Chapter 2: Research Approaches in Education

Educational research is typically classified into two broad categories: quantitative and qualitative research. Each approach has its own methodology and terminology.

Quantitative research uses objective measurement to gather numeric data that are used to answer questions or test predetermined hypotheses. It generally requires a well-controlled setting. **Qualitative research,** in contrast, focuses on understanding social phenomena from the perspective of the human participants in natural settings. It does not begin with formal hypotheses, but it may result in hypotheses as the study unfolds.

Philosophical assumption behind these two approaches

Quantitative research originated in **positivism,** a philosophic view formulated in Europe in the 19th century. Positivists believe that general principles or laws govern the social world as they do the physical world and that through objective procedures researchers can discover these principles and apply them to understand human behavior. The positivists, such as Francis Bacon, stressed ***observation*** as the primary source of dependable knowledge. Positivism is often considered the traditional scientific method, which involves hypothesis testing and objective data gathering to arrive at findings that are systematic, generalizable, and open to replication by other investigators.

Qualitative research is based on a different philosophical approach, which sees the ***individual and his or her world as so interconnected*** that essentially the one has no existence without the other. It sees social reality as unique; thus, researchers can only understand human behavior by focusing on the meanings that events have for the people involved. You must look not only at **what people do** but also at **how they think and feel**, and you must attempt to **understand their reality**. The intended result of a qualitative research study is a narrative ***report so rich and comprehensive*** that you can understand the social reality experienced by the participants. Furthermore, because researchers do not know in advance how naturally occurring events will unfold or ***what variables may be important***, they do not begin a study with hypotheses.

Wikipedia Source

Qualitative research is a broad methodological approach that encompasses many research methods. The aim of qualitative research may vary with the disciplinary background, such as a psychologist seeking to gather an in-depth understanding of human behavior and the reasons that govern such behavior. Qualitative methods examine the *why* and *how* of decision making, not just *what*, *where*, *when*, or "who", and have a strong basis in the field of sociology to understand government and social programs.

A popular method of qualitative research is the case study which examines in depth "purposive samples" to better understand a phenomenon; hence, smaller but focused samples are more often used than large samples.

Qualitative researchers face many choices for techniques to generate data ranging from grounded theory[14] development and practice, narratology, storytelling, transcript poetry, classical ethnography, state or governmental studies, research and service demonstrations, focus groups, case studies, participant observation, qualitative review of statistics in order to predict future happenings, or shadowing, among many others. Qualitative methods are used in various methodological approaches, such as action research which has sociological basis, or actor-network theory.

Alternative Approach in Educational Research

In the late 20th century scholars began to call for an alternative to the quantitative approach in educational research (Guba & Lincoln, 1988). They believed that using quantitative methods in highly controlled settings ignored the participants' perspectives and experiences. **Qualitative research was the alternative**. For a time, the relationship between quantitative and qualitative researchers was somewhat adversarial, but gradually there was a trend toward rapprochement as researchers began to see quantitative and qualitative methodology as complementary.

A new methodology in which the same study uses both quantitative and qualitative approaches is called **mixed methods research.** The end result of mixed methods research is findings that may be more dependable and provide a more complete explanation of the research problem than either method alone could provide. It provides us with a more informative picture of the phenomenon.

Mixed method Approach Characteristics

It may be **more difficult to carry out** a mixed methods study because one must have knowledge and an understanding of both quantitative and qualitative methodology. A mixed method study also typically involves **more extensive data collection and analysis** and thus will require more

Quantitative research may be further classified as either experimental or nonexperimental.

Experimental research

Experimental research involves a study of the effect of the systematic manipulation of one variable(s) on another variable. The manipulated variable is called the **experimental treatment** or the **independent variable.** The observed and measured variable is called the **dependent variable.**

To have a "**true" experiment**, researchers must use a ***random process*** such as a coin toss to assign available subjects to the experimental treatments. Sometimes, however, researchers

cannot randomly assign subjects to experimental treatments for a study. Instead, the experimenter must use already assembled groups such as classes. In this case, the research is called **quasi-experimental.**

Nonexperimental Research

In **nonexperimental quantitative research,** the researcher identifies variables and may look for relationships among them but does not manipulate the variables. Major forms of nonexperimental research are ***relationship studies*** including (a) ex post facto, (b) correlational research and (**c**) survey research.

Given **ex-post facto** design, because there is no manipulation or control of the independent variable, one must be very careful regarding the conclusions that are drawn about any observed relationship.

