Child Care & Early Education

RESEARCH CONNECTIONS

http://www.researchconnections.org

Research Connections Research Methods Pages

NCCP ICPSR OPRE

SECTION 1. INTRODUCTION

This document provides basic information about the study designs, data collection and analytic methods commonly used in early child care policy and other social science research. It is intended to give users a general orientation to research designs and methodologies to aid in understanding research in this field. Below are links to descriptions and discussions of different research designs, data collection approaches, and analytic methods and strategies. In addition to the information and discussion contained in each of these sections, links are provided to resources where additional information on a topic can be found. Terms related to each of these areas are defined in the Research Glossary.

Section 2. Study Design and Analysis

The Administration for Children & Families (ACF) Common Framework for Research and Evaluation outlines different types of studies that generate information and answer empirical questions related to the human services provided by ACF and the recipients of those services. These types of studies fall into two overarching research categories: Descriptive, and Impact or Causal. While descriptive studies provide descriptions of programs, individuals and families, and identify how different characteristics of programs and program recipients are related to one another, impact or causal studies estimate the causal effect of an intervention.

There are two basic types of causal study design: randomized experiments and quasi-experiments. According to the ACF guidelines, a causal study design should include randomly assigned treatment and comparison/control groups when feasible. Quasi-experimental designs should only be used when random assignment is not possible.

Part I. Descriptive Research Studies
Discussion of the purposes of descriptive research, the methods used and the strengths and limitations of this type of research.

Part II. Causal Studies
Discussion of different randomized and quasi-experimental designs, the validity of their results, and advantages and disadvantages of the different designs.

Part III. Data Analysis
Different statistics and the methods and strategies used to describe the characteristics of the members of a sample or population, explore the relationships between variables and test research hypotheses are described. Approaches used to visually represent data are described.

Section 3. Data Collection

Part I. Survey Research and Questionnaires
Descriptions of key issues in survey research and questionnaire design. Descriptions of different types of sampling designs and data collection approaches with sources of error identified and described.

Part II. Administrative Data
Descriptions of key issues in research using administrative data records and the advantages and disadvantages of using administrative data.

Part III. Field Research
Descriptions of methods of field research and the advantages and disadvantages of these.

Section 4. Assessing Research Quality

Guidance on how to assess the quality of quantitative and qualitative research and tools to assist in evaluating research studies is provided. Principles of ethical research and how these are incorporated into research with human subjects are also discussed.

See Also

ACF Common Framework for Research and Evaluation
This document outlines the roles of various types of research and evaluation in generating information and answering empirical questions related to the human services provided by the Administration for Children & Families (ACF).

SECTION 2. STUDY DESIGN AND ANALYSIS

Part I. Descriptive Research Studies
Part II. Causal Studies
Part III. Data Analysis

Part I. Descriptive Research Studies

Descriptive research is a type of research that is used to describe the characteristics of a population. It collects data that are used to answer a wide range of what, when, and how questions pertaining to a particular population or group. For example, descriptive studies might be used to answer questions such as: What percentage of Head Start teachers have a bachelor's degree or higher? What is the average reading ability of 5-year-olds when they first enter kindergarten? What kinds of math activities are used in early childhood programs? When do children first receive regular child care from someone other than their parents? When are children with developmental disabilities first diagnosed and when do they first receive services? What factors do programs consider when making decisions about the type of assessments that will be used to assess the skills of the children in their programs? How do the types of services children receive from their early childhood program change as children age?
Descriptive research does not answer questions about why a certain phenomenon occurs or what the causes are**.** Answers to such questions are best obtained from randomized and quasi-experimental studies. However, data from descriptive studies can be used to examine the relationships (correlations) among variables. While the findings from correlational analyses are not evidence of causality**,** they can help to distinguish variables that may be important in explaining a phenomenon from those that are not. Thus, descriptive research is often used to generate hypotheses that should be tested using more rigorous designs.

A variety of data collection methods may be used alone or in combination to answer the types of questions guiding descriptive research. Some of the more common methods include surveys, interviews, observations, case studies, and portfolios. The data collected through these methods can be either quantitative or qualitative. Quantitative data are typically analyzed and presenting using descriptive statistics. Using quantitative data, researchers may describe the characteristics of a sample or population in terms of percentages (e.g., percentage of population that belong to different racial/ethnic groups, percentage of low-income families that receive different government services) or averages (e.g., average household income, average scores of reading, mathematics and language assessments). Quantitative data, such as narrative data collected as part of a case study, may be used to organize, classify, and used to identify patterns of behaviors, attitudes, and other characteristics of groups.

Descriptive studies have an important role in early care and education research. Studies such as the National Survey of Early Care and Education and the National Household Education Surveys Program have greatly increased our knowledge of the supply of and demand for child care in the U.S. The Head Start Family and Child Experiences Survey and the Early Childhood Longitudinal Study Program have provided researchers, policy makers and practitioners with rich information about school readiness skills of children in the U.S.

Each of the methods used to collect descriptive data have their own strengths and limitations. The following are some of the strengths and limitations of descriptive research studies in general.

Strengths:

- Study participants are questioned or observed in a natural setting (e.g., their homes, child care or educational settings).
- Study data can be used to identify the prevalence of particular problems and the need for new or additional services to address these problems.
- Descriptive research may identify areas in need of additional research and relationships between variables that require future study. Descriptive research is often referred to as "hypothesis generating research."
- Depending on the data collection method used, descriptive studies can generate rich datasets on large and diverse samples.

Limitations:

- Descriptive studies cannot be used to establish cause and effect relationships.
- Respondents may not be truthful when answering survey questions or may give socially desirable responses.
- The choice and wording of questions on a questionnaire may influence the descriptive findings.
- Depending on the type and size of sample, the findings may not be generalizable or produce an accurate description of the population of interest.

Part II. Causal Studies

Researchers conduct experiments to study cause and effect relationships and to estimate the impact of child care and early childhood programs on children and their families. There are two basic types of experiments:

A. Randomized experiments
B. Quasi-experiments

An experiment is a study in which the researcher manipulates the treatment, or intervention, and then measures the outcome. It addresses the question “if we change X (the treatment or intervention), what happens to Y (the outcome)?” Conducted both in the laboratory and in real life situations, experiments are powerful techniques for evaluating cause-and-effect relationships. The researcher may manipulate whether research subjects receive a treatment (e.g., attendance in a Head Start program: yes or no) or the level of treatment (e.g., hours per day in the program).

Suppose, for example, a group of researchers was interested in the effect of government-funded child care subsidies on maternal employment. They might hypothesize that the provision of government-subsidized child care would promote such employment. They could then design an experiment in which some mothers would be provided the option of government-funded child care subsidies and others would not. The researchers might also manipulate the value of the child care subsidies in order to determine if higher subsidy values might result in different levels of maternal employment.

The group of participants that receives the intervention or treatment is known as the "treatment group," and the group that does not is known as the "control group" in randomized experiments and "comparison group" in quasi-experiments.

The key distinction between randomized experiments and quasi-experiments lies in the fact that in a randomized experiment participants are randomly assigned to either the treatment or the control group whereas participants are not in a quasi-experiment.

A. Randomized Experiments

Random assignment ensures that all participants have the same chance of being in a given experimental condition. Randomized experiments (also known as RCT or randomized control trials) are considered to be the most rigorous approach, or the "gold standard," to identifying causal effects because they theoretically eliminate all preexisting differences between the treatment and control groups. However, some differences might occur due to chance. In practice, therefore, researchers often control for observed characteristics that might differ between individuals in the treatment and control groups when estimating treatment effects. The use of control variables improves the precision of treatment effect estimates.

Cluster-randomized experiments

Despite being the "gold standard" in causal study design, randomized experiments are not common in social science research because it is often impossible or unethical to randomize individuals to experimental conditions. Cluster-randomized experiments, in which groups (e.g., schools or classes) instead of individuals are randomized, often encounter less objections out of ethical concerns and therefore are more feasible in real life. They also prevent treatment spill over to the control group. For example, if students in the same class are randomly assigned to either the treatment or control group with the treatment being a new curriculum, teachers may introduce features of the treatment (i.e., new curriculum) when working with students in the control group in ways that might affect the outcomes.

One drawback of cluster-randomized experiments is a reduction in statistical power. That is, the likelihood that a true effect is detected is reduced with this design.

B. Quasi-Experiments

Quasi-experiments are characterized by the lack of randomized assignment. They may or may not have comparison groups. When there are both comparison and treatment groups in a quasi-experiment, the groups differ not only in terms of the experimental treatment they receive, but also in other, often unknown or unknowable, ways. As a result, there may be several "rival hypotheses" competing with the experimental manipulation as explanations for observed results.

There are a variety of quasi-experiments. Below are some of the most common types in social and policy research, arranged in the order of weak to strong in terms of their capabilities of addressing threats to a statement that the relationship between the treatment and the outcome of interest is causal.

One group only:

- **One-group pretest-posttest**

A single group that receives the treatment is observed at two time points, one before the treatment and one after the treatment. Changes in the outcome of interest are presumed to be the effect of the treatment. For example, a new fourth grade math curriculum is introduced and students' math achievement is assessed in the fall and spring of the school year. Improved scores on the assessment are attributed to the curriculum. The biggest weakness of this design is that a number of events can happen around the time of the treatment and influence the outcome. There can be multiple plausible alternative explanations for the observed results.

- **Interrupted time series**

A single group that receives the treatment is observed at multiple time points both before and after the treatment. A change in the trend around the time of the treatment is presumed to be the treatment effect. For example, individuals participating in an exercise program might be weighed each week before and after a new exercise routine is introduced. A downward trend in their weight around the time the new routine was introduced would be seen as evidence of the effectiveness of the treatment. This design is stronger than one-group pretest-posttest because it shows the trend in the outcome variable both before and after the treatment instead of a simple two-point-in-time comparison. However, it still suffers the same weakness that other events can happen at the time of the treatment and be the alternative causes of the observed outcome.

Two groups:

- **Static-group comparison**

A group that has experienced some treatment is compared with one that has not. Observed differences between the two groups are assumed to be the result of the treatment. For example, fourth graders in some classrooms in a school district are introduced to a new math curriculum while fourth graders in other classrooms in the district are not. Differences in the math scores of the two groups assessed in the spring of the school year only are assumed to be the result of the new curriculum. The weakness of this design is that the treatment and comparison groups may not be truly comparable because participants are not randomly assigned to the groups and there may be important differences in the characteristics and experiences of the groups, only some of which may be known. If the two groups differ in ways that affect the outcome of interest, the causal claim cannot be presumed.

