30
Research Design

Sharon Anderson Dannels

The definition of research design is deceptively simple: it is a plan that provides the underlying structure to integrate all elements of a quantitative study so that the results are credible, free from bias, and maximally generalizable. "Research design provides the glue that holds the research project together" (Trochim, 2006, Design, ¶1). The research design determines how the participants are selected, what variables are included and how they are manipulated, how data are collected and analyzed, and how extraneous variability is controlled so that the overall research problem can be addressed. Regardless of the sophistication of the statistical analysis, the researcher's conclusions may be worthless if an inappropriate research design has been used. Thus, design decisions both constrain and support the ultimate conclusions (Miles & Huberman, 1994).

Research designs may be identified as a specific design (e.g., a pretest-posttest control group design or a nonequivalent control group design) or by the broader category of experimental, quasi-experimental, or nonexperimental. *Experimental* designs are used in experiments to investigate cause and effect relationships. In contrast, *nonexperimental* designs are used in more naturalistic studies or in situations where the primary purpose is to describe the current status of the variables of interest. The latter designs are distinguished by the absence of manipulation by the researcher, with an emphasis on observation and measurement. Between these two broad categories are *quasi-experimental* designs which lack the randomization of *experimental* designs yet seek to address causal relations.

The adequacy of the research design to produce credible results, most notably to make causal inferences, is evaluated in terms of two primary types of validity: internal and external (Campbell & Stanley, 1963). *Internal validity* refers to the confidence that the specified causal agent is responsible for the observed effect on the dependent variable(s). *External validity* is the extent to which the causal conclusions can be generalized to different measures, populations, environments, and times. In addition, *statistical conclusion validity* is considered with internal validity and refers to the appropriate use of statistics. *Construct validity*, the ability to generalize the research operations to hypothetical constructs is a companion to external validity (Cook & Campbell, 1979).

Campbell and Stanley (1963) and Cook and Campbell (1979) produced the foundational works defining *quasi-experimental* design (see Desideratum 3) from which much of the literature on research design is extrapolated. Shadish, Cook, and Campbell (2002) revisited the initial works, providing greater attention to external validity, randomized designs, and specific design elements

Table 30.1 Desiderata for Research Design.

Desideratum	*Manuscript Section(s)*[*]
1. The research design is foreshadowed and follows logically from the general problem statement and associated research questions.	I
2. The research problem is clearly articulated and researchable.	I
3. The research design is appropriate to address the research problem and is clearly articulated.	M
4. Variables are identified and operationalized; sampling, instrumentation, procedures, and data analysis are detailed.	M, R
5. The research design is internally consistent (e.g., the data analysis is consistent with the sampling procedures).	M, R, D
6. The design is faithfully executed or, if applicable, explanations of necessary deviations are provided.	I, M, R
7. Extraneous variability is considered and appropriately controlled.	M, R
8. Potential rival hypotheses are minimized. Threats to internal validity and statistical conclusion validity, and the adequacy of the counterfactual, are considered.	M, D
9. Conclusions as to what occurred within the research condition are appropriate to the design.	M, D
10. Generalizations, if any, are appropriate. External validity and construct validity are considered elements of the design.	M, D
11. The limitations of the design are articulated and appropriately addressed.	D

[*] I = Introduction, M = Methods, R = Results, D = Discussion.

to be used as a counterfactual rather than prescribed research designs. A white paper prepared for the American Educational Research Association (AERA) by Schneider, Carnoy, Kilpatrick, Schmidt, and Shavelson (2007) specifically addressed the issue of causal inferences using large-scale datasets, experimental, and nonexperimental designs. Texts by Keppel, Saufley, and Tokunaga (1998), Fraenkel and Wallen (2006), or Huck, Cormier, and Bounds (1974) provide more thorough introductory treatments, whereas the text by Keppel and Wickens (2004) presents more advanced coverage. Table 30.1 contains specific desiderata to guide reviewers and authors as they make decisions regarding quantitative research design.

1. The Research Design Is Foreshadowed

From within a quantitative social and behavioral science research framework, the discussion of research design usually begins with the methods or procedures used to conduct the study (e.g., the selection and/or assignment of participants, the operationalization of the variables, the procedures for data collection and analysis) and the subsequent implications for conclusions. However, the research design and methods utilized should not come as a surprise at the end of the Introduction, but rather should be an extension of the foundation that has been developed therein.

In describing research design for qualitative research, Maxwell (2005) identified five components that comprise his model for research design, much of which is applicable, yet has remained only implicit, for quantitative research design. The five interacting components that Maxwell identified include the goals, conceptual framework (which includes the theoretical framework and research literature), research question(s), methods, and validity. Although Maxwell included elements within these that are typically not appropriate to quantitative research (e.g., the inclusion of personal experience within the conceptual framework), and he envisioned these elements dynamically interacting rather than the more sequential linear procedure of quantitative research, he did make explicit the need to evaluate the conclusions of a study within this larger context. It is in the Introduction that researchers should identify what variables will be attended to, which will be ignored, and which need to be controlled. The congruence of the Introduction, including its review of the literature, with the research design is necessary to evaluate the overall contribution of the study.

2. The Research Problem

It is impossible to evaluate the adequacy of a particular research design if there is no clearly articulated statement of the research problem. The research problem may be expressed in the form of research question(s) and/or hypotheses, and serves to formalize the research topic into an operational guide for the study, connecting the conceptual framework to the methods (Fraenkel & Wallen, 2006; Maxwell, 2005). The description of the research problem should identify the target population, the variables, and the nature of any anticipated relation between the variables and thereby focus the data collection and presage the data analysis. Hypotheses are not necessary, but are often stated when a specific prediction to be tested is made.