Correlational research gathers data from individuals on two or more variables and then seeks to determine if the variables are related (correlated). *Correlation* means the extent to which the two variables vary **directly** (positive correlation) or **inversely** (negative correlation). The degree of relationship is expressed as a numeric index called the *coefficient of correlation.*

Both ***ex post facto*** and ***correlational research***es investigate relationships between variables. The major distinction between the two is that in ex post facto research one categorizes the participants into at least two groups on **one** variable and then compares them on the other variable. In correlational research, a researcher deals with one group of individuals measured on at least **two** continuous variables (A **continuous variable** is one which can take on infinitely many, uncountable values. A **continuous variable** is the opposite of a discrete **variable**, which can only take on a certain number of values).

Survey research (also called **descriptive research**) uses instruments such as questionnaires and interviews to gather information from groups of individuals. Surveys permit the researcher to summarize the characteristics of different groups or to measure their attitudes and opinions toward some issue. Researchers in education and the social sciences use surveys widely.

Qualitative Research

There are many different types of qualitative research; we consider briefly eight of the most widely used approaches.

Case Studies (it helps researchers to build theory**)**

A case study is a type of **ethnographic research** study that focuses on a single unit, such as one individual, one group, one organization, or one program. The goal is to arrive at a **detailed description** and understanding of the entity (the case). In addition, a case study can result in data

from which **generalizations to theory are possible**. Freud, for example, used the case study extensively in building his theory of personality. Case studies use ***multiple methods***, such as interviews, observation, and archives, to gather data.

Document or Content Analysis

Content analysis focuses on analyzing and interpreting recorded material to learn about **human behavior**. The material may be public records, textbooks, letters, films, tapes, diaries, themes, reports, or other documents. Content analysis usually **begins with a question** that the researcher believes can best be answered by studying documents.

Ethnography

Ethnography is an in-depth study of naturally occurring behavior within a culture or social group. Social scientists sometimes call ethnography ***field research*** because it is conducted in a natural setting or 'field". The ***researcher observes group behavior*** as it occurs naturally in the setting, without any simulation or imposed structure. Ethnography requires a variety of data-gathering procedures, such as **prolonged observation** of the setting, **interviewing** members of the culture, and **studying documents and artifacts**. Researchers interpret the data in the **context** of the situation in which they gathered the data.

Grounded Theory

Grounded theory research is designed to **develop a theory** of social phenomena based on the field data collected in a study. Experience with the data generates insights, hypotheses, and questions, which researchers pursue with further data collection. From an inductive analysis of the data, the researcher constructs concepts. He or she then forms a theory by proposing plausible relationships among the concepts. ***The theory is thus said to be grounded in the data***. For example, a researcher interested in mainstreaming in elementary school could observe a number of classrooms and conduct interviews with teachers and students. Analysis of the data could lead to a theory about mainstreaming in the elementary school.

Historical Research

Historical research analyzes documents and artifacts and or uses interviews with eyewitnesses to gain insight into past events. The success of historical research depends on the accuracy and completeness of the source material. The researcher must establish the authenticity of the documents used, as well as the validity of their contents.

An educational researcher might want to investigate the trends in kindergarten education in a particular school district from its beginnings to the present. Also, one might investigate the methods used to teach reading in the past or study school practices and policies such as grade retention.

Narrative Inquiry

In **narrative inquiry,** researchers examine the stories people tell about their lives and co-construct a narrative analysis of those stories. The researcher and those telling their stories have **an equal voice** in determining the meanings attributed to the experiences. Narrative analysis has also been referred to using terms such as life stories. A researcher investigating teacher reflection or teacher pathways into teaching might use narrative inquiry approaches.

Typical Stages in Research

- ***Selecting a Research Problem***; quantitative researcher construct question with "**what**" often trying to investigate or question the relationship between variables. Qualitative researchers answer the "**why**" and "**how**" questions of certain phenomena.
- **Reviewing the Literature on the Problem;** Researchers review the literature to gain more insight into problem and to determine what research may already have been done.
- **Designing the Research;** the investigator next plans how to conduct research to answer the question. The design is the researcher's plan for the study, which includes the method to be used, what data will be gathered, where, how, and from whom.

In qualitative research, the **design** is **flexible** and may change during the investigation if appropriate. The design of qualitative research is thus often described as "**emergent"**..