- **Difference-in-differences**

Both treatment and comparison groups are measured before and after the treatment. The difference between the two before-after differences is presumed to be the treatment effect. This design is an improvement of the static-group comparison because it compares outcomes that are measured both before and after the treatment is introduced instead of two post-treatment outcomes. For example, the fourth graders in the prior example are assessed in both the fall (pre-treatment) and spring (post-treatment). Differences in the fall-spring scores between the two fourth grade groups are seen as evidence of the effect of the curriculum. For this reason, the treatment and comparison groups in difference-in-differences do not have to be perfectly comparable. The biggest challenge for the researcher is to defend the parallel trend assumption, namely the *change* in the treatment group would be the same as the change in the comparison group in the absence of the treatment.

- **Regression discontinuity**

Participants are assigned to experimental conditions based on whether their scores are above or below a cut point for a quantitative variable. For example, students who score below 75 on a math test are assigned to the treatment group with the treatment being an intensive tutoring program. Those who score at or above 75 are assigned to the comparison group. The students who score just above or below the cut point are considered to be on average identical because their score differences are most likely due to chance. These students therefore act as if they were randomly assigned. The difference in the outcome of interest (e.g., math ability as measured by a different test after the treatment) between the students right around the cut point is presumed to be the treatment effect.
Regression discontinuity is an alternative to randomized experiments when the latter design is not possible. It is the only recognized quasi-experimental design that meets the Institute of Education Sciences standards for establishing causal effects. Although considered to be a strong quasi-experimental design, it needs to meet certain conditions.

Resources

See the following for additional information on randomized and quasi-experimental designs.
The Core Analytics of Randomized Experiments for Social Research
Experimental and Quasi-Experimental Designs for Research
Experimental and Quasi-Experimental Designs for Generalized Causal Inference

Instrumental Variables (IV) Approach

An instrumental variable is a variable that is correlated with the independent variable of interest and only affects the dependent variable through that independent variable. The IV approach can be used in both randomized experiments and quasi-experiments.

In randomized experiments, the IV approach is used to estimate the effect of treatment receipt, which is different from treatment offer. Many social programs can only offer participants the treatment, or intervention, but not mandate them to use it. For example, parents are randomly assigned by way of lottery to a school voucher program. Those in the treatment group are offered vouchers to help pay for private school, but ultimately it is up to the parents to decide whether or not they will use the vouchers. If the researcher is interested in estimating the impact of voucher usage, namely the effect of treatment receipt, the IV approach is one way to do so. In this case, the IV is the treatment assignment status (e.g., a dummy variable with 1 being in the treatment group and 0 being in the control group), which is used to predict the probability of a parent using the voucher, which is in turn used as the independent variable of interest to estimate the effect of voucher usage.

In quasi-experiments, the IV approach is used to address the issue of endogeneity, namely that the treatment status is determined by participants themselves (self selection) or by criteria established by the program designer (treatment selection). Endogeneity is an issue that plagues quasi-experiments and often a source of threats to the causal claim. The IV approach can be used to tease out the causal impact of an endogenous variable on the outcome. For example, researchers used cigarette taxes as an instrumental variable to estimate the effect of maternal smoking on birth outcomes (Evans and Ringel, 1999). Cigarette taxes affect how much pregnant mothers smoke but not birth outcomes. They therefore meet the condition of being an IV, which correlates with the independent variable/treatment (i.e., maternal smoking habit) and only affects the dependent variable (i.e., birth outcomes) through that independent variable. The estimated effect is, strictly speaking, a local average treatment effect, namely the effect of treatment (maternal smoking) among those mothers affected by the IV (cigarette taxes). It does not include mothers whose smoking habit is not affected by the price of cigarettes (e.g., chain smokers who may be addicted to nicotine).

An instrumental variable needs to meet certain conditions to provide a consistent estimate of a causal effect.

Resources

See the following for additional information on instrumental variables.

An introduction to instrumental variable assumptions, validation and estimation
An Introduction to Instrumental Variables

Validity of Results from Causal Designs

The two types of validity are internal and external. It is often difficult to achieve both in social science research experiments.

- **Internal Validity**
    - Internal validity refers to the strength of evidence of a causal relationship between the treatment (e.g., child care subsidies) and the outcome (e.g., maternal employment).
    - When subjects are randomly assigned to treatment or control groups, we can assume that the treatment caused the observed outcomes because the two groups should not have differed from one another at the start of the experiment.
    - For example, take the child care subsidy example above. Since research subjects were randomly assigned to the treatment (child care subsidies available) and control (no child care subsidies available) groups, the two groups should not have differed at the outset of the study. If, after the intervention, mothers in the treatment group were more likely to be working, we can assume that the availability of child care subsidies promoted maternal employment.

  One potential threat to internal validity in experiments occurs when participants either drop out of the study or refuse to participate in the study. If individuals with particular characteristics drop out or refuse to participate more often than individuals with other characteristics, this is called differential attrition. For example, suppose an experiment was conducted to assess the effects of a new reading curriculum on the reading achievement of 10th graders. Schools were randomly assigned to use the new curriculum in all classrooms (treatment schools) or to continue using their current curriculum (control schools). If many of the slowest readers in treatment schools left the study before it was completed (e.g., dropped out of school or transferred to a school in another state), schools with the new curriculum would experience an increase in the average reading scores. The reason they experienced an increase in reading scores, however, is because weaker readers left the school, not because the new curriculum improved students' reading skills. The effects of the curriculum on the achievement of 10th graders might be overestimated, if schools in the control schools did not experience the same type of attrition.

- **External Validity**

    - External validity, or generalizability, is also of particular concern in social science experiments.
    - It can be very difficult to generalize experimental results to groups that were not included in the study.
    - Studies that randomly select participants from the most diverse and representative populations are more likely to have external validity.

For example, a study shows that a new curriculum improved reading comprehension of third-grade children in Iowa. To assess the study's external validity, the researcher would consider whether this new curriculum would also be effective with third graders in New York or with children in other elementary grades.

Advantages and Disadvantages of Experimental and Quasi-Experimental Designs

- Randomized experiments
    - Advantages
        - Yield the most accurate assessment of cause and effect.
        - Typically have strong internal validity.
        - Ensure that the treatment and control groups are truly comparable and that treatment status is not determined by participant characteristics that might influence the outcome.
    - Disadvantages
        - In social policy research, it can be impractical or unethical to conduct randomized experiments.
        - They typically have limited external validity due to the fact that they often rely on volunteers and are implemented in a somewhat artificial experimental setting with a small number of participants.
        - Despite being the “gold standard” for identifying causal impacts, they can also be faced with threats to internal validity such as attrition, contamination, cross-overs, and Hawthorne effects.
- Quasi-experiments
    - Advantages
        - Often have stronger external validity than randomized experiments because they are typically implemented in real-world settings and on larger scale.
        - May be more feasible than randomized experiments because they have fewer time and logistical constraints often associated with randomized experiments.
        - Avoid the ethical concerns associated with random assignment.
        - Are often less expensive than randomized experiments.
    - Disadvantages
        - They often have weaker internal validity than randomized experiments.
        - The lack of randomized assignment means that the treatment and control groups may not be comparable and that treatment status may be driven by

participant characteristics or other experiences that might influence the outcome.
- Conclusions about causality are less definitive than randomized experiments due to the lack of randomization and reduced internal validity.
- Despite having weaker internal validity, they are often the best option available when it is impractical or unethical to conduct randomized experiments.

Part III. Data Analysis

Different statistics and methods used to describe the characteristics of the members of a sample or population, explore the relationships between variables, to test research hypotheses, and to visually represent data are described. Terms relating to the topics covered are defined in the Research Glossary.

A. Descriptive Statistics
B. Tests of Significance
C. Graphical/Pictorial Methods
D. Analytical Techniques

A. Descriptive Statistics

Descriptive statistics can be useful for two purposes:

1. To provide basic information about the characteristics of a sample or population. These characteristics are represented by variables in a research study dataset.
2. To highlight potential relationships between these characteristics, or the relationships among the variables in the dataset.

The four most common descriptive statistics are:

- Proportions, Percentages and Ratios
- Measures of Central Tendency
- Measures of Dispersion
- Measures of Association

Proportions, Percentages and Ratios

One of the most basic ways of describing the characteristics of a sample or population is to classify its individual members into mutually exclusive categories and counting the number of cases in each of the categories. In research, variables with discrete, qualitative categories are called nominal or categorical variables. The categories can be given numerical codes, but they cannot be ranked, added, or multiplied. Examples of nominal variables include gender (male,

female), preschool program attendance (yes, no), and race/ethnicity (White, African American, Hispanic, Asian, American Indian). Researchers calculate proportions, percentages and ratios in order to summarize the data from nominal or categorical variables and to allow for comparisons to be made between groups.

Proportion--The number of cases in a category divided by the total number of cases across all categories of a variable.

Percentage--The proportion multiplied by 100 (or the number of cases in a category divided by the total number of cases across all categories of a value times 100).

Ratio--The number of cases in one category to the number of cases in a second category.

Example:

A researcher selects a sample of 100 students from a Head Start program. The sample includes 20 White children, 30 African American children, 40 Hispanic children and 10 children of mixed-race/ethnicity.

Proportion of Hispanic children in the program = 40 / (20+30+40+10) = .40.

Percentage of Hispanic children in the program = .40 x 100 = 40%.

Ratio of Hispanic children to White children in the program = 40/20 = 2.0, or the ratio of Hispanic to White children enrolled in the Head Start program is 2 to 1.

Measures of Central Tendency

Proportions, percentages and ratios are used to summarize the characteristics of a sample or population that fall into discrete categories. Measures of central tendency are the most basic and, often, the most informative description of a population's characteristics, when those characteristics are measured using an interval scale. The values of an interval variable are ordered where the distance between any two adjacent values is the same but the zero point is arbitrary. Values on an interval scale can be added and subtracted. Examples of interval scales or interval variables include household income, years of schooling, hours a child spends in child care and the cost of child care.

Measures of central tendency describe the "average" member of the sample or population of interest. There are three measures of central tendency:

Mean -- The arithmetic average of the values of a variable. To calculate the mean, all the values of a variable are summed and divided by the total number of cases.

Median -- The value within a set of values that divides the values in half (i.e. 50% of the variable's values lie above the median, and 50% lie below the median).

Mode -- The value of a variable that occurs most often.

Example:

The annual incomes of five randomly selected people in the United States are $10,000, $10,000, $45,000, $60,000, and $1,000,000.

Mean Income = (10,000 + 10,000 + 45,000 + 60,000 + 1,000,000) / 5 = $225,000.