Terms used in the research problem must be defined in such a way that the questions are focused and testable. For example, "What is the best treatment for anxiety?" is not a testable question. Without defining "best" there is insufficient information to guide the study. Does "best" mean the most economical, the most consistent, or possibly the most permanent? The question also does not identify the population (e.g., children, teens, adults), or what types of treatment will be evaluated (e.g., psychotherapeutic, pharmacological, social behavioral), or what type of anxiety (e.g., self-report, clinically diagnosed, theoretically defined, physiologically measured). Without these further clarifications, it is not possible to assess whether or not the research design is appropriate to address the research problem.

Not only does the research problem suggest the appropriate research design, it also clarifies the specific type of data to be collected and thereby influences the data collection procedures. Questions can be classified as instrumentalist or realist (Cook & Campbell, 1979; Maxwell, 2005). *Instrumentalist* questions rely on the utilization of observable measures and require direct observation or measurement. *Realist* questions are about feelings, attitudes, or values that cannot be directly observed. The type of question, instrumentalist or realist, should connect the purpose of the study with the type of data collected. For example, if the purpose of the study is to provide information about teaching effectiveness, an instrumentalist question would be posed. It then would not be appropriate to collect the data using a survey within a survey research design to garner information from teachers as to their *perceived* effectiveness.

Designs developed for use with instrumentalist questions require greater inference and therefore might be more susceptible to bias. Yet as Tukey (1986) stated, and is often quoted, "Far better an approximate answer to the right question which is often vague, than an exact answer to the wrong question, which can always be made precise" (p. 407). Within the quantitative paradigm authors should clearly state their efforts to minimize bias and/or include appropriate caveats urging caution when interpreting and generalizing the results.

3. Articulation of the Research Design

The type of question(s), realist or instrumentalist (see Desideratum 2), will determine the type of data collected (e.g., self-report or performance). However, more fundamentally the research question(s) will determine the appropriate type of research design. Questions about relations among variables or questions about the current status of variables can be answered with a nonexperimental design. Experimental designs dominate the discussion when it comes to questions about cause and effect. However, alternative designs have challenged the notion that experimental designs are the only type of design appropriate for causal inference.

The types of research design are distinguished by the degree to which the researcher is able to control the research environment. Four types of control are evaluated: (a) the researcher's ability to control the selection and/or assignment of participants to groups, (b) the manipulation of the independent variable(s), (c) how any dependent variables are measured, and (d) the timing of the measurement(s). The types of research designs vary significantly with regard to the type of control that the researcher is able to exert. Nonexperimental designs offer very little control, and experimental designs require more control.

Questions of cause and effect should be addressed using experimental designs. They are *experiments* in the sense that the researcher is able to control or deliberately manipulate conditions in order to observe the varying outcomes. As Shadish et al. stated:

> Experiments require (1) variation in the treatment, (2) posttreatment measures of outcomes, (3) at least one unit on which observation is made, and (4) a mechanism for inferring what the outcome would have been without treatment— the so-called "counterfactual inference" against which we infer that the treatment produced an effect that otherwise would not have occurred.
>
> (Shadish et al., 2002, p. xvii)

Within the category of experimental designs, *randomized* designs (sometimes referred to as *true experimental* designs), are distinguished by the researcher's ability to control the experimental conditions, most specifically the random assignment of participants to conditions. *Quasi-experimental* designs comprise a separate category because, although the researcher can manipulate the proposed causal variable and determine what, when, and who is measured, he/she lacks the freedom to randomly assign the experimental units or participants to the treatment conditions. Without this random assignment the researcher must be more circumspect when making causal inferences (Cook & Campbell, 1979; Shadish et al., 2002).

In addition to the randomized and quasi-experimental designs, methodologists have referred to *pre-experimental* or *pseudo-experimental* designs (Cook & Campbell, 1979; Huck et al., 1974) as forms of experimental designs. These designs are separated from quasi-experimental because of their lack of experimenter control and subsequent weaker claims of causality. It is imperative that researchers and reviewers attend to how the various types of control (and more specifically, the lack of control) can impact both the internal validity (see Desideratum 8) and external validity (see Desideratum 10).

Questions of cause and effect require a comparison. The ideal, but impossible, comparison is the *counterfactual* (Cook & Sinha, 2006). Whereas the experimenter is able to measure what occurs when a treatment is introduced, he/she cannot say what would have occurred to that individual had the treatment not been introduced—the counterfactual. Thus, any experiment requires an approximation of the true counterfactual: "The better the counterfactual's approximation to the true counterfactual, the more confident causal conclusion will be" (Cook & Sinha, 2006, p. 551).

Specific experimental research designs are distinguished by how this counterfactual is constructed. Some designs use a control group and/or an alternate treatment, whereas others use a pretest measure to compare to the outcome measure. Some designs combine more than one approach (e.g., pre-test–post-test control group design) to improve the quality of the counterfactual. The logic of this approach is that the counterfactual represents what would be in the absence of the treatment. Unfortunately, this belief cannot always be justified. For example, the use of a control group presumes that this group is identical to the treatment group in all ways except for the existence of the treatment. Similarly, the use of a pretest presumes that all else remains the same, except the exposure to the treatment. Clearly, these assumptions cannot always be defended and the researcher should provide as much evidence as is reasonable to support his/her claims of the adequacy of counterfactual that serves as the comparative. Additional variables should be tested to further support the argument of equivalence of a control group to which the participants have not been randomly assigned.

