Chapter 7 Evaluation Research Design: Critical Issues

All research designs are subject to bias and design threats which can adversely affect the internal validity (i.e., accuracy) and external validity (i.e., generalizability) of study results from the sample back to its parent population. The evaluation research designer must also ensure statistical and social validity.

Bias can enter the evaluation research design process as a result of researcher and/or subject preconceptions or behaviors; these preconceptions or behaviors can (intentionally or unintentionally) distort study findings. Researcher preconceptions can include beliefs about the subject, positive or negative; stereotypes, prejudices, cultural incompetence or insensitivity, etc. We tend to be more concerned about researcher bias than subject bias. Most subject bias can be engineered out of the evaluation design by fully informing subjects of the purpose of the study, any benefits or potential risks to them, and stressing the importance of their full participation.

In addition to being as free from researcher or subject bias as possible, an evaluation study must be internally valid so that attributions about the behavior of or changes in the dependent variable can be confidentially asserted to be due to the independent variable and not bias or internal validity design defects. Internal design validity threats may affect studies using either a probability or purposeful sampling strategy. A sampling strategy is the plan used to recruit subjects or sampling units (anything studied that isn’t human) into the study. The threats to internal and external design (or experimental) validity are drawn chiefly from Campbell and Stanley (1963). Integrated into the Campbell and Stanley schema are selected threats advanced by Martella, Nelson, and Marchand-Martella (1999, pp. 135-36, 146-47, & 154-55).

If we have drawn a sample from a parent population and we intend to generalize from the sample back to that population, we must also be concerned about external design validity, more commonly referred to as generalizability. External validity is a concern for studies using a probability sampling strategy; studies using a purposeful sampling strategy should not generalize their findings.

Internal and external validity design threats must be controlled. By controlled, we mean that the effects of any internal design validity and/or the external validity threats are equalized (if two or more groupings are involved), mitigated or eliminated to fullest extent possible.

Following is a discussion of specific research biases, internal design validity threats, external design validity threats, statistical and social validity, and sampling.

I. Bias and Control in Evaluation Research Design

A. Every sample (the group of people, things, places, etc. which is studied) has some degree of bias or error introduced into the research design or process.

1. Bias is defined as any single factor (e.g., sampling error) or combination of factors (e.g., sampling error and uncontrolled internal validity design threats) which distorts data from what should have been obtained under optimum circumstances.
2. The researcher's intent is to keep bias, regardless of source, as low as possible. Bias may stem from several sources.
   a. A <u>researcher</u> may have a pre-disposition to favor or disfavor a particular group of subjects or a preference for a particular outcome. Independently or combined, these biases (intentional or not) may affect management of the research project, data interpretation, or recommendations.
   b. <u>Subjects</u> can be a source of bias because of how they were selected (sampling), their behavior during the research process, etc. Remember, we are comparing what happens against an ideal, which we may or may not be able to actually know. Thus, it is essential to properly follow the prescribed research methodology.
   c. The research question should drive <u>research design</u> selection. If the wrong design is selected (i.e., the selected design will not produce data to answer the research question or test the hypothesis), useful data are unlikely.
   d. A <u>research plan poorly executed</u> (i.e. the manner in which the study was conducted) will possess little internal design validity and external validity or generalizability. In short, garbage in is garbage out.

B. Biases Frequently Encountered in Evaluation Research

1. Selection Bias
   a. Socio-behavioral research relies primarily on volunteers and referrals. Selection bias assumes that a sample has been selected and the intent of the study is to generalize the findings or results from that sample back to its parent population. Care must be taken to ensure that any treatment, control or comparison groups are actually representative of the population to which findings or results will be generalized. Generalization requires a randomly selected sample. See the Section on Probability Sampling in this chapter.
   b. Another selection bias is subject attrition, i.e., dropping out of a study in such a way that the treatment, comparison, or control group isn't any longer representative of its parent population; this applies only to designs using a probability sampling strategy. Subject attrition in survey research is called non-respondent or non-response bias. Either of these biases may operate or exert an effect in action research as well.

2. Common Bias
   a. <u>Central Tendency bias</u> is the trend to rate or score each subject towards the center of a rating scale. For example, on a scale from 1 to 10, 90% of the subjects are rated between 4 and 6. This bias is most common when

performance assessment checklists are used. This bias may also be seen in self-rating scales completed by subjects.

b. Leniency/Severity bias occurs when a rater (researcher or subject) is either extremely lenient or severe in rating performance or attitudes typically using Likert style scales. In other words, the "true or usual" rating is not given. Leniency bias is related to Social desirability bias.

c. Social Desirability bias operates when a respondent reports or rates behavior, feelings, or attitudes, because he or she wants to either please the researcher or appear to possess desirable characteristics.

d. Stereotyping bias operates when a researcher or subject allows opinions or feelings about a group to affect his or her responses, ratings, or observations.

3. The Expectancy (Pygmalion) Effect

a. The Pygmalion effect is the proverbial self-fulfilling prophesy or expectation bias. Suppose a researcher was evaluating the effects of instructional technology on improving the academic performance of school children from poor, rural families. Let's say, the children's teacher communicated to the students that because they were disadvantaged, he didn't think they would perform well regardless of the technology. At the conclusion of the study, we found that the students did not perform well. One possible explanation for this under-performance could be the expectation from the teacher.

b. The moral is that as evaluation researchers, we must be cognizant of both the conscious and the subtle messages we send.