Median Income = $45,000.

Modal Income = $10,000.

The mean is the most commonly used measure of central tendency. Medians are generally used when a few values are extremely different from the rest of the values (this is called a skewed distribution). For example, the median income is often the best measure of the average income because, while most individuals earn between $0 and $200,000 annually, a handful of individuals earn millions.

Measures of Dispersion

Measures of dispersion provide information about the spread of a variable's values. There are three key measures of dispersion:

- Range
- Variance
- Standard Deviation

Range is simply the difference between the smallest and largest values in the data. Researchers often report simply the values of the range (e.g., 75 – 100).

Variance is a commonly used measure of dispersion, or how spread out a set of values are around the mean. It is calculated by taking the average of the squared differences between each value and the mean. The variance is the standard deviation squared.

Standard deviation, like variance, is a measure of the spread of a set of values around the mean of the values. The wider the spread, the greater the standard deviation and the greater the range of the values from their mean. A small standard deviation indicates that most of the values are close to the mean. A large standard deviation on the other hand indicates that the values are more spread out. The standard deviation is the square root of the variance.

Example:

Five randomly selected children were administered a standardized reading assessment. Their scores on the assessment were 50, 50, 60,75 and 90 with a mean score of 65.

Range = 90 - 50 = 40.

Variance = $[(50 - 65)^2 + (50 - 65)^2 + (60 - 65)^2 + (75 - 65)^2 + (90 - 65)^2] / 5 = 300$.

Standard Deviation = Square Root (150,540,000,000) = 17.32.

Skewness and Kurtosis

The range, variance and standard deviation are measures of dispersion and provide information about the spread of the values of a variable. Two additional measures provide information about the shape of the distribution of values.

Skew is a measure of whether some values of a variable are extremely different from the majority of the values. Skewness refers to the tendency of the values of a variable to depart from symmetry. A distribution is symmetric if one half of the distribution is exactly equal to the other half. For example, the distribution of annual income in the U.S. is skewed because most people make between $0 and $200,000 a year, but a handful of people earn millions. A variable is positively skewed (skewed to the right) if the extreme values are higher than the majority of values. A variable is negatively skewed (skewed to the left) if the extreme values are lower than the majority of values. In the example of students' standardized test scores, the distribution is slightly positively skewed.

Kurtosis measures how outlier-prone a distribution is. Outliers are values of a variable that are much smaller or larger than most of the values found in a dataset. The kurtosis of a normal distribution is 0. If the kurtosis is different from 0, then the distribution produces outliers that are either more extreme (positive kurtosis) or less extreme (negative kurtosis) than are produced by the normal distribution.

Measures of Association

Measures of association indicate whether two variables are related. Two measures are commonly used:

- Chi-square test of independence
- Correlation

Chi-Square test of independence[1] is used to evaluate whether there is an association between two variables.

- It is most often used with nominal data (i.e., data that are put into discrete categories: e.g., gender [male, female] and type of job [unskilled, semi-skilled, skilled]) to determine whether they are associated. However, it can also be used with ordinal data.
- Assumes that the samples being compared (e.g., males, females) are independent.
- Tests the null hypothesis of no difference between the two variables (i.e., type of job is not related to gender).

To test for associations, a chi-square is calculated in the following way: Suppose a researcher wants to know whether there is a relationship between gender and two types of jobs, construction worker and administrative assistant. To perform a chi-square test, the researcher counts the number of female administrative assistants, the number of female construction workers, the number of male administrative assistants, and the number of male construction workers in the data. These counts are compared with the number that would be expected in each category if there were no association between job type and gender (this expected count is based on statistical calculations). The association between the two variables is determined to be significant (the null hypothesis is rejected), if the value of the chi-square test is greater than or equal to the critical value for a given significance level (typically .05) and the degrees of freedom associated with the test found in a chi-square table. The degrees of freedom for the chi-square are calculated using the following formula: $df = (r-1)(c-1)$ where r is the number of rows and c is the number of columns in a contingency or cross-tabulation table. For example, the critical value for a 2 x 2 table with 1 degree of freedom ([2-1][2-1]=1) is 3.841.

Correlation coefficient is used to measure the strength and direction of the relationship between numeric variables (e.g., weight and height).

- The most common correlation coefficient is the Person's product-moment correlation coefficient (or simply **Pearson's r)**, which can range from -1 to +1.
- Values closer to 1 (either positive or negative) indicate that a stronger association exists between the two variables.
- A positive coefficient (values between 0 and 1) suggests that larger values of one of the variables are accompanied by larger values of the other variable. For example, height and weight are usually positively correlated because taller people tend to weigh more.
- A negative association (values between 0 and -1) suggests that larger values of one of the variables are accompanied by smaller values of the other variable. For example, age and hours slept per night are often negatively correlated because older people usually sleep fewer hours per night than younger people.

[1] The chi-square test can also be used as a measure of goodness of fit, to test if data from a sample come from a population with a specific distribution, as an alternative to Anderson-Darling and Kolmogorov-Smirnov goodness-of-fit tests.

B. Tests of Significance

The findings reported by researchers are typically based on data collected from a single sample that was drawn from the population of interest (e.g., a sample of children selected from the population of children enrolled in Head Start or Early Head Start). If additional random samples of the same size were drawn from this population, the estimated percentages and means calculated using the data from each of these other samples might differ by chance somewhat from the estimates produced from one sample. Researchers use one of several tests to evaluate whether their findings are statistically significant.

Statistical significance refers to the probability or likelihood that the difference between groups or the relationship between variables observed in statistical analyses is not due to random chance (e.g., that differences between the average scores on a measure of language development between 3- and 4-year-olds are likely to be "real" rather than just observed in this sample by chance). If there is a very small probability that an observed difference or relationship is due to chance, the results are said to reach statistical significance. This means that the researcher concludes that there is a real difference between two groups or a real relationship between the observed variables.

Significance tests and the associated *p*- value only tell us how likely it is that a statistical result (e.g., a difference between the means of two or more groups, or a correlation between two variables) is due to chance. The p-value is the probability that the results of a statistical test are due to chance. In the social and behavioral sciences, a p-value less than or equal to .05 is usually interpreted to mean that the results are statistically significant (that the statistical results would occur by chance 5 times or fewer out of 100), although sometimes researchers use a p-value of .10 to indicate whether a result is statistically significant. The lower the p-value, the less likely a statistical result is due to chance. Lower p-values are therefore a more rigorous criteria for concluding significance.

Researchers use a variety of approaches to test whether their findings are statistically significant or not. The choice depends on several factors, including the number of groups being compared, whether the groups are independent from one another, and the type of variables used in the analysis. Three widely used tests are the t-test, F-test, and Chi-square test.

Three of the more widely used tests of statistical significance are described briefly below.

- **Chi-Square test** is used when testing for associations between categorical variables (e.g., differences in whether a child has been diagnosed as having a cognitive disability by gender or race/ethnicity). It is also used as a goodness-of-fit test to determine whether data from a sample come from a population with a specific distribution.

- **t-test** is used to compare the means of two independent samples (independent t-test), the means of one sample at different times (paired sample t-test) or the mean of one

sample against a known mean (one sample t-test). For example, when comparing the mean assessment scores of boys and girls or the mean scores of 3- and 4-year-old children, an independent t-test would be used. When comparing the mean assessment scores of girls only at two time points (e.g., fall and spring of the program year) a paired t-test would be used. A one sample t-test would be used when comparing the mean scores of a sample of children to the mean score of a population of children. The t- test is appropriate for small sample sizes (less than 30) although it is often used when testing group differences for larger samples. It is also used to test whether correlation and regression coefficients are significantly different from zero.

- **F-test** is an extension of the t-test and is used to compare the means of three or more independent samples (groups). The F-test is used in Analysis of Variance (ANOVA) to examine the ratio of the between groups to within groups variance. It is also used to test the significance of the total variance explained by a regression model with multiple independent variables.

Significance tests alone do not tell us anything about the size of the difference between groups or the strength of the association between variables. Because significance test results are sensitive to sample size, studies with different sample sizes with the same means and standard deviations would have different t statistics and *p* values. It is therefore important that researchers provide additional information about the size of the difference between groups or the association and whether the difference/association is substantively meaningful.

Resources

See the following for additional information about descriptive statistics and tests of significance:

- Descriptive analysis in education: A guide for researchers
- Basic Statistics
- Descriptive Statistics
- Tests of Significance
- Effect Sizes and Statistical Significance
- Summarizing and Presenting Data
- Skewness

C. Graphical/Pictorial Methods

There are several graphical and pictorial methods that enhance understanding of individual variables and the relationships between variables. Graphical and pictorial methods provide a visual representation of the data. Some of these methods include:

- Bar charts
- Pie charts
- Line graphs
- Scatter plots
- Geographical Information Systems (GIS)
- Sociograms

Bar charts

- **Bar charts** visually represent the frequencies or percentages with which different categories of a variable occur.
- Bar charts are most often used when describing the percentages of different groups with a specific characteristic. For example, the percentages of boys and girls who participate in team sports. However, they may also be used when describing averages such as the average boys and girls spend per week participating in team sports.
- Each category of a variable (e.g., gender [boys and girls], children's age [3, 4, and 5]) is displayed along the bottom (or horizontal or X axis) of a bar chart.
- The vertical axis (or Y axis) includes the values of the statistic on that the groups are being compared (e.g., percentage participating in team sports).
- A bar is drawn for each of the categories along the horizontal axis and the height of the bar corresponds to the frequency or percentage with which that value occurs.

Pie charts

- A **pie chart** (or a **circle chart**) is one of the most commonly used methods for graphically presenting statistical data.
- As its name suggests, it is a circular graphic, which is divided into slices to illustrate the proportion or percentage of a sample or population that belong to each of the categories of a variable.
- The size of each slice represents the proportion or percentage of the total sample or population with a specific characteristic (found in a specific category). For example, the percentage of children enrolled in Early Head Start who are members of different racial/ethnic groups would be represented by different slices with the size of each slice proportionate to the group's representation in the total population of children enrolled in the Early Head Start program.

Line graphs

- A **line graph** is a type of chart which displays information as a series of data points connected by a straight line.
- Line graphs are often used to show changes in a characteristic over time.
- It has an X-axis (horizontal axis) and a Y axis (vertical axis). The time segments of interest are displayed on the X-axis (e.g., years, months). The range of values that the

characteristic of interest can take are displayed along the Y-axis (e.g., annual household income, mean years of schooling, average cost of child care). A data point is plotted coinciding with the value of the Y variable plotted for each of the values of the X variable, and a line is drawn connecting the points.

