begg, G.B. (1985).Treatment allocation methods in clinical trials a review. *Statistics and Medicine*, 4, 129–144.

Kenny, D. & Campbell, D. (1999). *A primer on regression artifacts*. New York: Guilford.

Kott, P.S. (1991). A model based look at linear regression with survey data. *American Statistician, 45*, 107-112.

Lakens, D., Scheel, A. & Isager, P. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1, 259–269.

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics, 7*, 405-424.

Lohr, S. L & Liu, J. (1994). A comparison of weighted and unweighted analyses in the National Crime Victimization Survey. *Journal of Quantitative Criminology, 10*, 343-360.

Maxwell, S. E., Kelley, K. & Rausch, J. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, 59, 537-563.

Mauro, R. (1990). Understanding LOVE (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin*, 108, 314-329.

Mee, R. & Chua, T. (1991). Regression toward the mean and the paired sample t test. *The American Statistician, 45*, 39-42.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.

Muthén, L.K. & Muthén, B,O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 4, 599–620.

Peugh, J., Strotman, D., McGrady, M., Rausch, J. & Kashikar-Zuck, S. (2017). Beyond intent to treat (ITT): A complier average causal effect (CACE) estimation primer. *Journal of School Psychology*, 60, 7-24.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, 61*, 317-337.

Pfeffermann, D. & Sverchkov, M. (2009). Inference under informative sampling. *Sample Surveys: Inference and Analysis, 29*, 455-489.

Reichardt, C.S. (2019). *Quasi-experimentation: A guide to design and analysis*. New York: Guilford.

Rosenberger, W. & Lachin, J. (2015). *Randomization in clinical trials: Theory and practice*. New York: Wiley.

Rosenthal, R. & Forde, K. (1963) The effect of experimenter bias on the performance of the albino rat. *Behavioral Science*, 8, 183-189.

Shadish, W. Cook, T. & Campbell, D. (2001). Experimental and quasi-experimental designs. New York, Houghton.

Suresh, K. (2011). An overview of randomization techniques: an unbiased assessment of outcome in clinical research. *Journal of Human Reproductive Sciences, 4*, 8–11.

Wellek, S. (2010). Testing statistical hypotheses of equivalence and noninferiority. New York: Chapman and Hall.

Wilcox, R. (2017). *Introduction to robust estimation and hypothesis testing*. San Diego: Academic Press (Fourth edition).

Winship, C. & Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods and Research, 23*, 230-257.

Wolf, E., Harrington, K. Clark, S. & Miller, M. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, 76, 913–934.