4. The Hawthorne or Halo Effect

a. This is also called guinea pig, novelty, or "the-gee-whiz" effect. The Martella, Nelson, and Marchand-Martella (1999, p. 48) definition is similar. If research participants change their behavior, attitudes, self-reported learning, etc. so as to affect the dependent variable (think back to chapter 2) just because they are involved in a research project, then the Hawthorne or Halo Effect is said to be operating.

b. There was a series of studies conducted in the 1920's by Elton Mayo at a General Electrical wiring plant. Dr. Mayo's team found that regardless of how working conditions were changed, the workers produced high numbers of wired harnesses. When asked about the continually increasing production levels, the workers (both the treatment and control groups) reported that they increased production levels because they were in the study and considered the study to be important. Thus, Dr. Mayo was unable to ascertain the effects on production levels of various working conditions, as subjects deliberately increased production partly because they were in a study.

5. The John Henry Effect
   a. The John Henry Effect occurs when the control group (e.g., Figures 8.6 or 8.7) learns it's in a research project and then performs differently in order to deliberately achieve the desired effect on the dependent variable. Subjects didn't behave as they "would normally;" they tried to be like the experimental group. Martella, Nelson, and Marchand-Martella (1999, p. 39) refer to this effect as compensatory rivalry by the control group and compensatory equalization of treatments.
   b. In other words, the control group attempts to emulate or mimic the experimental group. Thus, the researcher is unable to clearly assess the effect of the independent variable or treatment on the dependent variable.

C. Bias Control Strategies

1. The control strategies for these biases depend on the sampling strategy used to recruit subjects for the evaluation study.
   a. For studies using a probability sampling strategy, large representative samples, random selection of subjects, and random assignment of subjects to either the experimental or control groups are primary control tools.
   b. For studies using a purposeful sampling strategy, employment of a comparison group (e.g., Figures 8.3 or 8.4) is helpful.
   c. Regardless of sampling strategy, careful research design, "clean" instrumentation, careful study project management, and a rich (i.e., detailed) description of the sample and research methodology, will document and report whether or not these biases operate and alert the reader to interpret data with appropriate caution.

2. Blinding
   a. One common control strategy for many of the biases is called blinding.
   b. In blinding, to the maximum extent possible, subjects, data collectors, raters, etc. are kept ignorant of (1) the research question or hypothesis, (2) which group (e.g., treatment or control; see Figures 8.6 or 8.7) a subject is in, and (3) what treatment (e.g., program) is being administered, etc.
   c. Subjects know they are participating in a research project and have given their consent, through a process called "Informed Consent," but only know enough about the project to meet their participation responsibilities.

II. Internal and External Evaluation Design Validity

A. Evaluation research designers must design studies in such a way as to ensure internal design validity as well as external design validity, where generalization of results from a sample back to its parent population will occur. Specific internal and external design threats are summarized in Table 7.1.

1. There are two categories of sampling strategies (see the discussion on sampling below) which are important to our present discussion; where probability samples are drawn, external design validity is critical and where purposeful sampling strategies are used, external design validity is of less importance, as there is usually little interest in results generalization.

2. However, regardless of the sampling strategy used to recruit study subjects, internal design validity is critical.

3. The Evaluation Study Designer's Responsibilities
    a. Studies employing a probability sampling strategy (see sampling discussion below) are assumed to have controlled (i.e., equalized, minimized, or eliminated) internal and external design validity threats to the maximum extent possible. The evaluation research designers' responsibility is to carefully review all aspects of the study to ensure that no undocumented and unreported internal or external design validity threats have crept into the project.
    b. Studies employing a purposeful sampling strategy (see sampling discussion below) are likely to experience internal design validity threats.
        (1) With respect to internal design validity threats, the evaluation designer should make a conscious effort to ferret out those specific threats likely to operate (i.e., potentially effect change in the dependent variable which is unrelated to the independent variable), so that their existence and their potential influence is identified. Armed with such information, an evaluator or stakeholder is appropriately cautioned when interpreting study data.
        (2) If there is no intention to generalize study findings beyond those subjects studied, then there is no need to be concerned about external design validity threats; the evaluation designers' responsibility is to indicate that there is no intention to generalize results and to caution others against doing the same.

Table 7.1

Internal and External Validity Design Threat

Threat Category	**Specific Threat**
Internal Design Threats	History, Maturation, Testing, Instrumentation, Statistical Regression, Selection, Mortality or Attrition, Resentful Demoralization of the Control or Comparison Group, Diffusion of Treatment
External Design Validity Threats	The Reactive or Interactive Effective Testing; The Interaction Effects of Selection, Instrumentation, History, or Maturation biases and the Treatment (or Independent) Variable; Reactive Effects of Experimental Arrangements; Multiple Treatment Interference; Experimenter Effects
Other Threats	Statistical Validity; Social Validity

B. Internal Design Validity

1. Internal Design Validity: The Importance
   a. Evaluation studies high in internal validity produce valid and reliable data.
   b. Specific threats to internal evaluation study design validity are history, maturation, testing, instrumentation, statistical regression, selection, mortality, resentful demoralization of the control group, and diffusion of treatment. These threats potentially apply to studies using either a probability or purposeful sampling strategy.
   c. The extent to which any of these threats operate within a study is the extent to which the internal design validity of the study is threatened; any one or combination of these threats may affect the dependent variable so that it becomes difficult, if not impossible, to accurately measure the effect of the independent variable on the dependent variable, rendering the study potentially useless.