Scatter plots

- **Scatter plots** display the relationship between two quantitative or numeric variables by plotting one variable against the value of another variable
- The values of one of the two variables are displayed on the horizontal axis (x axis) and the values of the other variable are displayed on the vertical axis (y axis)
- Each person or subject in a study would receive one data point on the scatter plot that corresponds to his or her values on the two variables.

  For example, a scatter plot could be used to show the relationship between income and children's scores on a math assessment. A data point for each child in the study showing his or her math score and family income would be shown on the scatter plot. Thus, the number of data points would equal the total number of children in the study.

Geographic Information Systems (GIS)

- A Geographic Information System is computer software capable of capturing, storing, analyzing, and displaying geographically referenced information; that is, data identified according to location.
- Using a GIS program, a researcher can create a map to represent data relationships visually. For example, the National Center for Education Statistics creates maps showing the characteristics of school districts across the United States such as the percentage of children living in married couple households, median family incomes and percentage of population that speaks a language other than English. The data that are linked to school district location come from the American Community Survey.

Sociograms

- Display networks of relationships among variables, enabling researchers to identify the nature of relationships that would otherwise be too complex to conceptualize.

See the following for additional information about different graphic methods:

- Graphical Analytic Techniques
- MapED
- Geographic Information Systems

D. Analytical Techniques

Researchers use different analytical techniques to examine complex relationships between variables. There are three basic types of analytical techniques:

- Regression Analysis
- Grouping Methods
- Multiple Equation Models

Regression Analysis

Regression analysis assumes that the dependent, or outcome, variable is directly affected by one or more independent variables. There are four important types of regression analyses:

Ordinary least squares (OLS) regression

- OLS regression (also known as linear regression) is used to determine the relationship between a dependent variable and one or more independent variables.
- OLS regression is used when the dependent variable is continuous. Continuous variables, in theory, can take on any value with a range. For example, family child care expenses, measured in dollars, is a continuous variable.
- Independent variables may be nominal, ordinal or continuous. Nominal variables, which are also referred to as categorical variables, have two or more non-numeric or qualitative categories. Examples of nominal variables are children's gender (male, female), their parents' marital status (single, married, separated, divorced), and the type of child care children receive (center-based, home-based care). Ordinal variables are similar to nominal variables except it is possible to order the categories and the order has meaning. For example, children's families' socioeconomic status may be grouped as low, middle and high.
- When used to estimate the associations between two or more independent variables and a single dependent variable, it is called multiple linear regression.
- In multiple regression, the coefficient (i.e., standardized or unstandardized regression coefficient for each independent variable) tells you how much the dependent variable is expected to change when that independent variable increases by one, holding all the other independent variables constant.

Logistic regression

- Logistic regression (or logit regression) is a special form of regression analysis that is used to examine the associations between a set of independent or predictor variables and a dichotomous outcome variable. A dichotomous variable is a variable with only two possible values, e.g. child receives child care before or after the Head Start program day (yes, no).

- Like linear regression, the independent variables may be either interval, ordinal, or nominal. A researcher might use logistic regression to study the relationships between parental education, household income, and parental employment and whether children receive child care from someone other than their parents (receives nonparent care/does not receive nonparent care).

Hierarchical linear modeling (HLM)

- Used when data are nested. Nested data occur when several individuals belong to the same group under study. For example, in child care research, children enrolled in a center-based child care program are grouped into classrooms with several classrooms in a center. Thus, the children are nested within classrooms and classrooms are nested within centers.
- Allows researchers to determine the effects of characteristics for each level of nested data, classrooms and centers, on the outcome variables. HLM is also used to study growth (e.g., growth in children's reading and math knowledge and skills over time).

Duration models

- Used to estimate the length of time before a given event occurs or the length of time spent in a state. For example, in child care policy research, duration models have been used to estimate the length of time that families receive child care subsidies.
- Sometimes referred to as survival analysis or event history analysis.

Resources

See the following for additional information about regression methods and models:

- Review of Regression Techniques
- The Little Handbook of Statistical Practice
- Duration Models for Repeated Events

Grouping Methods

Grouping methods are techniques for classifying observations into meaningful categories. Two of the most common grouping methods are discriminant analysis and cluster analysis.

Discriminant analysis

- Identifies characteristics that distinguish between groups. For example, a researcher could use discriminant analysis to determine which characteristics identify families that seek child care subsidies and which identify families that do not.

- It is used when the dependent variable is a categorical variable (e.g., family receives child care subsidies [yes, no], child enrolled in family care [yes, no], type of child care child receives [relative care, non-relative care, center-based care]). The independent variables are interval variables (e.g., years of schooling, family income).

Cluster analysis

- Used to classify similar individuals together. It uses a set of measured variables to classify a sample of individuals (or organizations) into a number of groups such that individuals with similar values on the variables are placed in the same group. For example, cluster analysis would be used to group together parents who hold similar views of child care or children who are suspended from school.
- Its goal is to sort individuals into groups in such a way that individuals in the same group (cluster) are more similar to each other than to individuals in other groups.
- The variables used in cluster analysis may be nominal, ordinal or interval.

Resources

See the following for additional information about grouping methods:

- Review of Discriminant Function Analysis
- Review of Cluster Analysis

Multiple Equation Models

Multiple equation modeling, which is an extension of regression, is used to examine the causal pathways from independent variables to the dependent variable. For example, what are the variables that link (or explain) the relationship between maternal education (independent variable) and children's early reading skills (dependent variable)? These variables might include the nature and quality of mother-child interactions or the frequency and quality of shared book reading.

There are two main types of multiple equation models:

- Path analysis
- Structural equation modeling

Path analysis

Path analysis is an extension of multiple regression that allows researchers to examine multiple direct and indirect effects of a set of variables on a dependent, or outcome, variable. In path analysis, a direct effect measures the extent to which the dependent variable is influenced by an independent variable. An indirect effect measures the extent to which an independent variable's influence on the dependent variable is due to another variable.

- A path diagram is created that identifies the relationships (paths) between all the variables and the direction of the influence between them.
- The paths can run directly from an independent variable to a dependent variable (e.g., X→Y), or they can run indirectly from an independent variable, through an intermediary, or mediating, variable, to the dependent variable (e.g. X1→X2→Y).
- The paths in the model are tested to determine the relative importance of each.
- Because the relationships between variables in a path model can become complex, researchers often avoid labeling the variables in the model as independent and dependent variables. Instead, two types of variables are found in these models:
    - **Exogenous variables** are not affected by other variables in the model. They have straight arrows emerging from them and not pointing to them.
    - **Endogenous variables** are influenced by at least one other variable in the model. They have at least one straight arrow pointing to them.

Structural equation modeling (SEM)

Structural equation modeling expands path analysis by allowing for multiple indicators of unobserved (or latent) variables in the model. Latent variables are variables that are not directly observed (measured), but instead are inferred from other variables that are observed or directly measured. For example, children's school readiness is a latent variable with multiple indicators of children's development across multiple domains (e.g., children's scores on standardized assessments of early math and literacy, language, scores based on teacher reports of children's social skills and problem behaviors).

There are two parts to a SEM analysis. First, the measurement model is tested. This involves examining the relationships between the latent variables and their measures (indicators). Second, the structural model is tested in order to examine how the latent variables are related to one another. For example, a researcher might use SEM to investigate the relationships between different types of executive functions and word reading and reading comprehension for elementary school children. In this example, the latent variables word reading and reading comprehension might be inferred from a set of standardized reading assessments and the latent variables cognitive flexibility and inhibitory control from a set of executive function tasks. The measurement model of SEM allows the researcher to evaluate how well children's scores on the standardized reading assessments combine to identify children's word reading and reading comprehension. Assuming that the results of these analyses are acceptable, the researcher would move on to an evaluation of the structural model, examining the predicted relationships between two types of executive functions and two dimensions of reading.

SEM has several advantages over traditional path analysis:

- Use of multiple indicators for key variables reduces measurement error.
- Can test whether the effects of variables in the model and the relationships depicted in the entire model are the same for different groups (e.g., are the direct and indirect

effects of parent investments on children's school readiness the same for White, Hispanic and African American children).
- Can test models with multiple dependent variables (e.g., models predicting several domains of child development).

See the following for additional information about multiple equation models:

- Finding Our Way: An Introduction to Path Analysis (Streiner)
- An Introduction to Structural Equation Modeling (Hox & Bechger)

SECTION 3. DATA COLLECTION

Part I. Survey Research and Questionnaires
Part II. Administrative Data
Part III. Field Research

Part I. Survey Research and Questionnaires

Descriptions of key issues in survey research and questionnaire design are highlighted in the following sections. Modes of data collection approaches are described together with their advantages and disadvantages. Descriptions of commonly used sampling designs are provided and the primary sources of survey error are identified. Terms relating to the topics discussed here are defined in the Research Glossary.

A. Survey Research
B. Questionnaire Design
C. Modes of Survey Administration
D. Sampling
E. Sources of Error

A. Survey Research

Survey research is a commonly used method of collecting information about a population of interest. The population may be comprised of a group of individuals (e.g., children under age five, kindergartners, parents of young children) or organizations (e.g., early care and education programs, k-12 public and private schools). There are many different types of surveys, several ways to administer them, and different methods for selecting the sample of individuals or organizations that will be invited to participate. Some surveys collect information on all members of a population and others collect data on a subset of a population. Examples of the former are the National Center for Education Statistics' Common Core of Data and the Administration for Children and Families' Survey of Early Head Start Programs. A survey may be administered to a sample of individuals (or to the entire population) at a single point in time (cross-sectional survey) or the same survey may be administered to different samples from the population at different time points (repeat cross-sectional). Other surveys may be administered to the same sample of individuals at different time points (longitudinal survey). The Survey of Early Head Start Programs is an example of a cross-sectional survey and the National Household Education Survey Program is an example of a repeat cross-sectional survey. Examples of longitudinal surveys include the Head Start Family and Child Experiences Survey and the Early Childhood Longitudinal Study, Birth and Kindergarten Cohorts. Regardless of the type of survey, there are two key features of survey research:

- Questionnaires -- a predefined series of questions used to collect information from individuals.
- Sampling -- a technique in which a subgroup of the population is selected to answer the survey questions. Depending on the sampling method, the information collected may or may not be generalized to the entire population of interest.

The American Association for Public Opinion Research (AAPOR) offers recommendations on how to produce the best survey possible: Best Practices for Survey Research.

AAPOR also provides guidelines on how to assess the quality of a survey: Evaluating Survey Quality in Today's Complex Environment.