- Collecting Data;
- Analyzing Data;

The analysis of the **numerical data in quantitative** research provides evidence that supports or fails to support the hypothesis of the study. Qualitative data generally take the form of words (descriptions, observations, impressions, recordings, and the like). The researcher must organize and categorize or code the large mass of data so that they can be described and interpreted.

- Interpreting the Findings and Stating Conclusions

The quantitative researcher typically **makes statements about the probability** that such a finding is due to chance and reaches a conclusion about the hypothesis. Qualitative researchers present their interpretations and explanations in narrative form. They do not talk about probability but try to **emphasize the trustworthiness and credibility of the findings**.

- Reporting the Results

Questions That Educational Researchers Ask

Theoretical Questions (deal with developing new theories or testing existing theories**)**

Questions of a theoretical nature are those asking "What is it?" or "How does it occur?" or "Why does it occur?" Educational researchers formulate "what" questions more specifically as "What

is intelligence?" or "What is creativity?" Typical "how" questions are "How does the child learn?" or "How does personality develop?" "Why" questions might ask "Why does one forget?" or "Why are some children more achievement-oriented than other children?"

Research with a theoretical orientation may focus on either **developing new theories or testing existing theories**. The former involves a type of study in which researchers seek to discover generalizations about behavior, with the goal of clarifying the nature of relationships among variables. They may believe that certain variables are related and thus conduct research to describe the nature of the relationship. From the findings, they may begin to formulate a theory about the phenomenon.

Probably more common in ***quantitative educational research*** are studies that aim to test already ***existing theories***. .

Basic and Applied Research

Basic research is research aimed at obtaining empirical data used to formulate and expand theory. Basic research is not oriented in design or purpose toward the solution of practical problems. Its essential aim is to expand the frontiers of knowledge without regard to practical application.

Applied research aims to solve an immediate practical problem. It is a research performed in relation to actual problems and under the conditions in which they appear in practice.

Applied research may not provide the general knowledge to solve other problems. For example, an elementary school teacher may study the effect of a new method of teaching fractions. She or he conducts the research to answer a practical question, **not** necessarily to make broad **generalizations** or to help develop a theory.

Note:

This classification of research is not always distinct, however, because there are varying degrees on the basic‑applied continuum.

Basic research often has practical benefits in the long term. For example, progress in educational practice is related to progress in **discovering general laws through basic** psychological, educational, and sociological research.

Language of Research

Empirical research

Empirical research is research using empirical evidence. It is a way of gaining knowledge by means of direct and indirect observation or experience. Empiricism values such research more

than other kinds. Empirical evidence (the record of one's direct observations or experiences) can be analyzed quantitatively or qualitatively.

Scientists need terms at the empirical level to describe particular observations; they also need terms at the theoretical level for referring to hypothetical processes that may not be subject to direct observation.

What is the important role of Construct in research?

One of these terms is **construct**. To summarize their observations and to provide explanations of behavior, scientists create constructs. **Constructs** are abstractions that cannot be observed directly but are useful in interpreting empirical data and in theory building (**96 کنکور دکتری**).

Constructs in Educational Research

For example, people can observe that individuals differ in what they can learn and how quickly they can learn it. To account for this observation, scientists invented the construct called ***intelligence***. They hypothesized that intelligence influences learning and that individuals differ in the extent to which they possess this trait. Other examples of constructs in educational research are motivation, reading readiness, anxiety, underachievement, creativity, and self-concept.

Constructs may be defined in a way that gives their general meaning, or they may be defined in terms of the operations by which they will be measured or manipulated in a particular study. The former type of definition is called a *constitutive definition;* the latter is known as an *operational definition.*

Constitutive Definition

A **constitutive definition** is a formal definition in which a term is defined by using other terms. It is the dictionary type of definition. For example, intelligence may be defined as the ability to think abstractly or the capacity to acquire knowledge. This type of definition helps convey the **general meaning of a construct**, but it is not precise enough for research purposes.

What is the purpose of providing constitutive definition of a construct?

The researcher needs to define constructs so that readers know exactly what is meant by the term and so that other investigators can replicate the research. An operational definition serves this purpose.