2. Specific Internal Design Validity Threats
   a. <u>History</u> includes events which occur between observations or measurements of the dependent variable, which affect it, aside from the independent variable (Campbell & Stanley, 1963, p. 5; Martella, Nelson, & Marchand-Martella, 1999, p. 39).
      (1) Two Examples
         (a) An 8 year study of “naturally occurring” heart disease incidence (i.e. the expected number of heart disease cases that would "normally" be expected in study subjects) is likely to be effected by history. For example, the death of close friends may prod a study subject to change his or her diet, or encouragement by a spouse to live a much healthier lifestyle may cause study subjects to change dietary and exercise behaviors, thereby reducing heart disease risk and the number of cases.
         (b) Suppose an evaluator was studying the effects of a math education program (i.e., the independent variable) on underperforming 10th-graders. The study was to be completed over a nine-month academic year. There was a four-month teachers’ strike. It is likely that being deprived of four months’ math instruction will adversely affect these 10th-graders’ math performance (i.e., the dependent variable).
      (2) Research studies conducted over an extended period of time in a dynamic environment are more likely to be affected by history than those studies which are more quickly completed and/or in “stable” environments. An overly long or short study time line can be affected by “Timing Bias,” as the duration of the study may affect the results; to avoid, the researcher must be very familiar with the professional and empirical literature in order to know the generally accepted timelines associated with similar research.

(3) Some study subjects will fully comply with their study obligations, others will not; this is "Compliance Bias." Differential compliance with study protocols may affect study results.
(4) Subjects who are enrolled in other studies or who on their own engage in experiences which affect the independent variable or dependent, trigger what is called "Co-intervention or Parallel-intervention" bias.

b. <u>Maturation</u> includes emotional, psychological, or physiological processes, <u>within</u> study subjects which (across time) somehow affects the dependent variable (Campbell & Stanley, 1963, p. 5; Martella, Nelson, & Marchand-Martella, 1999, p. 39).
   (1) Two Examples
      (a) Studies of pre-school and school-aged children may be affected by the normal developmental processes experienced by these children. These effects may make it difficult to accurately measure the effect of the independent variable on the dependent variable.
      (b) For example, suppose a new method of teaching children to form letters is advanced. Any study of its effectiveness, over current instructional methods, must consider increased psychomotor coordination as a potential moderating variable. As children mature, they naturally become more coordinated and as a consequence are able to form letters more accurately.
   (2) Probability sampling strategies are likely to control (i.e., equalize, minimize, or eliminate) any effect of this threat in studies that use one (see the sampling discussion below). For studies using a purposeful sampling strategy, the researcher should report in detail any changes in either the treatment group or comparison group.

c. <u>Testing</u> is taking a test or completing an assignment or task a second or third or more time; the experience is likely to affect subjects' performance or recall on subsequent testing (Campbell & Stanley, 1963, p. 5). In other words, taking a test over the same material two or three times will help the examinee perform better on the test at time 4 or 5.
   (1) Research designs with a pretest may artificially increase scores on the posttest due to the experience of the pretest; this is called pretest sensitization.
   (2) Testing is likely to be a potential internal design validity threat in designs with a pretest (e.g., Figures 8.2, 8.4, or 8.6) or any time series design (e.g., Figure 8.5).

d. <u>Instrumentation</u> includes changes to a test or data collection device (e.g., scoring rubric) or performance raters (e.g., a careless attitude) may produce changes in measurements of the dependent variable (Campbell & Stanley, 1963, p. 5). This is particularly possible in designs which rely on multiple observations (e.g., test administrations).
   (1) Two Examples

(a) For example, suppose it is discovered, after the administration of a pretest, that one item is a cue to the correct responses on three other items. Correcting the flawed item may produce "deflated" posttest scores as compared to pretest scores. Thus, the effect of the independent variable on the dependent variable (what the posttest is intended to measure) may be understated.

(b) Or, in an evaluation study, where the effect of a specific type of exercise on weight loss is being assessed, we find that the weight scale correctly weighed participants at program entry; but, understated weight by 6 pounds at program exit. Thus, the effects of the exercise program on weight are overstated.

(2) An instrument or data collection tool can be poorly calibrated (e.g., a scale) or constructed (e.g., a poorly written test); this is also called instrument bias.

(3) A data collection device can be so vaguely worded that its items fail to be sensitive to what should be measured; this is insensitive instrumentation. If more the items were more precisely worded, the instrument would be more sensitive.

(4) Some instruments are long, detailed, and/or tedious; this can give rise to attention (i.e., loss of interest or focus) bias and can lead to non-response bias. Instruments that require the respondent to recall events are subject to memory decay or recall bias; as the length of time between the event to be recalled and data collection increases, so does the chance of memory decay or recall bias.

(5) Evaluation research designers must take great care to ensure that all instrumentation is free of any defects and if mechanical, properly calibrated.

e. <u>Statistical Regression</u> occurs when a group of subjects has been assembled based on extreme scores or other extreme measurements, these outliers are likely to "move" or regress toward the group's mean (i.e., grand mean, N) on subsequent testing or measurement, using the same or similar instruments (Campbell & Stanley, 1963, p. 5). If extreme case or maximum variation sampling is used, statistical regression will more likely occur as "extremes" tend to move to the center; see the sampling section below for definitions of these non-probability sampling strategies.