Advantages and Disadvantages of Survey Research

Advantages

- Surveys are a cost-effective and efficient means of gathering information about a population.
- Data can be collected from a large number of respondents. In general, the larger the number of respondents (i.e., the larger the sample size), the more accurate will be the information that is derived from the survey.
- Sampling using probability methods to select potential survey respondents makes it possible to estimate the characteristics (e.g., socio-demographics, attitudes, behaviors, opinions, skills, preferences and values) of a population without collecting data from all members of the population.
- Depending on the population and type of information sought, survey questionnaires can be administered in-person or remotely via telephone, mail, online and mobile devices.

Disadvantages

- Questions asked in surveys tend to be broad in scope.
- Surveys often do not allow researchers to develop an in-depth understanding of individual circumstances or the local culture that may be the root cause of respondent behavior.
- Respondents may be reluctant to share sensitive information about themselves and others.
- Respondents may provide socially desirable responses to the questions asked. That is, they may give answers that they believe the researcher wants to hear or answers that shed the best light on them and others. For example, they may over-report positive behaviors and under-report negative behaviors.
- A growing problem in survey research is the widespread decline in response rates, or percentage of those selected to participate who chose to do so.

B. Questionnaire Design

The two most common types of survey questions are closed-ended questions and open-ended questions.

- **Closed-Ended Questions**

    - The respondents are given a list of predetermined responses from which to choose their answer.
    - The list of responses should include every possible response and the meaning of the responses should not overlap.
    - An example of a close-ended survey question would be, "Please rate how strongly you agree or disagree with the following statement: 'I feel good about my work on the job.' Do you strongly agree, somewhat agree, neither agree nor disagree, somewhat disagree, or strongly disagree?"
    - A Likert scale, which is used in the example above, is a commonly used set of responses for closed-ended questions.
    - Closed-ended questions are usually preferred in survey research because of the ease of counting the frequency of each response.

- **Open-Ended Questions**

    - Survey respondents are asked to answer each question in their own words. An example would be, “In the last 12 months, what was the total income of all members of your household from all sources before taxes and other deductions?” Another would be, “Please tell me why you chose that child care provider?”
    - It is worth noting that a question can be either open-ended or close-ended depending on how it is asked. In the previous example, if the question on household income asked respondents to choose from a given set of income ranges instead, it would be considered close-ended.
    - Responses are usually categorized into a smaller list of responses that can be counted for statistical analysis.

A well-designed questionnaire is more than a collection of questions on one or more topics. When designing a questionnaire, researchers must consider a number of factors that can affect participation and the responses given by survey participants. Some of the things researchers must consider to help ensure high rates of participation and accurate survey responses include:

- It is important to consider the order in which questions are presented.
    - Sensitive questions, such as questions about income, drug use, or sexual activity, should generally be placed near the end of the survey. This allows a level of trust

or psychological comfort to be established with the respondent before asking questions that might be embarrassing or more personal.
    - Researchers also recommend putting routine questions, such as age, gender, and marital status, at the end of the questionnaire.
    - Questions that are more central to the research topic or question and that may serve to engage the respondent should be asked early. For example, a survey on children's early development that is administered to parents should ask questions that are specific to their children in the beginning or near the beginning of the survey.
- Double-barreled questions, which ask two questions in one, should never be used in a survey. An example of a double-barreled question is, "Please rate how strongly you agree or disagree with the following statement: 'I feel good about my work on the job, and I get along well with others at work.'" This question is problematic because survey respondents are asked to give one response for their feelings about two conditions of their job.
- Researchers should avoid or limit the use of professional jargon or highly specialized terms, especially in surveys of the general population.
- Question and response option text should use words that are at the appropriate reading level for research participants.
- The use of complex sentence structures should be avoided.
- Researchers should avoid using emotionally loaded or biased words and phrases.
- The length of a questionnaire is always a consideration. There is a tendency to try to ask too many questions and cover too many topics. The questionnaire should be kept to a reasonable length and only include questions that are central to the research question(s). The length should be appropriate to the mode of administration. For example, in general, online surveys are shorter than surveys administered in-person.

Questionnaires and the procedures that will be used to administer them should be pretested (or field-tested) before they are used in a main study. The goal of the pretest is to identify any problems with how questions are asked, whether they are understood by individuals similar to those who will participate in the main study, and whether response options in close-ended questions are adequate. For example, a parent questionnaire that will be used in a large study of preschool-age children may be administered first to a small (often non-random) sample of parents in order to identify any problems with how questions are asked and understood and whether the response options that are offered to parents are adequate. Based on the findings of the pretest, additions or modifications to questionnaire items and administration procedures are made prior to their use in the main study.

Resources

See the following for more information about questionnaire design:

- A Brief Guide to Questionnaire Development
- Survey Design

C. Modes of Survey Administration

Surveys can be administered in four ways: through the mail, by telephone, in-person, or online. When deciding which of these approaches to use, researchers consider: the cost of contacting the study participant and of data collection, the literacy level of participants, response rate requirements, respondent burden and convenience, the complexity of the information that is being sought and the mix of open-ended and close-ended questions.

Some of the main advantages and disadvantages of the different modes of administration are summarized below.

Mail Surveys

- Advantages: Low cost; respondents may be more willing to share information and to answer sensitive questions; respondent convenience, can respond on their own schedule
- Disadvantages: Generally lower response rates; only reaches potential respondents who are associated with a known address; not appropriate for low literacy audiences; no interviewer, so responses cannot be probed for more detail or clarification; participants' specific concerns and questions about the survey and its purpose cannot be addressed

Telephone Surveys

- Advantages: Higher response rates; responses can be gathered more quickly; responses can be probed; participants' concerns and questions can be addressed immediately
- Disadvantages: More expensive than mail surveys; depending on how telephone numbers are identified, some groups of potential respondents may not be reached; use of open-ended questions is limited given limits on survey length

In-Person Surveys

- Advantages: Highest response rates; better suited to collecting complex information; more opportunities to use open-ended questions and to probe respondent answers; interviewer can immediately address any concerns participant has about the survey and answer their questions
- Disadvantages: Very expensive; time-consuming; respondents may be reluctant to share personal or sensitive information when face-to-face with an interviewer

Online Surveys

- Advantages: Very low cost; responses can be gathered quickly; respondents may be more willing to share information and to answer sensitive questions; questionnaires are programmed, which allows for more complex surveys that follow skip patterns based on previous responses; respondent convenience, can respond on their own schedule
- Disadvantages: Potentially lower response rates; limited use of open-ended questions; not possible to probe respondents' answers or to address their concerns about participation

Increasingly, researchers are using a mix of these methods of administration. Mixed-mode or multi-mode surveys use two or more data collection modes in order to increase survey response. Participants are given the option of choosing the mode that they prefer, rather than this being dictated by the research team. For example, the Head Start Family and Child Experience Survey (2014-2015) offers teachers the option of completing the study's teacher survey online or using a paper questionnaire. Parents can complete the parent survey online or by phone.

Resources

See the following for additional information about survey administration:

Four Survey Methodologies: A Comparison of Pros and Cons
Collecting Survey Data
Improving Response to Web and Mixed-Mode Surveys

D. Sampling

In child care and early education research as well as research in other areas, it is often not feasible to survey all members of the population of interest. Therefore, a sample of the members of the population would be selected to represent the total population. A primary strength of sampling is that estimates of a population's characteristics can be obtained by surveying a small proportion of the population. For example, it would not be feasible to interview all parents of preschool-age children in the U.S. in order to obtain information about their choices of child care and the reasons why they chose certain types of care as opposed to others. Thus, a sample of preschoolers' parents would be selected and interviewed, and the data they provide would be used to estimate the types of child care parents as a whole choose and their reasons for choosing these programs. There are two broad types of sampling:

- **Nonprobability sampling**: The selection of participants from a population is not determined by chance. Each member of the population does not have a known or given chance of being selected into the sample. Findings from nonprobability (nonrandom) samples cannot be generalized to the population of interest. Consequently, it is problematic to make inferences about the population. Common nonprobability sampling techniques include convenience sampling, snowball sampling, quota sampling and purposive sampling.

- **Probability sampling**: The selection of participants from the population is determined by chance and with each individual having a known, non-zero probability of selection. It provides accurate descriptions of the population and therefore good generalizability. In

survey research, it is the preferred sampling method. Three forms of probability sampling are described here:

- **Simple Random Sampling**
  This is the most basic form of sampling. Every member of the population has an equal chance of being selected. This sampling process is similar to a lottery: the entire population of interest could be selected for the survey, but only a few are chosen at random.

  For example, researchers may use random-digit dialing to perform simple random sampling for telephone surveys. In this procedure, telephone numbers are generated by a computer at random and called to identify individuals to participate in the survey.

- **Stratified Sampling**
  Stratified sampling is used when researchers want to ensure representation across groups, or strata, in the population. The researchers will first divide the population into groups based on characteristics such as race/ethnicity, and then draw a random sample from each group. The groups must be mutually exclusive and cover the population. Stratified sampling provides greater precision than a simple random sample of the same size.

- **Cluster Sampling**
  Cluster sampling is generally used to control costs and when it is geographically impossible to undertake a simple random sample. For example, in a household survey with face-to-face interviews, it is difficult and expensive to survey households across the nation using a simple random sample design. Instead, researchers will randomly select geographic areas (for example, counties), then randomly select households within these areas. This creates a cluster sample, in which respondents are clustered together geographically.

Survey research studies often use a combination of these probability methods to select their samples. **Multistage sampling** is a probability sampling technique where sampling is carried out in several stages. It is often used to select samples when a single frame is not available to select members for a study sample. For example, there is no single list of all children enrolled in public school kindergartens across the U.S. Therefore, researchers who need a sample of kindergarten children will first select a sample of schools with kindergarten programs from a school frame (e.g., National Center for Education Statistics' Common Core of Data) (Stage 1). Lists of all kindergarten classrooms in selected schools are developed and a sample of classrooms selected in each of the sampled schools (Stage 2). Finally, lists of children in the sampled classrooms are compiled and a sample of children is selected from each of the classroom lists (Stage 3). Many of the national surveys of child care and early education (e.g., the Head Start Family and Child Experiences Survey and the Early Childhood Longitudinal Survey-Kindergarten Cohort) use a multistage approach.

Multistage, cluster and stratified sampling require that certain adjustments be made during the statistical analysis. Sampling or analysis weights are often used to account for differences in the probability of selection into the sample as well as for other factors (e.g., samplig frame undercoverage and nonresponse). Standard errors are calculated using methodologies that are different from those used for a simple random sample. Information on these adjustments is provided by the National Center for Education Statistics through its Distance Learning Dataset Training System.