Operational Definition

An **operational definition** ascribes meaning to a construct by **specifying operations** that researchers must perform to measure or manipulate the construct. Operational definitions may not be as rich as constitutive definitions but are essential in research because investigators must

collect data in terms of observable events. Scientists may deal on a theoretical level with such constructs as learning, motivation, anxiety, or achievement, but before studying them empirically, scientists must specify observable events to represent those constructs and the operations that will supply relevant data. Operational definitions help the researcher bridge the gap between the **theoretical** and the **observable**. Although investigators are guided by their own experience and knowledge and the reports of other investigators, the operational definition of a concept is to some extent **arbitrary**. Often, investigators choose from a variety of possible operational definitions those that best represent their own approach to the problem. ---,

Operational definitions are essential to research because they permit investigators to **measure abstract constructs** and permit scientists to move from the **level of constructs** and **theory** to the **level of observation,** on which science is based. It is important to remember that although researchers report their findings in terms of abstract constructs and relate these to other research and to theory, what they have actually found is a relationship between two sets of observable and measurable data that they selected to represent the constructs.

Variables

Researchers, especially quantitative researchers, find it useful to think in terms of variables. A **variable** is a construct or a characteristic that can take on different values or scores. Researchers study variables and the relationships that exist among variables.

Types of Variables

Variables can be **categorical**, or they can be **continuous**. When researchers classify subjects by sorting them into mutually exclusive groups, the attribute on which they base the classification is termed a categorical variable. Home language, county of residence, father's principal occupation, and school in which enrolled are examples of categorical variables. The simplest type of categorical variable has only two mutually exclusive classes and is called a **dichotomous variable.** Male‑female, citizen‑alien, and pass‑fail are dichotomous variables. Some categorical variables have more than two classes; examples are educational level, religious affiliation, and state of birth.

When an attribute has an **infinite number of values** within a range, it is a **continuous variable.** As a child grows from 40 to 41 inches, he or she passes through an infinite number of heights. Height, weight, age, and achievement test scores are examples of continuous variables. The most important classification of variables is on the basis of their *use* within the research under consideration, when they are classified as independent variables or dependent variables.

Constants

The opposite of variable is **constant.** A constant is a fixed value within a study. If all subjects in a study are eighth-graders, then grade level is a constant. In quantitative research, constructs are quantified and take on different values. Thus, they are referred to as *variables.*

Chapter 3: Research Problem

A research problem is not a nuisance; it is a step toward new knowledge.

Nuisance /ˈnjuːs(ə)ns/; a person or thing causing inconvenience or annoyance.

> "it's a nuisance having all those people clomping through the house"

Sources of Research Problems

Although there are **no set rules for locating a problem**, certain suggestions can help. Three important sources for research problems are **experience**, **deductions from theory**, and **related literature**.

A *theory* may be defined as a set of interrelated statements, principles, and propositions that specify the relationships among variables. The application of the general principles embodied in a theory to specific educational problems is only **hypothetical**, however, until research empirically confirms them.

The characteristics of a good theory

1. *An essential characteristic of a good theory is that it is **testable***.
2. A good theory is not only testable but also **falsifiable**. Being falsifiable means that it is **capable of being proven wrong**. It is possible to gather evidence that contradicts the theory.
3.

Problem Statement in Quantitative Research

The **problem statement** in quantitative research specifies the variables and the population of interest. The problem statement can be a **declarative one** such as "This study investigates the effect of computer simulations on the science achievement of middle school students." The statement can **ask a question** about a relationship between the two (or more) variables. The problems could be further clarified by **operationally defined** the variables involved.

Why is the question form preferred largely?

Because it is straightforward and psychologically seems to orient the researcher to the task at hand−namely, to find the answer to the question

The Problem Statement in Qualitative Research

Qualitative researchers state it much **more broadly** than in quantitative research. Formulation of a qualitative problem begins with the **identification of a general topic** or an area you want to know more about. This general topic of interest is sometimes referred to by qualitative

researchers as the **focus of inquiry**. This initial broad focus provides the **framework** but allows for changes as the study proceeds. As the researcher gathers data and discovers new meanings, the general problem narrows to more specific topics and new questions may arise.

In quantitative research, the problem is specified in the beginning, but in qualitative it is stated broadly.

In qualitative research, the statement may be somewhat general in the beginning, but it will become more focused as the study proceeds. After exploring the **sites**, the **people**, and the **situations**, the researcher narrows the options and states the research problem more specifically.

Although the qualitative researcher intuitively **arrives at hunches** about the phenomenon, he or she does not formulate an initial hypothesis that the study tests.

Second Step: Evaluating the problem

Does the question warrant an expenditure of time and effort to investigate?