(1) We don't know why outlier scores (i.e. scores which are significantly higher or lower than the general cluster of scores or measurements) move more to the center of the score distribution.

(2) Thus, statistical regression may effect measurement of the dependent variable in these purposeful sampling strategies: maximum variation sampling and extreme case sampling.

(3) In samples drawn, using probability sampling strategies (see the sampling discussion below), the random selection and assignment of subjects to either an experimental or control group is likely to control

(i.e., equalize, minimize, or eliminate) any impact this threat may have on either the independent or dependent variables.

f. Subject selection biases (e.g., selection criteria that lead to the experimental and control groups differing on critical variables or inconsistently applied assignment procedures to either the experimental or control groups, etc.) may be more responsible for changes in the dependent variable rather than the independent variable (Campbell & Stanley, 1963, p. 5). When this occurs, the internal validity threat, "selection" is said to operate.
   (1) In samples drawn, using probability sampling strategies (see sampling discussion below), the random selection and assignment of subjects to either an experimental or control group is likely to control (i.e., equalize, minimize, or eliminate) any impact this threat may have on either the independent or dependent variables.
   (2) In samples, based on purposeful sampling strategies, selection bias does likely operate because the treatment or comparison groups (e.g., Figure 8.4) are not randomly selected or assigned. This is only a problem if the two groups differ on important characteristics (considering the study purpose and research question or hypothesis), such as grade level, age, academic ability, etc. The researcher should thoroughly describe the treatment and comparison groups to show that their composition is quite similar.

g. Mortality or Attrition is the differential loss of study subjects from research groups is called mortality or attrition (Campbell & Stanley, 1963, p. 5).
   (1) The longer the time duration of a study, the greater is the likelihood that mortality, also called attrition or withdraw bias, will operate as subjects may be lost due to loss of interest, moving away, illness, death, etc.
   (2) The risk that mortality poses is that the experimental or control groups will become so different on important characteristics that these differences may be what causes changes in the dependent variable, rather than the independent variable.
   (3) Probability sampling strategies are likely to control any effect in studies that use one. For studies using a purposeful sampling strategy, the researcher should report in detail any changes in either the treatment group or comparison group.

h. Resentful Demoralization of the Control or Comparison Group ("Contamination Bias."). There are instances where a control or comparison group does not perform to its full potential (e.g., doesn't complete data collection activities honestly and accurately or doesn't put forth what should be its expected effort to lean content or skill) because they are not receiving the treatment (i.e., the independent variable) and they are angry about it. Thus, differences between the treatment and control or comparison group are likely inflated compared to what should have been observed.

(1) Such atypical or unusual performance is likely to either deflate or inflate differences between the experimental and control or the treatment and comparison groups (Martella, Nelson, & Marchand-Martella, 1999, p. 39).

(2) Purposeful sampling should control (i.e., equalize, minimize, or eliminate) the effects of this threat, where used. In studies using purposeful sampling, the researcher should report in detail any changes in either the treatment group or comparison group.

i. Diffusion of Treatment occurs in studies with either a control or comparison groups where differences between groups on measurements of the dependent variable may be due to unintended independent variable exposure (Martella, Nelson, & Marchand-Martella, 1999, p. 39).

(1) Two Examples

(a) Suppose in an AIDS drug effectiveness treatment study, several members of the control group were treated with the experimental drug and did not become as ill as often as other members of the control group. Such an experience would likely understate the effectiveness of the experimental drug.

(b) Suppose, there was an evaluation of a 6th grade math education program taught during the school day to students in School A (treatment) and not to 6th graders in School B (comparison). School A 6th graders were bused to School B for an after-school enrichment program. It is very probable that these students discussed their math lessons resulting in School B students learning and using some elements of the new math curriculum which was being evaluated. This could change the comparison group's math achievement scores.

(2) This is a difficult internal validity threat to control, regardless of sampling strategy. In any case, the researcher should thoroughly document and report instances where the independent variable or treatment was diffused to either the control or comparison group. Effective study project management might prevent such occurrences as cited in the examples provided.

C. Threats to External Validity (i.e., Generalizability)

1. External Design Validity

a. Study results and/or conclusions which can be validly generalized from a sample (e.g., a small group of Florida voters' opinion on a proposed constitutional amendment) back to its parent population (e.g., all registered Florida voters) are said to possess external validity. A probability sampling strategy is required; see the discussion on sampling below. These threats potentially apply primarily to studies using a probability sampling strategy.

b. Bracht and Glass (1968) have explained that external validity is a combination of "population" and "ecological" validity.

(1) Population validity is defined as the extent to which findings and/or conclusions can be generalized from a sample back to its parent population. It would be unwise to generalize the nutritional benefits of a low-fat diet drawn from a sample of vegetarians back to the general population; the general population consumes a great deal of red meat and vegetarians don't. Population validity also concerns subjects' living conditions, demographic, and/or socioeconomic status.

(2) Ecological validity concerns results generalization to other (similar or dissimilar) environmental conditions. If a study was conducted during the winter in upstate New York, where it is quite cold, then those findings might be validly generalized back to similar cold weather climates; generalizing those findings to South Beach in Miami during the winter will be inaccurate as the environmental conditions are dissimilar. Ecological validity could also include the political context within which the study took place, prevailing economic conditions, local culture, etc. Ecological validity centers primarily on the context within which the study was conducted.