Resources

See the following for additional information about the different types of sampling approaches and their use:

- National Center for Education Statistics Distance Learning Dataset Training System
- Sampling in Developmental Science: Situations, Shortcomings, Solutions, and Standards
- Nonprobability Sampling
- The Future of Survey Sampling
- Sampling Methods (StatPac)

E. Sources of Error

Estimates of the characteristics of a population using survey data are subject to two basic sources of error: sampling error and nonsampling error. The extent to which estimates of the population mean, proportion and other population values differ from the true values of these is affected by these errors.

- **Sampling error** is the error that occurs because all members of the population are not sampled and measured. The value of a statistic (e.g., mean or percentage) that is calculated from different samples that are drawn from the same population will not always be the same. For example, if several different samples of 5,000 people are drawn at random from the U.S. population, the average income of the 5,000 people in those samples will differ. (In one sample, Bill Gates may have been selected at random from the population, which would lead to a very high mean income for that sample.)

  Researchers use a statistic called the standard error to measure the extent to which estimated statistics (percentages, means, and coefficients) vary from what would be found in other samples. The smaller the standard error, the more precise are the estimates from the sample. Generally, standard errors and sample size are negatively related, that is, larger samples have smaller standard errors.

- **Nonsampling error** includes all errors that can affect the accuracy of research findings other than errors associated with selecting the sample (sampling error). They can occur

in any phase of a research study (planning and design, data collection, or data processing). They include errors that occur due to coverage error (when units in the target population are missing from the sampling frame), nonresponse to surveys (nonresponse error), measurement errors due to interviewer or respondent behavior, errors introduced by how survey questions were worded or by how data were collected (e.g., in-person interview, online survey), and processing error (e.g., errors made during data entry or when coding open-ended survey responses). While sampling error is limited to sample surveys, nonsampling error can occur in all surveys.

Measurement Error

Measurement error is the difference between the value measured in a survey or on a test and the true value in the population. Some factors that contribute to measurement error include the environment in which a survey or test is administered (e.g., administering a math test in a noisy classroom could lead children to do poorly even though they understand the material), poor measurement tools (e.g., using a tape measure that is only marked in feet to measure children's height would lead to inaccurate measurement), rater or interviewer effects (e.g., survey staff who deviate from the research protocol).

Measurement error falls into two broad categories: systematic error and random error. Systematic error is the more serious of the two.

- **Systematic error**
  - Occurs when the survey responses are systematically different from the target population responses. It is caused by factors that systematically affect the measurement of a variable across the sample.
  - For example, if a researcher only surveyed individuals who answered their phone between 9 and 5, Monday through Friday, the survey results would be biased toward individuals who are available to answer the phone during those hours (e.g., individuals who are not in the labor force or who work outside of the traditional Monday through Friday, 9 am to 5 pm schedule).
  - It can include both nonobservational and observational error.
    - Nonobservational error -- Error introduced when individuals in the target population are systematically excluded from the sample, such as in the example above.
    - Observational error -- Error introduced when respondents systematically answer survey question incorrectly. For example, surveys that ask respondents how much they weigh may underestimate the population's weight because some respondents are likely to report their weight as less than it actually is.
  - Systematic errors tend to have an effect on responses and scores that is consistently in one direction (positive or negative). As a result, they contribute to bias in estimates.

- **Random error**

    - Random error is an expected part of survey research, and statistical techniques are designed to account for this sort of measurement error. It is caused by factors that randomly affect measurement of the variable across the sample.
    - Random error occurs because of natural and uncontrollable variations in the survey process, i.e., the mood of the respondent, lack of precision in measures used, and the particular measures/instruments (e.g., inaccuracy in scales used to measure children's weight).
      For example, a researcher may administer a survey about marital happiness. However, some respondents may have had a fight with their spouse the evening prior to the survey, while other respondents' spouses may have cooked the respondent's favorite meal. The survey responses will be affected by the random day on which the respondents were chosen to participate in the study. With random error, the positive and negative influences on the survey measures are expected to balance out.
    - Unlike systematic errors, random errors do not have a consistent positive or negative effect on measurement. Instead, across the sample the effects are both positive and negative. Such errors are often considered noise and add variability, though not bias, to the data.

Resources

See the following for additional information about the different types and sources of errors:

- Nonresponse Error, Measurement Error and Mode of Data Collection
- Total Survey Error: Design, Implementation, and Evaluation
- Data Accuracy

Part II. Administrative Data

Administrative data are an important source of information for social science research. For example, school records have been used to track trends in student academic performance. Administrative data generally refers to data collected as part of the management and operations of a publicly funded program or service. Today, use of administrative data is becoming increasingly common in research about child care and early education. These data often are a relatively cost-effective way to learn more about the individuals and families using a particular service or participating in a particular program, but they do have some important limitations.

The advantages and disadvantages of using administrative data are described here. Issues pertaining to the access to such data are discussed. Terms relating to administrative data and its use in research studies are defined in the Research Glossary.

A. Advantages of Administrative Data
B. Data Limitations
C. Obtaining and Learning About Administrative Data
D. Sampling

A. Advantages of Administrative Data

- Administrative data make possible analyses at the state and local levels that are rarely possible using national survey data.
- Such data often contain detailed, accurate measures of participation in various social programs. They typically include large numbers of cases, making possible many different types of analyses.
- Data on the same individuals and/or same programs over a long period of time can be used for longitudinal and trend studies.
- Potential for linking data from several programs in order to get a more complete picture of individuals and the services received.
- At the state level, such data provide effective ways for assessing state-specific programs and can be useful for several forms of program evaluation.
- The large sample sizes allow small program effects to be more easily detected, and permit effects to be estimated for different groups.
- It is less expensive to obtain administrative data than to collect data directly on the same group.

B. Data Limitations

Administrative data are collected to manage services and comply with government reporting regulations. Because the original purpose of the data is not research, this presents several challenges.

- The administrative data only describe the individuals or families using a service and provide no information about similar people who do not use the service.
- The potential observation period for any subject being studied (e.g., a person, a family, a child care program) is limited to the period of time that the subject is using the service for which the data are being collected.
- Generally, only those services that are publicly funded are included in the administrative data. For example, a researcher cannot rely on subsidy data to learn about all child care

providers in the state or on non-subsidized forms of child care being used to augment child care that is subsidized.

- Many variables used in administrative data are not updated regularly, so it is important to learn how and when each variable is collected. For instance, an "earnings" variable in administrative data for subsidized child care generally is entered at the time that eligibility is determined and then updated when eligibility is redetermined. When this is the case, there is no way to know, using administrative data alone, what a family earns in the months between eligibility determination and redetermination.
- Important variables needed for a particular research study may not be collected in administrative data.
- Because the data are limited to data on program participants, information on those eligible for the program but who are not enrolled is often not available. Thus, administrative data may not be especially useful for estimating certain characteristics such as participation rates.
- Measurement error can pose a substantial challenge to analysts using administrative data. Factors affecting measurement error include:
    - Data that were improperly entered at the agency
    - Incomplete or inaccurate data items, particularly those items not required by the agency for management or reporting purposes
    - Missing values on variables that have been overwritten by updated versions when cases are reviewed
- Procedures for accessing the data for research purposes can be time consuming and difficult. Protecting the privacy of program participants and the confidentiality of the data when they are used for research is a major concern to program officials.

Researchers interested in using administrative data for the purpose of research should expect to invest considerable time learning about the details of the administrative data system, the specific data elements being used, the data entry process and standards, and changes in the data system and data definitions over time. It also takes time to transform administrative data into research datasets that can be used in statistical analyses.

C. Obtaining and Learning About Administrative Data

Important issues usually confront researchers who have decided to use administrative data records in their research. Among the most important of these issues are:

- Obtaining Data Access and Ensuring Confidentiality
  To obtain administrative data, a researcher and the agency responsible for the data must reach an agreement on how the data are to be used and processed, how confidentiality will be maintained and how the research results will be disseminated.
- Documentation of Source Data

    - Once researchers have obtained the administrative records of interest, they must become familiar with the idiosyncrasies of the data.
    - When combining data from more than one administrative database, researchers must be careful to assess the comparability of the data elements and the effectiveness of record matching procedures.
    - Researchers must also learn about, understand, and document changes in the definition and meaning of data elements over time as well the procedures for updating data values.
    - Researchers should carefully document variable definitions, value codes, any recodes that the agency implemented, changes in definitions and their effective dates and information on how the agency collected the data.
- Documentation of Program Parameters and Context. In addition to documenting the source data, investigators should also take great care to document the important parameters of the program that collected the data and to describe the policy context at the time the data were collected. For example, whether all those who were eligible and applied for the program were able to be served, when there were caps on available funding that may have limited service provision.

D. Sampling

Sampling is not often done when administrative data are used for research purposes since information are available on the entire population of recipients. However, in order to ensure the protection of subject confidentiality a subsample from the full population may be selected. Studies that combine the use of survey research and administrative data records may also select only a sample of the population in order to minimize data collection costs.

Resources

The Joint Center for Poverty Research offers many recommendations on using administrative data. It recognizes the following centers as having successfully used administrative data in their research efforts:

- Child Welfare Research Center
- Ray Marshall Center for the Study of Human Resources at the University of Texas at Austin
- Chapin Hall Center for Children at the University of Chicago
- University of Maryland School of Social Work

More resources on administrative data integration, analyses, management, confidentiality, and security can be found here: Working with Administrative Data. Also, see Profiles in Success of Statistical Uses of Administrative Data for more information on the use of administrative data.

Part III. Field Research

Field research is a qualitative method of research concerned with understanding and interpreting the social interactions of groups of people, communities, and society by observing and interacting with people in their natural settings. The methods of field research include: direct observation, participant observation, and qualitative interviews. Each of these methods is described here. Terms related to these and other topics in field research are defined in the Research Glossary.

A. Direct Observation
B. Participant Observation
C. Qualitative Interviews

A. Direct Observation

Direct observation is a method of research where the researcher watches and records the activities of individuals or groups engaged in their daily activities. The observations may be unstructured or structured. Unstructured observations involve the researcher observing people and events and recording his/her observations as field notes. Observations are recorded holistically and without the aid of a predetermined guide or protocol. Structured observation, on the other hand, is a technique where a researcher observes people and events using a guide or set protocol that has been developed ahead of time.