Nonexperimental designs are usually restricted to descriptive or associational research, where the main purpose is at most to provide evidence of relations between two or more variables. However, there are nonexperimental designs, or naturalistic designs, used to explore causal relationships (e.g., *ex post facto* or *causal-comparative*) (Fraenkel & Wallen, 2006) and studies utilizing advanced statistical modeling procedures. Due to the absence of researcher control over not only the assignment of participants, but also the manipulation of a hypothesized causal agent, casual conclusions can only be tenuously advanced from studies that use a causal-comparative design. Studies that utilize some form of statistical modeling (e.g., structural equation modeling; see Chapter 33, this volume) rely upon an *a priori* theory and stochastic assumptions to make causal claims. The use of hierarchical data or multilevel modeling share the same methodological issues as other designs in addition to some unique issues, In their review of 99 studies using "traditional" multilevel models, Dedrick et al. (2009) identified four broad issues: model development, data considerations, estimation procedures, and hypothesis testing and statistical significance. Insufficient information about the procedures, reliance upon the data and statistical procedures to guide decisions, and the large number of statistical tests make the results difficult to evaluate (see Chapter 22 of this volume and Dedrick et al., 2009, for discussion and specific recommendations for reporting and evaluating studies using multilevel methods).

Although the ability to determine causality has traditionally been randomization, a number of statistical procedures are used to equate groups when random assignment is not possible (or to evaluate the equivalency of groups formed via random assignment). Propensity score matching uses logistic regression to assign each individual a score from a list of potential confounding variables used as the predictors. It is presumed that by matching groups on the span of their propensity scores the influence of the nuisance factors have been eliminated and the groups are equivalent. However, this assumes that all of the confounding variables have been included in the determination of the propensity score; an assumption that warrants evaluation. A second issue emerges when the span of scores for the groups only marginally overlap. Only participants with scores within the overlap should be included, which may have the potential of increasing the probability of a Type II error due to the reduced sample size. This also introduces a potential bias in that individuals who remain may not be representative of their group. Individuals with the lowest scores will be eliminated from the group with the initial overall lower propensity scores, whereas the group with the higher span will eliminate individuals with the highest propensity scores. There exist a number of methods for equating the groups, which suggests a lack of consensus:

> That there are so many alternative methods reflects the fact that no single approach is ideal and each has limitations. It is also disconcerting that, because each technique includes different subsets of people, it is quite possible to get different results depending on the choice.
>
> (Steiner & Norman, 2012, p. 1380)

A *regression discontinuity design* is used when the probability of treatment is determined by where one scores on one or more naturally occurring or arbitrarily defined threshold(s) (e.g., remedial course assignment below a set cut-score; eligibility determined by a specific age, geographic location, or income level). The assumption is made that those participants contiguous to the threshold are homogeneous and therefore the threshold arbitrarily serves to assign them to the control and treatment condition. The design is appropriate when the treatment is dichotomous with those on one side of the threshold becoming the control or comparison group and those on the other side comprising the treatment/experimental group. There are variations of the design (e.g., sharp versus fuzzy; Imbens & Lemieux, 2008), which have implications for the assumptions, data analysis, and interpretation of results. The design, when executed appropriately has strong internal validity. However, as with other quasi-experimental designs, the procedure does not ensure that the results are due to the treatment. Of particular concern is the possibility of selection bias due to an unidentified variable that is related to the threshold of the covariate or other changes at the cutoff (e.g., history threat). The validity of the regression discontinuity design is dependent upon the integrity of treatment assignment. The ability to manipulate an individual's score to alter the treatment assignment has serious implications as does compliance with treatment. The reviewer should evaluate the potential for misidentification, which if deemed extreme would invalidate the findings. Data analysis varies from relatively straight forward to very complex regression procedures, heuristically derived from visual inspection of the data. Attention to model specification is essential and evidence should be provided of efforts to assure unbiased and efficient estimates. Sensitivity tests should be conducted to assess the robustness of the results. Trochim (2006) presented an excellent introduction to the regression discontinuity design and analysis; Imbens and Lemieux (2008) provided coverage of designs and analysis with greater complexity.

The use of instrumental variables (IVs) occurs with both natural (observational) and randomized designs and with a number of statistical methods. By definition an IV must be correlated with the independent variable and uncorrelated with the error of the dependent variable; the effect on the dependent variable must be through the IV's relationship with the independent variable. Of critical importance is that use of an IV potentially changes the population to which the results can be generalized. The improvement in estimation is only for those who comply with the manipulation of the IV (i.e., the local average treatment effect, or LATE). In some cases, the research question of interest is addressed by the evaluation of the LATE, whereas for other questions the LATE represents only a subset (i.e., the compliers) of the population of interest. The ability to confidently determine the ratio of compliers to defiers has implications for the strength of the IV as well as the disparity between the LATE and the effect in the total population. If the treatment effects can be assumed homogenous then the LATE and the average treatment effect for the sample are the same. Sovey and Green (2011) suggested that although empirical evidence cannot be provided to assess the homogeneity, the researcher should present an argument based on the outcome of studies using different populations or IVs. An evaluation of the IV assumptions is required to determine the adequacy of the IV and whether an unbiased estimate of the effect has been achieved. In the randomized IV design (frequently an "encouragement design") the IV rather than the independent variable is randomly assigned. In nonexperimental studies IVs can be classified and evaluated on a gradient of plausible randomness (Dunning as cited by Sovey & Green, 2011). The randomness of the IV suggests that the IV is independent of the other predictors, however it does not guarantee that the IV effect on the criterion is only through the mediating factor. It is incumbent upon the researcher to defend that the IV satisfies both parts of the IV definition. Random assignment of the IV, a theoretical argument asserting the logic of the independence, as well as statistical tests to probe for potential correlations between the IV and other predictors can help to convince of the possibility that the IV is independent of unobserved variables related to the criterion. Whether or not the IV has a direct effect on the outcome should also be addressed by the researcher as he/she considers possible explanations. Weak instruments have the potential for bias and the strength of the IV should be tested. Although debated, the general guideline is that a single IV should have an