(3) The central caution concerning population and ecological validity is that result findings can only be validly generalized to similar subjects who experienced conditions similar to those under which the study was conducted.

c. A study high in external design validity typically generates a more accurate generalization, i.e., the generalizing of a study's findings from a sample back to its parent population.

d. Generally, evaluation research studies conducted at operational levels (e.g., a worksite, classroom, or local social service beneficiaries) are not particularly concerned about generalizability, as the evaluation research question concerns itself only or primarily with those employees, students, or clients. It is when the evaluation results are part of a multisite study or are intended to be generalized from a sample to a population, that generalizability and thus external validity are important.

e. Specific threats to external design validity include the reactive or interactive effect of testing, reactive effects of experimental arrangements, multiple treatment interference, or experimenter effects.

2. Specific Threats to External Design Validity

a. According to Campbell and Stanley (1963, p. 5-6) the Reactive or Interactive Effect of Testing occurs when a dependent variable may be effected by subjects' prior testing (pretest) experience Posttest scores may be either inflated or deflated due to how subjects react to the pretest experience. This threat has also been called pretest or posttest sensitization (Martella, Nelson, & Marchand-Martella, 1999, p. 48). Study results generalization may be valid only to situations where a pretest and posttest was administrated. This is a real problem for time series designs (see Figure 8.5).

b. Interaction Effects of Selection, Instrumentation, History, or Maturation biases and the Treatment (or Independent) variable are the combined effects of specific internal validity threats. This effect is associated with how subjects were selected, how subjects interacted with study instrumentation, the historical experiences of subjects in the study, and/or subject maturation. Combined, these can affect both the independent and dependent variables so that accurate measurement is not possible (Campbell & Stanley, 1963, p. 5-6; Martella, Nelson, & Marchand-Martella, 1999, p. 40). This external validity threat is quite difficult to detect.

c. Reactive effects of an experimental arrangement may occur when an experiment or study is conducted using unnatural conditions (such as in a science lab) which are rarely duplicated "outside the lab." These contrived conditions may limit the generalizability of results (Campbell & Stanley, 1963, p. 5-6). Martella, Nelson, and Marchand-Martella (1999, pp. 48 & 51) refer to this as "verification of the independent variable" and include novelty and disruption effects as well as the Hawthorne effect. In other words, to accurately generalize results, we have to be sure that the conditions study subjects experienced are the same as the population to which we are trying to generalize. This speaks to ecological validity.

d. Multiple Treatment Interference occurs when subjects are exposed to multiple treatments (i.e., different independent variables) or repeated exposures to the same treatment, it is virtually impossible to separate the effects of the various treatments or determine the cumulative effect of repeated exposures to the independent variable or treatment (Campbell & Stanley, 1963, p. 5-6). Martella, Nelson, & Marchand-Martella (1999, p.48) offer a similar definition. This is a real problem for time series (see Figure 8.5) and factorial designs (see chapter 12).

e. Experimenter Effects may influence results in research designs where the researcher plays a significant role (e.g., interviews), he or she may exert an effect on the dependent variable in addition to or in place of the independent variable. This, of course, will limit generalizability (Martella, Nelson, & Marchand-Martella, 1999, p. 48). When the researcher is also the interviewer, the interviewee may provide "socially desirable" responses so as not to offend or may just want to please the researcher.

3. Probability sampling strategies (see the sampling discussion below) are likely to control (i.e., equalize, minimize, or eliminate) any effect of these threats and ensure external validity.
    a. However, probability sampling is not a 100% guarantee; the evaluation study designer should be familiar with these threats and be an expert on the empirical and professional literature relevant to the study in order to be able to identify and report whether or not any of these threats are present. If so, then the results should be interpreted cautiously and perhaps not generalized.
    b. Studies using a purposeful sampling strategy should make no attempt to generalize their results beyond the study subjects or sampling units.

D. Statistical and Social Validity

1. With respect to statistical validity, the researcher is most concerned with the "fit" between the statistical procedure applied to data given the type of data, design of the study, and research question or hypothesis. It is essential to ensure that the "correct" statistical procedure has been applied. Another consideration in ascertaining statistical validity is effect size. The larger the effect size indices, the potentially more significant are the results (effect sizes are discussed in subsequent chapters.)
    a. An evaluation designer can consider statistical validity as a potential separate and distinct study internal validity design threat.
    b. The controls for this design threat are to ensure that the study is well designed (internal validity design threats are adequately controlled) and that the statistical procedure(s) "fits" the data.

2. Social validity is more subjective than statistical validity (Wolf, 1978). It speaks directly to the social relevance of the research.
    a. Wolf outlined a three pronged test for assessing a study's social validity
        (1) Are the goals of the study socially relevant?
        (2) Were the study's procedures worth its findings, i.e., does the end justify the means?
        (3) Are the study's effects worth the cost of the study and noteworthy for society?
    b. Social validity has taken on an increasingly significant role in society as competition for resources becomes more intense and accountability demands grow. It might be said that the greater the degree of social validity possessed by a study, the more political and/or economic importance it may have.