Other features of direct observation include:

- The observer does not actively engage the subjects of the study in conversations or interviews, but instead strives to be unobtrusive and detached from the setting.
- Data collected through direct observation may include field notes, checklists and rating scales, documents, and photographs or video images.
- Direct observation is not necessarily an alternative to other types of field methods, such as participant observation or qualitative interviews. Rather, it may be an initial approach to understanding a setting, a group of individuals, or forms of behavior prior to interacting with members or developing interview protocols.
- Direct observation as a research method is most appropriate in open, public settings where anyone has a right to be or congregate. Conducting direct observation in private or closed settings -- without the knowledge or consent of members -- is more likely to raise ethical concerns.

B. Participant Observation

Participant observation is a field research method whereby the researcher develops an understanding of a group or setting by taking part in the everyday routines and rituals alongside its members. It was originally developed in the early 20th century by anthropologists researching native societies in developing countries. It is now the principal research method

used by ethnographers -- specialists within the fields of anthropology and sociology who focus on recording the details of social life occurring in a setting, community, group, or society. The ethnographer, who often lives among the members for months or years, attempts to build trusting relationships so that he or she becomes part of the social setting. As the ethnographer gains the confidence and trust of the members, many will speak and behave in a natural manner in the presence of the ethnographer.

Data from participant observation studies can take several forms:

- Field notes are the primary type of data. The researcher takes notes of his/her observations and experiences and later develops them into detailed, formal field notes.
- Frequently, researchers keep a diary, which is often a more intimate, informal record of the happenings within the setting.
- The practice of participant observation, with its emphasis on developing relationships with members, often leads to both informal, conversational interviews and more formal, in-depth interviews. The data from these interviews can become part of field notes or may consist of separate interview transcripts.

There are a number of advantages and disadvantages to direct and participant observation studies. Here is a list of some of both. While the advantages and disadvantages apply to both types of studies, their impact and importance may not be the same across the two. For example, researchers engaged in both types of observation will develop a rich, deep understanding of the members of the group and the setting in which social interactions occur, but researchers engaged in participant observation research may gain an even deep understanding. And, participant observers have a greater chance of witnessing a wider range of behaviors and events than those engaged in direct observation.

Advantages of observation studies (observational research):

- Provide contextual data on settings, interactions, or individuals.
- A useful tool for generating hypotheses for further study.
- Source of data on events and phenomena that do not involve verbal interactions (e.g., mother-child nonverbal interactions and contact, physical settings where interactions occur).
- The researcher develops a rich, deep understanding of a setting and of the members within the setting.

Disadvantages of observation studies:

- Behaviors observed during direct observation may be unusual or atypical.
- Significant interactions and events may take place when observer is not present.
- Certain topics do not necessarily lend themselves to observation (e.g., attitudes, emotions, affection).

- Reliability of observations can be problematic, especially when multiple observers are involved.
- The researcher must devote a large amount of time (and resources).
- The researcher's objectivity may decline as he or she spends more time among the members of the group.
- The researcher may be faced with a dilemma of choosing between revealing and not revealing his or her identity as a researcher to the members of the group. If he or she introduces him/herself as a researcher, the members may behave differently than if they assume that he or she is just another participant. On the other hand, if the researcher does not, they may feel betrayed upon learning about the research.

C. Qualitative Interviews

Qualitative interviews are a type of field research method that elicits information and data by directly asking questions of individuals. There are three primary types of qualitative interviews: informal (conversational), semi-structured, and standardized, open-ended. Each is described briefly below along with advantages and disadvantages.

Informal (Conversational) Interviews

- Frequently occur during participant observation or following direct observation.
- The researcher begins by conversing with a member of the group of interest. As the conversation unfolds, the researcher formulates specific questions, often spontaneously, and begins asking them informally.
- Appropriate when the researcher wants maximum flexibility to pursue topics and ideas as they emerge during the exchange

Advantages of informal interviewing:

- Allows the researcher to be responsive to individual differences and to capture emerging information.
- Information that is obtained is not constrained by a predetermined set of questions and/or response categories.
- Permits researcher to delve deeper into a topic and what key terms and constructs mean to study participants.

Disadvantages of informal interviewing:

- May generate less systematic data, which is difficult to classify and analyze.
- The researcher might not be able to capture everything that the interviewee is saying and therefore there is potential for important nuance or information to be lost. For example, the researcher might not have a tape recorder at that moment due to the spontaneous nature of these interviews.
- Quality of the information obtained depends on skills of the interviewer.

Semi-Structured Interviews

- Prior to the interview, a list of predetermined questions or probes, also known as an interview guide, is developed so that each interviewee will respond to a similar series of questions and topics.
- Questions are generally open-ended to elicit as much detail and meaning from the interviewee as possible.
- The researcher is free to pursue and probe other topics as they emerge during the interview.

Advantages of semi-structured interviewing:

- Systematically captures data across interviewees.
- The researcher is able to rephrase or explain questions to the interviewee to ensure that everyone understands the questions the same way and probe (follow-up) a response so that an individual's responses are fully explored.
- Interviewee is allowed the freedom to express his or her views in their own words.

Disadvantages of semi-structured interviewing:

- Does not offer as much flexibility to respond to new topics that unfold during the interview as the informal interview.
- Responses to questions that have been asked in slightly different ways can be more difficult to compare and analyze.
- Quality of the information obtained depends on skills of the interviewer.
- Interviewer may unconsciously send signals about the types of answers that are expected.

Standardized, Open-Ended Interviews

- Similar to a survey since questions are carefully scripted and written prior to the interview, which serves to minimize variability in question wording and the way questions are asked.
- The researcher asks a uniform series of questions in the same order to each interviewee.
- The questions are open-ended to capture more details and individual differences across interviewees.
- Particularly appropriate for qualitative studies involving multiple interviewers.

Advantages of standardized interviewing:

- All questions are asked the same to each study participant. Data are comparable across interviewees.
- Reduces interviewer effects when several interviewers are used.
- Standardization helps to facilitate the processing and analysis of the data.

Disadvantages of standardized interviewing:

- Does not offer as much flexibility to respond to and probe new topics that unfold during the interview.
- Standardized wording of questions may limit the responses of those being interviewed.

Both standardized and semi-structured interviews involve formally recruiting participants and are typically tape-recorded. The researcher should begin with obtaining informed consent from the interviewee prior to starting the interview. Additionally, the researcher may write a separate field note to describe the interviewee's reactions to the interview, or events that occurred before or after the interview.

Resources

See the following for additional information about field research and qualitative research methods.

Ethnography, Observational Research and Narrative Inquiry
An Introduction to Qualitative Research

SECTION 4. ASSESSING RESEARCH QUALITY

The quality of social science and policy research can vary considerably. It is important that consumers of research keep this in mind when reading the findings from a research study or when considering whether or not to use data from a research study for secondary analysis. This section includes information and tools to help evaluate the quality of a research study. It also includes information on the ethics of research.

Part I. Key Questions to Ask
Part II. Research Assessment Tools
Part III. Ethics of Research

Part I. Key Questions to Ask

A. **Was the research peer reviewed?**
Peer reviewed research studies have already been evaluated by experienced researchers with relevant expertise. Most journal articles, books and government reports have gone through a peer review process. Keep in mind that there are many types of peer reviews. Reports issued by the federal government have been subject to many levels of internal review and approval before being issued. Articles published in professional journals with peer review have been evaluated by researchers that are experts in the field and who can vouch for the soundness of the methodology and the analysis applied. As a result, peer-reviewed research is usually of high quality. A research consumer, however, should still critically evaluate the study's methodology and conclusions.

B. **Can a study's quality be evaluated with the information provided?**
Every study should include a description of the population of interest, an explanation of the process used to select and gather data on study subjects, definitions of key variables and concepts, descriptive statistics for main variables, and a description of the analytic techniques. Research consumers should be cautious when drawing conclusions from studies that do not provide sufficient information about these key research components.

C. **Are there any potential threats to the study's validity?**
A valid study answers research questions in a scientifically rigorous manner. Threats to a study's validity are found in three areas: Internal, external and content validity.

1. **Internal validity** refers to whether the outcomes observed in a study are due to the independent variables or experimental manipulations investigated in the study and not to some other factor or set of factors. To determine whether a research study has internal validity, a research consumer should ask whether changes in the outcome could be attributed to alternative explanations that are

not explored in the study. For example, a study may show that a new curriculum had a significant positive effect on children's reading comprehension.
The study must rule out alternative explanations for the increase in reading comprehension, such as a new teacher, in order to attribute the increase in reading comprehension to the new curriculum. Studies that specifically explain how alternative explanations were ruled out are more likely to have internal validity. Threats to a study's internal validity can compromise the confidence consumers have in the findings from a study and include:

I. The introduction of events while the study is being conducted that may affect the outcome or dependent variable of the study. For example, while studying the effectiveness of children's participation in an early childhood program, the program was closed for an extended period of time due to damage from a hurricane.

II. Changes in the dependent variable due to normal developmental processes in study participants. For example, young children's performance on a battery of outcome measures (e.g., reading and math assessments) may decline during the testing or observation period due to fatigue or other factors.

III. The circumstances around the testing that is used to assess the dependent variable. For example, preschool children's performance on a standardized test may be questionable if test items are presented to children in unfamiliar ways or in group settings.

IV. Participants leaving or dropping out of the study before it is completed. This can be especially problematic if those who leave the study are different from those who stay. For example, in a longitudinal study of the effects of a school lunch program on children's academic achievement, the validity of the findings could be problematic if the most disadvantaged children in the program left the study at a higher rate than other children.

V. Changes to or inconsistencies in how the dependent and independent variables were measured. For example, changing the way in which children's math skills are measured at two time points could introduce error if the two measures were developed using different assessment frameworks (i.e., they were developed to assess different math content and processes). Inconsistencies are also introduced when different staff follow different procedures when administering the same measure. For example, when administering an assessment to bilingual children, some staff give children credit for answering correctly in English or Spanish, and other staff only give credit for answering correctly in English.

VI. Statistical regression or regression to the mean can affect the outcome of a study. It is the movement of test scores (post-test scores) toward the mean (average score), independent of any effect of an independent variable. It is especially a concern when assessing the skills of low performing individuals and comparing their skills to those with average or above average performance. For example, kindergarten children with the weakest reading skills at the start of the school year may show the greatest gains in their skills over the school year (e.g., between fall and spring assessments) independent of the instruction they received from their teachers.