F greater or equal to 10. The effect of the IV must be monotonic (i.e., no one in the control condition receives the treatment). This again may depend on an argument rather than empirical evidence to justify why this assumption is met by the research design. The stable value treatment assumption (SUTVA) requires that the treatment or assignment of one unit not affect that of another. The use of more sophisticated sampling strategies (e.g., cluster sampling) may have implications for the satisfaction of this assumption. Four final considerations for the IV design: (a) the use of IV requires large sample sizes, (b) the estimated effect may be biased if the IV is not dichotomous, (c) the various statistical procedures also have assumptions that must be evaluated, and (d) different analyses models may result in different results.

Despite the statistical sophistication, methodologists remain divided on whether nonexperimental designs can provide convincing evidence that warrant claims of causality (see, for example, Shaffer, 1992).

Researchers who rely upon extant databases should be attentive to the quality of the original research design and how the design decisions impacted the data collected. For example, large datasets frequently result from sampling strategies that have implications for how the data should be evaluated (e.g., weighting). Researchers using existing data, including those performing statistical modeling, should (a) disclose information relevant to how the data were obtained, (b) provide sufficient detail of the *a priori* theory or theories, (c) faithfully execute the chosen statistical procedure after adequately addressing associated underlying assumptions, and (d) acknowledge the limitations of the study to make claims of causality.

The selection of the research design should consider the research problem within the larger context of the research topic. Careful consideration should be given to whether a longitudinal within-subjects design (see Chapter 2, this volume) or a cross-sectional between-subjects design (see Chapter 1, this volume) would be better suited to address the research problem. For example, either design can answer the question of whether or not there is a difference in performance on some defined measure of knowledge of teenagers and septuagenarians. However, if the hypothetical construct being measured is long term memory, the longitudinal design will enable greater confidence that the difference in test performance is due to memory rather than learning. If, however, the hypothetical construct represented by the test performance is learning, the less time and cost consuming, cross-sectional between subjects design would be adequate and consistent with research in this field.

Once determined, the research design helps authors to coordinate how participants are selected, how variables are manipulated, how data are collected and analyzed, and how extraneous variability is controlled. Discussion of specific designs can be found in Cook and Campbell (1979), Huck et al. (1974), Shadish et al. (2002), Trochim (2006), or Creswell (2005). Each element of the research design should be described with sufficient detail that the study can be replicated. All variables (i.e., independent, dependent, moderator, mediator, or control) should be defined, and the measurement should be congruent with the presentation in the Introduction. The type of design (i.e., experimental, quasi-experimental, or nonexperimental) or the specific design (e.g., groups × trials mixed between-within design, or nonequivalent control group design) should be stated. Adherence to a specific design is not required and the inclusion of additional procedures to control extraneous variability is encouraged (e.g., the inclusion of a pretest or a control group). In their follow-up text to Campbell and Stanley (1963) and Cook and Campbell (1979), Shadish et al. (2002) emphasized the value of design elements as counterfactuals rather than designs per se. In essence the design is constructed rather than selected from a prescribed list. The inclusion of each design element should be evaluated in terms of the potential impact on both internal and external validity (see Desiderata 8 and 10).

It is not uncommon for the researcher to omit any explicit reference as to what research design or design elements are used. Yet as Maxwell (1996, p. 3) noted, "Research design is like a philosophy of life; no one is without one, but some people are more aware of theirs and thus able to make more informed and consistent decisions." When design elements have not been explicated, the degree to which the researcher has made conscious design decisions is unknown. In this case, not only must the reviewer be vigilant in evaluating the credibility of what is reported, but he/she must also try to reconstruct the design that was used by what is reported in the Methods and Results sections. Without the aid of the author to define the counterfactuals used, the reviewer and reader are left to not only evaluate their effectiveness, but to also identify what they are. This is essential to determining whether the research design can support the stated conclusions.

4. Specific Design Elements

The first element of the research design is a description of the participants. The selection of participants should be consistent with the identified design. The type of design will determine first whether group assignment is necessary, and second, if so does assignment precede or follow selection. If a sample is used, the sampling frame and the population should be identified. The sampling procedure should be specified and there should be a justification of the sample size (see Chapter 35 this volume). An appropriate sample, in size and composition (i.e., representativeness), is foundational to the conclusions of the study (see Desideratum 9). In addition, how participants are assigned to treatment conditions (if appropriate) is important to the determination of the strength of any inference of causality (see Desideratum 8). Not only is it important to report what definition or instrument is used to select or assign participants, it is also important to report the reliability, validity, and cut scores of that instrument. This provides confidence that the participants met the criterion established and allows comparison to previous research. For example, "extraverts" as defined by the Eysenck Personality Inventory (EPI) are not the same as "extraverts" defined by the Myers–Briggs Type Indicator. Defining extraverts as those scoring above the sample mean on the EPI may not be the same as extraverts defined as those scoring above a normed score. The method by which participants are placed into groups (i.e., no groups, randomly assigned, or pre-existing groups) is essential to the type of research design being used and therefore to the conclusions that can be drawn (see Desiderata 8 and 9).