III. Sampling in Evaluation Research Design

A. Key Sampling Terms and Concepts

1. A population is a defined as an aggregation or grouping of scores, people, cases, etc. to be studied. A sample is a portion of a population which is representative of that population. There are two primary sampling strategies: probability or purposeful.

a. Probability or Random Sampling
   (1) Random samples are drawn, using specific strategies, so that they are representative of the population from which they are drawn.
   (2) Representative samples are suitable for generalizing from a sample back to its parent population.
   (3) So, we must take care and select an evaluation design where threats to external validity or generalizability are controlled.
   (4) Evaluation research designs which employ probability sampling designs are referred to as true experimental designs or randomized controlled trials. Two examples are found in Chapter 8: Pretest-Posttest Control Group Design and The Pretest Only Control Group Design. Each of these designs has an experimental group (which receives the treatment or independent variable) and a control group which does not.

b. Purposeful Sampling
   (1) Purposeful samples are drawn with a non-probability sampling model or strategy.
   (2) The sample is structured to be information rich. Selection of sample units or subjects is based on prior identified criteria and the evaluation researcher must be knowledgeable about subjects' characteristics, the variability between and among subjects, etc.
   (3) We would use purposeful sampling when measuring the effectiveness of an educational intervention with an intact group such as a classroom or to assess the effectiveness of an HIV transmission prevention program which uses needle exchanges among IV drug users.
   (4) The critical point is that study subjects are grouped for a specific purpose; there is little concern with generalizing the results beyond those study subjects.

c. The survey, which is a research design, may use either a probability or a purposeful (also called a non-probability) sampling strategy, depending on the purpose of the survey and its evaluation research question or questions.

2. Considerations in Sampling Strategy Selection
   a. The sampling model (i.e., probability or purposeful) selected by the researcher must be appropriate given the research question(s) or hypotheses.
   b. Every sample has some degree of bias; the researcher's intent is to keep bias, regardless of source, as low as possible.
   c. The larger the sample size, usually the lower the risk of bias, provided the sample is representative of its corresponding population when a probability sampling strategy is used.
   d. The two most common errors in generalizing from a sample to a population are over-generalization and generalizing from a "sample" to a similar population when using a purposeful sampling strategy.

B. Probability Sampling Strategies

1. Probability or random samples are drawn in such a manner that every member of the population has an equal or identical chance of being included in the sample. The processes used are random selection and assignment; together these strategies ensure that the sample is representative of the population and enables sample results to be generalized back to the population. Criteria for an adequate probability sample are:
   a. A sample must be representative of its population, i.e., free of systematic bias. Systematic bias is consistent across the sample and will distort findings, either by inflating or depressing them.
   b. Precision reflects the influence of chance in drawing a sample and is measured by the standard error of the estimate. The smaller the standard error of the estimate, the more precise is the estimate.

2. Strategies to Ensure Accuracy and Precision
   a. Random selection and assignment of subjects to the experimental and control groups are the primary strategies for controlling threats to external validity and ensuring generalizability.
   b. Random selection of subjects ensures that the sample is representative of the population so that results from the sample can be generalized back to the parent population.
   c. Random assignment of subjects occurs when control groups are employed. This contributes to ensuring that the experimental and control groups are either equivalent or nearly so. This is critical to ensure that the sample is representative of the population.

3. Presented in Figure 7.1 is a demonstration of random selection. Presented in Figure 7.2 is a depiction of random assignment.
   a. As presented in Figure 7.1, the population is selected (The evaluation research designer must be very careful to properly describe the population to be included in the evaluation study.). Next, the sample is selected using a probability sampling strategy. Third, data are collected, analyzed, and reported. Fourth, results are generalized from the sample back to its parent population.

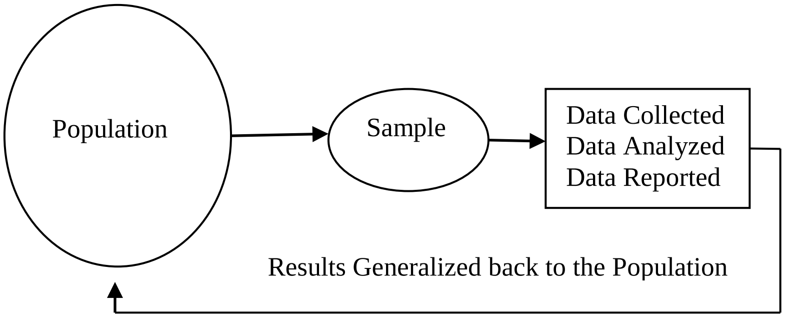

Figure 7.1 Random Selection

b. As presented in Figure 7.2, the population is selected. Next the sample was drawn using a probability sampling strategy. Third, subjects are randomly assigned to either the E-group (experimental group) or the C-group (control group). Fourth, data are collected, analyzed, and reported. Fifth, results are generalized back to the population from its sample.

3. **Probability (Random) Sampling Models**

a. In <u>simple random sampling</u> each sampling unit is assigned a number or other identifier; all numbers are pooled; then the specified number of units is selected. For example, from a population of 1000 high school seniors, we need to select 250 to be included in a survey sample. We could write each senior's name on a slip of paper of uniform size and then fold each slip of paper in the same way. Next, we could put all 1000 slips of paper into a big box and shake it vigorously. Finally, we would withdraw the first slip of paper and then shake the box vigorously again and pull out the second slip of paper. We would repeat this process until we had 250 names randomly selected from the population of 1000 high school seniors.

b. In <u>cluster sampling</u> the target area (usually geographical) is divided into sections, based on specified criteria related to the research question; then all sampling units within the selected clusters are studied. Clusters are usually randomly selected. For example, our evaluation consulting company, located in a large metropolitan area with 72 separate zip or postal code zones, has been asked to prepare a proposal to identify pressing needs for city services. The city wants a representative sample of all Metro residents. Accordingly, we have identified the most important social and demographic characteristics of the city; next, we have identified seven zip or postal code zones that closely match these social and demographic characteristics. Third, we randomly select three of the seven clusters to ensure that we have a representative sample. Fourth, we identify all households within each of the three selected zones. Fifth, our study teams will make attempts to interview each head of household in each zone.