2. **External validity** refers to the extent to which the results of a study can be generalized to other settings (ecological validity), other people (population validity) and over time (historical validity). To assess whether a study has external validity, a research consumer should ask whether the findings apply to individuals whose place and circumstances differ from those of study participants. For example, a research study shows that a new curriculum improved reading comprehension of third-grade children in Iowa. As a research consumer, you want to ask whether this new curriculum may also be effective with third graders in New York or with children in other elementary grades. Studies that randomly select participants from the most diverse and representative populations and that are conducted in natural settings are more likely to have external validity. Threats to a study's external validity come from several sources, including:

    I. The sample is not representative of the population of interest. As a result, findings from the study may be biased (sample selection bias) and do not accurately represent the population. Several factors can lead to a sample not being representative of the population.

        a. The list of all those in the population who are eligible to be sampled is incomplete or contains duplicates. For example, in a household survey, the list of housing units from which the sample will be drawn may be missing housing units (e.g., one or the two housing units in a duplex home). Or, an address list that will be used to drawn a sample may have some households listed twice.

        b. Some members of the population or members of certain groups may not be adequately represented in the sample (undercoverage). For example, a survey of adult education that relies on a published list of telephone numbers to select its sample may not get an accurate estimate of the participation of adults in different education programs because young adults who

have higher rates of participation are less likely to have landlines and to have numbers published.

c. Not all individuals who are sampled agree to participate in the study. When those who participate are different in meaningful ways from those who do not, there is the potential for the findings from the study to be biased (nonresponse bias). That is, the findings may not represent an accurate picture of the total population.

d. Selecting samples using non-probability methods (e.g., purposive sample, volunteer samples), which tend to over- or under-represent certain groups in the population. For example, volunteer surveys on controversial topics such as school vouchers and sex education are more likely to overrepresent individuals with strong opinions. And, shopping mall surveys in general only represent the small group of individuals who are shopping at a particular location and at specific times.

II. The findings from one study are difficult to replicate across locations, groups, and time. Despite the best efforts, it is extremely difficult to introduce and implement a program (treatment) exactly the same way in different locations. Similarly, it is difficult to conduct a study the same way each time. While researchers have control over many features of their studies, there are factors that are beyond their control (e.g., willingness of potential subjects to participate, scheduling conflicts that could lead to cancellations of data collection activities, data collection being suspended due to natural disasters). For example, the ability to carry out a study of school-age children's reading and math achievement in one school or in one school district may be affected by teachers' willingness to surrender instructional time for students to participate in a series of standardized assessments. In some cases, modifications to the study design (e.g., shorten the assessment, limit sensitive questions on a teacher or parent survey) must be made to accommodate the concerns of school and district leaders.

III. Changes in the behaviors and reported attitudes of study participants as a result of being included in a research study (Hawthorne effect). For example, parents participating in a research study on children's early development may change the ways in which they support their child's learning at home.

3. **Construct validity** refers to the degree to which a variable, test, questionnaire or instrument measures the theoretical concept that the researcher hopes to

measure. To assess whether a study has construct validity, a research consumer should ask whether the study has adequately measured the key concepts in the study. For example, a study of reading comprehension should present convincing evidence that reading tests do indeed measure reading comprehension. Studies that use measures that have been independently validated in prior studies are more likely to have construct validity.

There are many threats to construct validity. These can arise during: the planning and design stage, assessment or survey administration, and data processing and analysis. Some are attributed to researchers and others to the subjects of the research. Here are some of the more common threats:

I. Threats that occur during the planning and design stage include:

   a. Poorly defined constructs are perhaps the largest threat to construct validity. This applies to constructs that are too narrowly defined as well as those that are defined too broadly.

   b. Validity can also be affected by the measures a researcher chooses to measure a construct. Measures that include too few items to adequately represent the construct pose a threat as do measures that include items that tap other constructs. For example, a math assessment administered to four- and five-year old children that only includes items that require children to count would not be adequate to represent their math skills. A math assessment administered to this same group of children that was made up mostly of word problems would be tapping both their math and language skills. A valid measure should cover all aspects of the theoretical construct and only aspects of the theoretical construct.

   c. Assessment items or survey questions that are poorly written are threats to validity. Such items would include double-barreled questions that ask multiple questions within a single item (e.g., are you happily married and do you and your spouse argue?). Other examples of poorly written questions include those that use language that is above the reading level of most respondents, use professional jargon or are written in such a way as to trigger a socially desirable response.

   d. The validity of an assessment is threatened if there are too many items that are outside the ability of the individual being assessed (e.g., too many very easy items and too many very difficult items). For example, an early literacy assessment that only included passages that children were asked to read and answer questions

about would not result in a valid assessment of children's early literacy.

II. Threats that occur during administration include:

a. Threats that are introduced by interviewers and assessment staff. Actions by these individuals that can affect the reliability and thus the validity of the assessment occur when they deviate from the research protocol and when they signal a correct answer to the study participant through their actions. For example, an assessment of young children's English language vocabulary may specify that only responses in English are acceptable. However, when assessing bilingual children, some assessors comply with this rule while others accept responses in English or in the child's home language (e.g., Spanish). Assessors may unintentionally signal to children the correct responses on an assessment by 'staring' at the correct response to a multiple-choice item or by smiling and giving praise only when the child answers correctly.

b. Threats to validity can also be introduced by the research participants. These would include participant apprehension or anxiety that could result in poorer performance on an assessment or to incorrect or ambiguous responses to a series of interview items. These threats must be taken seriously and addressed when administering standardized assessments to young children, many of whom will have limited experience with these types of tests. The language used when administering an assessment can also threaten its validity, if subjects do not have the language skills to understand what they are being asked to do and the language skills needed to respond.

III. Threats that occur during data processing and analysis include:

a. Coding errors - Coding errors that are systematic as compared to those that are random are especially problematic.

b. Poor inter-coder or inter-rater reliability - When coding responses to open-ended survey items or assigning scores to behaviors observed during a video interaction, it is important that different coders or raters assign the same code or score for the same response or behavior. That is, the goal is high inter-coder or inter-rater reliability. When inter-coder or inter-rater reliability is poor, it can have an adverse effect on the validity of a measure. For example, the construct validity of an observation measure of the

quality of parent-child interactions could be compromised should individual members of a group coding a set of videotaped mother-child interactions apply different standards as to what they deem as intrusive parenting practices.

c. Inconsistencies in how data are analyzed and missing data handled - Missing data may be handled in a number of different ways, and the approach that is chosen could prove to be problematic for a construct, especially when the data are not missing at random. For example, if items tapping certain math skills are missing disproportionately, the validity of the measure could be jeopardized if a researcher assigns the mean score for those items or if he simply averages the scores for the non-missing items.

Part II. Research Assessment Tools

The purpose of the quantitative and qualitative research assessment tools is to provide users with a quick and simple means to evaluate the quality of research studies. The research assessment tools describe the information that should be available in study reports and the key features of a high quality study design. When using tools, higher scores indicate higher quality research.

Quantitative Research Assessment Tool
Qualitative Research Assessment Tool

Resources

See the following for additional information on assessing the quality of research.
Early childhood program evaluations: A decision-maker's guide
Early childhood assessment: Why, what, and how?
Education Commission of the States (ECS) and Mid-continent Research for Education and Learning (McREL)

Part III. Ethics of Research

Researchers are expected to adhere to the principles of ethical research. The Belmont Report provides a broad framework for the ethics of research involving human subjects. Three basic ethical principles are identified:

- Respect for persons -- requires that research subjects are not coerced into participating in a study and requires the protection of research subjects who have diminished autonomy.
- Beneficence -- requires that research does not harm research subjects, and that researchers minimize the risks for subjects while maximizing the benefits for them.
- Justice -- requires that all forms of differential treatment among research subjects be justified.

Applications of these principles lead to considerations of the following:

- **Informed Consent**
  Participants should give informed consent before participating in a study. In order for participants to give informed consent.

  - The researcher must inform the participants of the study's purpose, content, duration, and potential risks and benefits.
  - The researcher must inform the participants that they can stop participating in the study at any point.
  - In the event of survey research, the researcher must inform the participants that they do not have to answer all the survey questions.
  - If the participants are children under legal age, the researcher must seek consent from their parents or guardians.

- **Confidentiality**
  Unless consent is given otherwise, it is absolutely imperative that researchers keep participants' identities confidential. Confidentiality means that participants cannot be identified in any way. In survey research, this includes but is not limited to making sure that participants' identifiers are not linked to their survey responses. Common identifiers include names, social security numbers, addresses, and telephone numbers. Such Personal Identifying Information or PII must be safeguarded. When analyzing data collected from small groups or samples with small n's and when reporting the findings from these analyses, the researchers must be extra mindful of not revealing participants' identities. Cell sizes with fewer than three cases should not be reported because information about the individuals in this group could be obtained by subtraction.

- **Anonymity**
  Anonymity is an even stronger safeguard of participant privacy. If a researcher assures anonymity, it means that the researcher is unable to link participants' names to the information they provide.

In addition to the above principles, considerations of specific ethical issues are often required depending on the form and context of research. For example, when using administrative data, the researcher must keep in mind that there are many legal protections set by the federal and state governments that require the privacy of program applicant information. For instance, in 1977, the Privacy Protection Study Commission determined that records or information used for statistical research could not be used in an individually identifiable form and that researchers could not take any action that would affect the individual to whom the information pertains.

A main ethical issue confronting researchers engaged in participant observation research is deciding when and how to inform those being observed that they are part of a research study. In theory, a researcher should identify himself or herself as a researcher at the onset of participant observation. However, in reality this may not be feasible without inherently changing the interactions at the outset. If the researcher decides to do so, a general but forthright description of the aims of the research should be sufficient. As relationships with members deepen, any controversial aspects of the study should be revealed. A researcher must obtain informed consent from any member who agrees to a formal, in-depth interview.

Institutional Review Board (IRB) Review

In order to assure that research subject and participant rights and welfare are protected, all researchers should have their project reviewed by an IRB or comparable bodies. The National Institutes of Health supplies strict guidelines for project approval. Many of these guidelines are based on the Belmont Report.

Resources

See the following for additional information about research ethics and protecting the rights of study participants and their data:

- CASRO Code of Standards and Ethics for Survey Research
- AAPOR Code of Ethics
- National Center for Education Statistical Standards
- Tips on Informed Consent

Prepared by: Jerry West
Last updated: March 2019

Research Connections is a partnership between the National Center for Children in Poverty at the Mailman School of Public Health, Columbia University, and the Interuniversity Consortium for Political and Social Research at the Institute for Social Research, the University of Michigan, supported by a grant from the Office of Planning, Research and Evaluation in the Administration for Children and Families, U.S. Department of Health and Human Services. Contents are solely the responsibility of the authors and do not necessarily represent the official views of the Office of Planning, Research and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.