An integral element to the integrity of any study is the reliability and validity of the instrument(s) used to collect the data (see Chapter 29, this volume, for specifics on validity and reliability assessments). Not only is it necessary to provide evidence of the appropriate types of reliability and validity that have been established, but to make the case for why the author would expect that this evidence would apply to his/her use of the instrument. Citing extensive previous use is not sufficient evidence of reliability or validity.

The experimental and/or data collection procedures comprise the next element of the research design. There should be a detailed description of any experimental conditions, including any control conditions, if a treatment is introduced. This should include precise details of time intervals—duration of exposure to the treatment as well as time lapse between exposure and data collection, dosages, equipment settings, and research personnel. How and when the data are collected should be clearly described, making special note if the timing or the mode of collection could affect the response. For survey research designs this should include the number of reminder contacts, the timing of the reminders, and the mode of contact.

The final element of the research design before the discussion of study results is the presentation of the data management and data analysis. Data reduction and transformations, including the treatment of missing values should be articulated, highlighting any deviations from standard procedures. The data analysis should explicitly address demographic data that are useful for the discussion of appropriate generalizations (see Desideratum 10) or the equivalency of groups (see Desideratum 3). Data from instruments with total or scale scores should be analyzed for internal consistency reliability and compared to previous uses of the instrument. The specific test(s) used to address each research question and/or hypothesis should be named, including any information necessary for the reader to determine the appropriateness of the test or decision (e.g., degrees of freedom, alpha level, *p* values). *Post hoc* tests for the interpretation of omnibus test results (e.g., *post hoc* comparisons following an ANOVA) must be included and should be identified by name. When using samples, there should be a test at every point of decision. Look for words of comparison—most, greater, fewer, and verify that the appropriate statistical test has been conducted. When multiple tests are reported, consider the potential for an inflated Type I error rate. Evidence should be presented to confirm that the assumptions of statistical tests were met or that appropriate adjustments were made.

5. Internal Consistency of Research Design

A research design that lacks internal coherence creates problems for the interpretation of the results. This problem emerges particularly when the research design has not been explicated. Beginning with the Introduction, which should establish the need for the study and what precisely will be studied, through the statement of the research problem and research question(s), the way the sample is selected, the independent variable(s) are manipulated, the data collected, and how the data are analyzed, each design element should logically follow. If the researcher claims that a randomized design is used, it then follows that participants must be randomly assigned to the experimental conditions. If the research question is about differences between groups, the sampling plan must be such that ensures sufficient representation in each group and not left to random selection. The most blatant example of inconsistency is when the statistical analysis is not appropriate for the type of research question or how the data were collected (e.g., using a test of correlation to answer questions of cause and effect, especially when no temporal order has been established). Similarly, if participants are selected because they represent the two extremes of a grouping variable, it would be inappropriate to use correlation to evaluate the relationship between the two variables.

6. Design Execution

Details that researchers present in the Methods section must be consistent with what they intended at the outset of the research, as expressed in the Introduction. The procedures detailed in the Methods section should be evaluated to verify that they were faithfully executed. Small departures from the original design are often unavoidable—even anticipated, however, they require explanation.

A common problem is that the number of anticipated observations is not equal to the sample size upon which the conclusions are based, likely reducing the desired power level (see Chapter 35 this volume). This issue is particularly prevalent in survey research. Even in studies where researchers over-sample in an effort to achieve the desired sample size, the total response rate is often less than desired; the response rate for an individual survey item (i.e., missing response) may be considerably lower (Jackson, 2002). Often, studies are designed with equal or proportional sample size in each cell, yet frequently when the results are reported the cells are uneven. This has implications for how missing data are treated, for statistical assumptions, for the power of the test, as well as for other design implications. It is therefore incumbent upon the researcher to account for missing values and evaluate the implications for the design. One consideration is whether the nonresponses adversely affect the representativeness of the sample. More specifically, consideration must be given to whether missing values represent a threat to internal validity (see Desideratum 8). The disproportionate loss of participants from one treatment condition might suggest a threat to internal validity (Cook & Campbell, 1979). This is not only true of quasi-experimental designs but also of randomized designs, which by virtue of random assignment of participants are protected from most other threats.

Reviewers of manuscripts should be vigilant for evidence suggesting that procedures were counter to stated claims. For example, if a researcher claims that participants were randomly assigned, but then later suggests that the treatment was assigned to pre-existing groups, this changes the design from a randomized design to a quasi-experimental design with all the attendant issues that must be addressed.

Valid conclusions about a causal relation between treatment and outcome are dependent upon the treatment (and control) condition(s) being faithfully delivered and the dependent variable reliably measured. There should be evidence that the researcher (or whoever is providing the treatment) has been trained and dependably delivers the specified treatment. Evidence in the form of manipulation checks should be provided to verify that experimental manipulations were effective. For example, experiments that rely on deception require that the participants are indeed deceived, and a well-designed study will provide evidence to this effect. In addition to ensuring that the experimental conditions are consistent with what is reported there should be evidence that the researcher has sufficient training to collect the data (e.g., training for interviewers, inter-rater reliability).