c. In systematic sampling the first sampling unit or subject is randomly selected and then all other sampling units are selected from a list or its equivalent, every *k'th* point on until the desired sample size is reached. The researcher needs a list of all population members and needs to know how the list was constructed and to verify that no bias was built into the list. Suppose we wanted to survey every professional historian in the United States. Now, this would be very difficult due to cost and not knowing how to reach each historian; so, we elected to draw a random sample of historians who were members of state or national professional history associations. Accordingly, we contacted every national and state professional association to which historians belong and purchased membership lists. These lists were manually entered into a computer in order to eliminate duplicate names and then we randomly ordered all the names into one comprehensive list. Next, we randomly picked a number from 1 to 10, which was "8." Finally, starting with the eighth name on the list we picked every eighth name (8, 16, 24, 32, 38, 46, etc.) on the list until, we had the desired sample size.

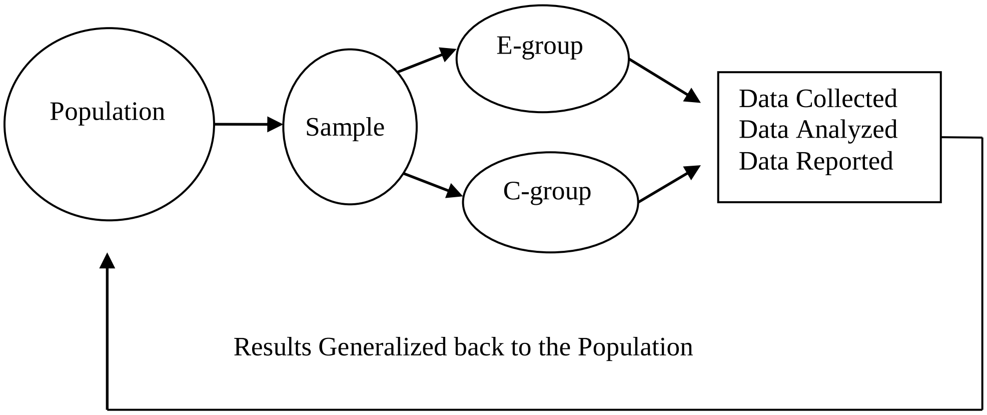

Figure 7.2 Random Assignment

d. Stratified Random Sampling

(1) In stratified random sampling, we divide the population into subgroups called strata. Our purpose is to ensure that each stratum is proportionately represented in the sample in the same manner as in the population. This strategy helps to ensure that the sample is representative of the population.

(2) The methodology used in stratified random sampling is profiled below:

(a) First, the sampling fraction is determined, by n/N, where: n = sample size and N = population size. The sampling fraction is the ratio of the sample size to the population size.

(b) Second, the sample size is determined, say 5% of the population.

(c) Applying the concept of proportional allocation, the sampling fraction is the same for all strata, which mirrors its population proportion.

(3) Suppose, we worked in the institutional research office of a university which had seven colleges, enrolling 15,823 students. We were tasked with conducting a survey of student satisfaction with university services.

(a) In order to ensure that the sample was representative of each college's enrollment within the University, we decided to draw the sample using stratified random sampling. We wanted a 5% sample size. So to determine the sample size for each stratum, we would take the stratum's enrollment and multiply it by 0.05. This way we can ensure that each college (stratum) is proportionally represented in the sample at the same level as the population (University students). Using this methodology, we computed the desired sample size to be 792 students.

(b) Next, using simple random sampling, we randomly selected 273 (5,461 x 0.05) students from the College of Arts and Sciences to be included in the sampling frame. We repeated this process until we achieved the desired sample size and proportional representation in the sample. See Table 7.2.

Table 7.2

Stratified Random Sampling Example

Strata (University Colleges)	**Strata Size**	**Strata Sample Size**
Arts and Sciences	5,461	273
Business Administration	1,850	93
Community Services	2,092	105
Education	3,508	175
Engineering	2, 112	106
Law	318	16
Pharmacy	482	24
Total	15,823 (N)	792 (n)

4. **Sample Size**

a. The question of what constitutes an adequate sample size is actually a bit complicated. A general rule of thumb is to have the largest possible sample that resources permit. Generally, the larger the sample, the more representative it is of the population.

b. Specific factors which determine sample size include desired statistical power, probability of a Type I statistical error, expected effect size, statistical procedure, and whether a one-tail or two-tail statistical test is to be applied to the data. A useful work (but a bit complicated) is Cohen (1999). Cohen presents sample size tables sorted by statistical procedure which combine the above factors to state specific sample sizes.

c. Sample Size Guidance
   (1) A sampling expert or someone who is experienced with Cohen's (1999) work should be consulted to determine needed sample size, in those instances where the sample is purposeful (See below).
   (2) If you are using a probability sampling strategy in order to generalize findings from a sample back to a population and you are unable to consult an expert, review the relevant professional literature and employ a sample size which was used in studies similar to the one you plan to conduct. Dattalo (2008) may be helpful.
   (3) Some statistical textbooks provide formulas for computing a sample size for specific statistical procedures; these formulas may prove useful.
   (4) As large a sample size as can be afforded is usually better as there is a lower risk of bias, provided the sample is representative of its parent population.