In addition to considering whether the researcher has delivered treatment successfully, the researcher and the reviewer should consider the plausibility of participant noncompliance with treatment. Drug trials are dependent upon participants actually consuming the prescribed dose; training is dependent upon participant attendance. Without evidence of participant compliance there is insufficient evidence that the research design has been implemented.

7. Control of Extraneous Variability

Without appropriate control of extraneous variability it can be difficult to isolate and observe the effect(s) of the hypothesized causal variable(s) on the dependent variable(s). Control of extraneous variability is therefore one of the primary functions of a research design. Rigorous adherence to carefully designed research procedures can help to minimize the effect of unintended influences. However, the design element that has the greatest impact on the control of extraneous variability is the selection and/or assignment of participants. Random assignment to treatment conditions is the principal means by which a research design avoids the systematic influence of unintended variables. The advantage of this method is that it controls for the influence of a number of variables, even those unidentified. However, it alone might be insufficient if an extraneous variable has a stronger effect than the causal variable that is being considered. Manuscript reviewers should note any mention of one or more variables in the Introduction (or from previous content knowledge) that is known to have a strong relation to any of the dependent variables, and ensure that its influence is considered in the research design chosen by the study's authors. In fact, an alternative research design might have been more appropriate. For example, the effect of an extraneous variable might be controlled by including it as an additional variable in the design (i.e., randomized block design) or by restricting the population of the study to only one level of the extraneous variable (e.g., only include women in the study).

Matching is a procedure whereby participants are paired on their scores for a specific variable(s) and then each member of the pair is assigned to a different treatment condition. The intent of using this procedure is to equate the groups in terms of this specific variable, a variable that is believed to influence the dependent variable. This procedure should be used judiciously. Although it might equate the groups on that one specific variable, matching interferes with the ability to randomly assign participants, and thereby forfeits the benefit of randomization. The implications for the data analysis also must be considered. The matched pairs cannot be treated as independent observations and the data analysis must reflect this. The use of these procedures must be considered within the larger context of the overall design to ensure that their use is reflected in other design decisions (e.g., data analysis) and conclusions.

Sometimes, statistical procedures can also be used to control extraneous variability. The use of covariates can help adjust the scores on the dependent variable before testing for group differences if the extraneous variable is measured as a continuous variable. Propensity scores from a logistic regression (see Chapters 16 and 28, this volume) might also be used to evaluate the equivalence of treatment groups, improve matching, and/or be used as a covariate (Pasta, 2000). Although the use of propensity scores is a means of controlling for the effect of more than one extraneous variable, it is still limited to the control of only those variables that are identified and quantitatively measured. Rather than attempt to improve the equivalence of groups, Rosenbaum (1991) suggested a procedure (hidden bias sensitivity analysis) to assess how much bias would be necessary between the treatment and control groups for bias to be a viable alternative explanation for the treatment outcome. Shadish et al. (2002) warned that the use of these, or other advanced statistical procedures, is not a substitute for good design. Where possible, extraneous variability should be controlled by the research design, and then if appropriate augmented by available statistical procedures.

8. Internal Validity and Statistical Conclusion Validity

The careful construction and faithful execution of the research design provides the foundation for the research conclusions. Each element of the design relates to the validity of the study. Research conducted to test causal relations relies on the adequacy of the constructed counterfactual to represent the true counterfactual (see Desideratum 3). The adequacy is evaluated in terms of the ability to rule out rival hypotheses or alternative explanations for the outcome. In 1957, Campbell first coined the term *internal validity*, which was further elaborated by Campbell and Stanley (1963) as the confidence that the identified causal variable is responsible for the observed effect on the dependent variable and not due to other factors. They identified a list of threats to internal validity, which should be considered when constructing the design as well as when evaluating the conclusions. The list of threats to internal validity, with some modifications, can be found in most research design textbooks (also see Shadish et al., 2002; Shadish & Luellen, 2006; for discussions on threats relevant to specific designs see Cook & Campbell, 1979; Huck et al., 1974).

Threats to internal validity are usually discussed in terms of quasi-experimental designs because they result from the inability to randomly assign participants to treatment conditions. That is, random assignment reasonably protects the study from most threats to internal validity; however, such threats should be considered for any study that seeks to make causal inferences, with or without random assignment. Some threats (e.g., mortality or attrition, the disproportional loss of participants from one condition) occur after the assignment to experimental condition or as a result of something that occurs during treatment delivery, which thereby jeopardize causal interpretations of even a randomized design.

The potential of a threat to internal validity in and of itself is insufficient to dismiss a researcher's claims of causality. When evaluating the potential threats, Shadish and Luellen (2006, p. 541) advocated the consideration of three questions: "(a) How would the threat apply in this case? (b) Is the threat plausible rather than just possible? and (c) Does the threat operate in the same direction as the observed effect so that it could partially or totally explain that effect?" If it can be conceived how a specific threat would offer a rival hypothesis, which is probable—not just possible, and explains the direction of the outcome, only then would the internal validity be challenged. The careful researcher will consider these threats in the design of the study, anticipating those with potential relevance. If considered prior to the execution of the study, it may be possible to alter a design element(s) to avoid a potential threat, or additional data may be collected to provide evidence to argue against a threat's explanatory ability (see Desideratum 3).

Cook and Campbell (1979) further refined the discussion of internal validity by introducing *statistical conclusion validity* as a distinct form of validity related to internal validity. Statistical conclusion validity refers to the "appropriate use of statistics to infer whether the presumed independent and dependent variables covary. Internal validity referred to whether their covariation resulted from a causal relationship" (Shadish et al., 2002, p. 37). Threats to statistical conclusion validity provide reasons why the researcher might be wrong about (a) whether a relationship exists and (b) the strength of the relationship. A list of threats to statistical conclusion validity can be found in Shadish et al. (2002, p. 45). Attention to statistical power, assumptions of the statistical tests, inflated Type I error, and effect size, as well as issues related to the measurement and sampling, fall within the purview of statistical conclusion validity.

Generally, it is not appropriate to refer to the internal validity of nonexperimental designs, with one exception: specific designs that are being used to make causal inferences (e.g., causal comparative). However, the validity of conclusions reached still requires evaluation. Each design decision affects the validity, with decisions regarding the appropriate sampling, instrumentation, and statistical analysis of particular importance for the nonexperimental design. The sample size and representativeness of the population, the reliability and validity of measurement, and the appropriate statistical analysis are key to the conclusions of a nonexperimental study.

Authors and reviewers must keep in mind that "Validity is a property of inferences. It is *not* a property of designs or methods, for the same design may contribute to more or less valid inferences under different circumstances" (Shadish et al., 2002, p. 34). Executing a prescribed design does not guarantee valid inferences, nor does the rigid adherence to a checklist of potential threats to validity. Neither are adequate substitutes for the researchers' sound logic.

9. Conclusions Are Appropriate

Miles and Huberman (1994) noted that design decisions both support and constrain the conclusions of research. Just as the genesis of the research design is before the Methods section, its influence extends beyond the Results. Researchers are responsible for presenting conclusions that are consistent with and appropriate to the design. The adage "correlation is not causation" is just one example for the necessity to ensure that claims in the Discussion do not exceed what the research can support. Design decisions, such as the decision to control extraneous variability to only one level of an extraneous variable (e.g., women only) restrict the conclusions to only that group.

Careful articulation of the research design elements, with attention to potential threats to internal and statistical conclusion validity (see Desideratum 8), prepares the researcher to present the conclusions within the context of the existing literature. Causal claims should not be made without ruling out threats to internal validity. With a nonexperimental design utilizing only descriptive statistics to report the findings from a sample, it is inappropriate to make comparisons between groups (e.g., "women scored higher than men"). There must be a test at the point of decision.

Without appropriate supporting evidence it is inappropriate to draw conclusions from statistical nonsignificance. For example, when testing for mean differences between treatment populations, nonsignificance should not be interpreted to imply that there is no difference between the populations' means or that the population means are therefore equal. Statistical nonsignificance means that the researcher has failed to show a difference of sufficient magnitude that cannot be reasonably explained by chance alone. That the population means are equal is only one possible explanation. It is also possible that the sample size was insufficient or the measurement not sensitive enough to detect true differences.

The tendency to overstate findings is not limited to misrepresenting statistical conclusions or failing to recognize threats to internal validity, but also includes making claims beyond what was studied. For example, if a study using a survey to measure the level of teacher satisfaction shows that 65% of teachers report being *slightly dissatisfied* with teaching, it is inappropriate for the researcher to conclude that his/her study found that teachers will be leaving their schools, or that teachers should be paid higher salaries. The author should be diligent to ensure that recommendations from the study are not presented in a manner that they can be construed as findings.

10. External Validity and Construct Validity

Technically, the *external validity* of a research design refers to the degree to which a study's observed *causal* relations are generalizable; that is, it helps characterize "to what populations, settings, treatment variables, and measurement variables can this effect be generalized" (Campbell & Stanley, 1963, p. 5). Internal and external validity are considered to be complementary: whereas the former addresses the question of what can be inferred about cause and effect from this instance, the latter assesses the degree to which the causal findings can be generalized. Frequently what will increase internal validity may decrease external validity and vice versa. In their 1979 work, Cook and Campbell extended their dichotomous discussion of validity into the typology that comprised internal, statistical conclusion, external, and construct validity. Whereas internal and statistical validity (see Desideratum 8) are relevant to the inferences that derive from the specifics of the study procedures, construct and external validity relate to whether the inferences can be extended beyond the current situation. *Construct validity* generalizations refer to "inferences about the constructs that research operations represent" and external validity generalizations are "inferences about whether the causal relationship holds over variation in persons (or more generally: units), settings, treatment, and measurement variables" (Shadish et al., 2002, p. 20). From these definitions it becomes apparent that with nonexperimental designs that are used to describe the current status or noncausal relations between variables, it is inappropriate to discuss external validity. Instead, construct validity is the more appropriate consideration. Thus, nonexperimental designs that are not used to test causality should be evaluated for construct validity, and nonexperimental designs that are used to evaluate causal relations and experimental designs should be evaluated for both construct and external validity.

Construct validity is inherent in social and behavioral research and as an issue is twofold: definition and measurement. Every construct has multiple facets or features, with some being more central than others. Thus, defining the construct requires identifying multiple components, with the core being those features to which there is the greatest agreement. Once defined, the question becomes one of how to represent the construct, and more specifically how to measure it. Determining how multi-faceted constructs can be reduced to a manageable size, yet still represent the higher order construct, is the dilemma of construct validity. Each study uses a limited set of conditions in terms of the population, the treatment, the setting, and the outcome; from which the desire is to make statements about the higher order construct. Each element of the research design should be evaluated for the construct(s) it represents. Researchers tend to focus on only the treatment variable, if there is one, and the outcome measure. Clearly, discussions limited to the construct validity of the outcome measure are insufficient as they address only one of the constructs in the study. How will the sample selected reflect the larger construct that it represents? For example, how does a sample consisting of students two grades below reading level represent a population of "students at risk"?