C. Purposeful Sampling

1. Purposeful sampling is used mostly in action and/or qualitative research, where our primary focus is only on those subjects or sampling units being studied. Recall, generalization from a sample studied back to a parent population is usually not possible or even desirable; so, we are not really worried about external validity threats when purposeful sampling is used. So, the processes outlined in Figures 7.1 and 7.2 don't apply.
   a. Commonly used action research designs are referred to as either pre-experimental or quasi-experimental.
      (1) Three pre-experimental designs are presented in Chapter 8: One-Shot Case Study, One Group Pretest Posttest Design, And the Static Group Comparison.
      (2) Two quasi-experimental designs are presented in Chapter 8: The Non-Equivalent Control Group Design and The Time Series Design.
      (3) Pre-experimental and quasi-experimental designs may use a comparison group, not a control group. It is important to be clear on which group label is correct; otherwise, a reader may be confused.
   b. Purposeful sampling is a good strategy, provided it logically fits the researcher's purpose. Remember, with purposeful sampling, we lose our ability to generalize our results back any population; in effect, our sample is our population. This is okay, because we are really only interested in evaluating the program or other effects on this particular sample or group of subjects.

2. Purposeful Sampling Strategies
   a. Comprehensive sampling includes all study units or subjects with specified characteristics as in the sample. In other words, our purpose is to assess program effects on a particular target group that meets the purpose of the study. Suppose, we were evaluating the effects of an educational program to teach learning-disabled students to maximize their reading achievement; we would want to include as many of these students in our sample as possible.
   b. Maximum variation sampling is a selection process that includes sampling units or subjects so that differences on specified characteristics are maximized. For example, suppose we study differences among three high schools (one low performing, one "average" performing and one high-performing) in a school district to try to explain causes for the different levels of performance. Or, we could assess differences among local leadership characteristics at five manufacturing plants, owned by the same company, to determine why one plant is more efficient than the others.
   c. Extreme case sampling includes sampling units or subjects with extreme or unusual characteristics. For example, we could select an extremely poor performing or exemplary performing school or company with the intent to learn why it is performance is poor or exemplary.
   d. Typical case sampling takes the middle road, selecting sampling units or subjects that are considered "typical", "normal", or "usual" of the phenomenon under study. Suppose, we were commissioned to study the "average" Americans' perception of brand quality or the "typical" American moms' opinion of political candidates' family-friendly policy proposals.
   e. Homogeneous sampling is used when the purpose of the study is to focus on a particular subgroup within a larger community or group. For example, we could study beginning teachers' teaching practices (as opposed to midcareer or late career teachers) or new customer service representatives' customer care skills (as opposed to senior customer service representatives).
   f. Snowball or chain sampling is employed when it is very difficult to identify subjects to study. Studies which seek to measure HIV knowledge among illegal drug users would use snowball or chain sampling. Once an illegal drug user is identified, he or she is asked to provide the name and location of other illegal drug users. Once these illegal drug users have been located, they are asked to provide additional leads and so on.

Review Questions

Directions. Read each item carefully; either fill-in-the-blank or circle letter associated with the term that best answers the item.

1. The internal validity threat where cases (i.e., subjects) respond differently due to repeated measurements is most likely:
   a. History
   b. Mortality
   c. Testing
   d. Instrumentation

2. The internal validity threat where there are changes in cases between measurements is:
   a. History
   b. Maturation
   c. Mortality
   d. Regression

3. The internal validity threat where the measure changes over time is:
   a. History
   b. Testing
   c. Instrumentation
   d. Mortality

4. Which of the following are key factors in selecting a sampling strategy?
   a. Response rate
   b. Cost and time
   c. Experimental group attributes
   d. “a” and “b”

5. Within a _________ study, measurements of the dependent variable are conducted once.
   a. Longitudinal
   b. Ex post facto
   c. Cross-sectional
   d. Panel

6. Criteria for determining an adequate sample include:
   a. Representativeness
   b. Randomness
   c. Precision
   d. “a” and “c”

7. The internal validity threat where cases (i.e., subjects) respond differently due to repeated measurements is most likely:
   a. History
   b. Mortality
   c. Testing
   d. Instrumentation

8. The internal validity threat where there are changes in cases between measurements is:
   a. History
   b. Maturation
   c. Mortality
   d. Regression

9. The internal validity threat where the measure changes over time is:
   a. History
   b. Testing
   c. Instrumentation
   d. Mortality

10. Which of the following are key factors in selecting a sampling strategy?
    a. Response rate
    b. Cost and time
    c. Experimental group attributes
    d. “a” and “b”

Answers: 1. c, 2. b, 3. c, 4. d, 5. c, 6. d, 7. c. 8. b, 9. c, 10. d.

References

Bracht, G. W. & Glass, G. V. (1968). The external validity of experiments. *American Educational Journal, 5,* 437-474.

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

Cohen, J. (1999). *Statistical power analysis for the behavioral sciences* (3rd. ed.).Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Dattalo, P. (2008). *Determining sample size: Balancing power, precision, and practicality*. New York, NY: Oxford University Press.

Martella, R. C., Nelson, R. & Marchand-Martella, N. E., (1999). *Research methods: Learning to become a critical research consumer*. Boston, MA: Allyn & Bacon.

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis measurement is finding its heart. *Journal of Applied Behavior Analysis, 11,* 203-214.