How does conducting the study in the laboratory represent the larger construct of the settings where the conclusions would apply? These questions need to be considered in addition to the more obvious examination of how the treatment and outcome constructs are operationalized. There is no one-to-one correspondence of the operationalization of the study and the constructs; the question is: *How great is the disparity?* A list and further discussion of potential threats to construct validity is presented in Shadish et al. (2002).

External validity refers specifically to whether or not observed *causal* relations can be extended across individuals, settings, treatments, and/or outcome measures. The use of probability sampling is the foundation for external validity. Probability sampling requires that each item in the domain has a nonzero chance of being randomly selected. This condition is infrequently met when sampling participants for a study, much less when sampling from the domains that describe the other elements of an experiment (i.e., the setting, treatment conditions, outcome measures). Shadish et al. (2002) explicated a more heuristic approach to determine causal generalizations. They proposed five principles for consideration: surface similarity, ruling out irrelevancies, making discriminations, interpolation and extrapolation, and causal explanation. Too frequently external validity is only discussed in terms of generalizing to populations, either those internal or external to the study, and the ability to generalize to the other elements receives scant attention. With a design seeking to establish evidence of a causal relationship, the researcher and reviewer should examine the degree to which the design elements that are included represent a random sampling of the construct domain, be it the population, setting, treatment, or outcome. In the absence of random sampling, the principles described by Shadish et al. (2002) provide a systematic means to evaluate external validity. Shadish et al. also present a list of common threats to external validity.

11. Design Limitations

The diligent researcher will acknowledge weaknesses in the research design and present the implications of the shortcomings. For example, by recognizing in advance that the use of pre-existing groups compromises the internal validity (see Desideratum 8), the researcher has the opportunity to offer explanations, possibly even statistical evidence (see Desideratum 7), to argue the equivalence of the groups prior to the introduction of the treatment. By ignoring any reference to this design decision, the reviewer and reader are left to decide whether the potential pre-existing group differences are sufficient to explain the outcome. More significantly, this can create a lack of confidence. The question becomes: *If the researcher does not know enough to discuss the implications of the use of pre-existing groups, what other relevant information might he/she not recognize the necessity to reveal?* Does the researcher understand enough about the research design to adequately convey the information necessary for the reader to make an independent decision as to the appropriateness of the conclusions?

In theory, many of the weaknesses can be avoided by assiduous attention to the research design, yet this is not always the case. Weaknesses result from a lack of feasible alternatives, unforeseen occurrences during the study, and/or from poor research design. The credibility of the researcher is enhanced if he/she is able to eliminate him/herself from the latter category by anticipating and addressing criticism from the knowledgeable reviewer or reader.

References

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago, IL: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Cook, T. D., & Sinha, V. (2006). Randomized experiments in educational research. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 551–565). Mahwah, NJ: Erlbaum.

Creswell, J. (2005). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, NJ: Prentice Hall.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, *79*, 69–102.

Fraenkel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research in education* (6th ed.). New York: McGraw Hill.

Huck, S. W., Cormier, W., & Bounds, W. G. (1974). *Reading statistics and research*. New York: Harper & Row.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*, 615–635.

Jackson, G. B. (2002). *Sampling for social science research and evaluations*. Retrieved from www.gwu.edu/~gjackson/281_Sampling.PDF

Keppel, G., Saufley, W. H., Jr., & Tokunaga, H. (1998). *Introduction to design and analysis* (2nd ed.). New York: Freeman.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). New York: Pearson Prentice Hall.

Maxwell, J. A. (1996). *Qualitative research design: An interactive approach*. Thousand Oaks, CA: Sage.

Maxwell, J. A. (2005). *Qualitative research design: An interactive approach* (2nd ed.). Thousand Oaks, CA: Sage.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage.

Pasta, D. J. (2000). Using propensity scores to adjust for group differences: Examples comparing alternative surgical methods. In *Proceedings of the twenty-fifth annual SAS Users Group International Conference* (paper 261–25). Cary, NC: SAS Institute.

Rosenbaum, P. R. (1991). Discussing hidden bias in observational studies. *Annals of Internal Medicine*, *115*, 901–905.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). Estimating causal effects using experimental and observational designs: A think tank white paper. Washington, DC: Prepared under the auspices of the American Educational Research Association Grants Program.

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

Shadish, W. R., & Luellen, J. K. (2006). Quasi-experimental design. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research*. (pp. 539–550). Mahwah, NJ: Erlbaum.

Shaffer, J. P. (Ed.). (1992). *The role of models in nonexperimental social science: Two debates*. Washington, DC: American Educational Research Association and American Statistical Association.

Sovey, A. J., & Green, D. P. (2011). Instrumental variables estimation in political science: A readers' guide. *American Journal of Political Science*, *55*, 188–200.

Steiner, D. L., & Norman, G. R. (2012). The pros and cons of propensity scores. *Chest*, *142*, 1380–1382.

Trochim, W. M. K. (2006). Design. Retrieved from www.socialresearchmethods.net/kb/design.php.

Tukey, J. W. (1986). *The collected works of John W. Tukey: Philosophy and principles of data analysis 1949–1964, Volume III* (Ed. L. V. Jones). Boca Raton, FL: CRC Press.