Criteria for research problem evaluation:

1. The problem should have significance−that is, it should be one whose solution will make a contribution to educational theory or practice. The problem may fill in gaps in current knowledge or help resolve some of the inconsistencies in previous research.
2. The problem should be one that will lead to new problems and so to further research. A good study, while arriving at an answer to one question, usually generates a number of other questions that need investigation.
3. The problem must be researchable. A researchable problem is one that can be attacked empirically; that is, it is possible to gather data that answer the question.
4. The problem should be suitable for the researcher. The problem should be one in which the researcher has a genuine interest and about which you can be enthusiastic. It should be a problem whose solution is personally important.
5. The problem should be ethically appropriate. That is, the problem should be one that you can investigate without violating ethical principles. Three issues are important
    a. *Consent*. Researchers need to obtain consent from the intended subjects.
    b. *Protection from harm*. Do not plan research that may cause physical harm or psychological harm such as stress, discomfort, or embarrassment that could have lasting adverse /ˈadvəːs/ effects.
    c. *Privacy*. A researcher should invade the privacy of subjects as minimally as possible.

Points

- Research cannot answer questions of “should.”

Stating the Research Problem

After you have selected and evaluated the problem, the next task is to state the problem in a form amenable to investigation.

Problem Selection ⟶ Problem Evaluation ⟶ Problem Statement

The Problem Statement in Quantitative Research

- We have to specify the variables and population of interest
- The problem statement could be declarative such as, "This study investigates the effect of computer simulations on the science achievement of middle school students."
- It could be stated in the form of question(s)

Why do some researchers prefer to state the problem in the form of Questions?

Because it is straightforward and psychologically seems to orient the researcher to the task at hand−namely, to find the answer to the question.

The problem can be further clarified by operationally defining the variables involved.

What does it mean by operationally defining the variables??

The Problem Statement in Qualitative Research

Qualitative researchers also begin with a problem, but they state it much more **broadly** than in quantitative research. A qualitative problem statement or question indicates the *general* purpose of the study. Formulation of a qualitative problem begins with the identification of a general topic or an area you want to know more about. This general topic of interest is sometimes referred to by qualitative researchers as the **focus of inquiry**. This initial broad focus provides the framework but allows for changes as the study proceeds. As the researcher gathers data and discovers new meanings, the general problem narrows to more specific topics and new questions may arise.

More Broad Topic/Problem ⟶ More Specific Topic/Problem

Whereas the quantitative researcher always states the problem before collecting data, the qualitative researcher may formulate problems after beginning to collect data. In fact, the researcher often does not present the final statement of the problem−which typically specifies the setting, subjects, context, and aim of the study−until he or she has collected at least some data. In qualitative research, the statement may be somewhat general in the beginning, but it will become more focused as the study proceeds. After exploring the sites, the people, and the situations, the researcher narrows the options and states the research problem more specifically.

Chapter 4: Review of the Literature

The search for related literature plays a vital but quite different role in qualitative and quantitative research. **It must be completed early in quantitative research** but not in qualitative research.

The Role of Related Literature in Quantitative Research

Quantitative researchers are urged **not** to rush headlong into conducting their study. The search for related literature should be completed **before** the actual conduct of the study begins in order to **provide a context** and **background** that support the conduct of the study.

This literature review stage serves several important **functions** in quantitative research:

1. *Knowledge of related research enables investigators to **define the frontiers** of their field.*
2. *It enables researchers to place their questions in perspective.*
3. *It helps researchers **to limit their research question** and to clarify and define the concepts of the study.*

Successful reviews often result in the **formation of hypotheses** regarding the relationships among variables in a study. The hypotheses can provide direction and focus for the study.

4. *Through studying related research, investigators learn which **methodologies** have proven useful and which seem less promising.*
5. It *avoids **unintentional replication** of previous studies.*
6. *It places researchers in a better position to **interpret the significance** of their own results.*

As this discussion shows, quantitative research is built on a study of earlier work in the field, which helps the researcher refine his or her problem and place it in context. For qualitative researchers, the approach is very different. They are advised not to read in their area of interest because it is important that they approach their study **without any preconceived idea**s that might influence their work.

The Role of Related Literature in Qualitative and Mixed Methods Research

Barney G. Glaser, a pioneer in the grounded theory school within qualitative research, wrote (1978), "In our approach we **collect the data** first. Then start **analyzing it** and **generating theory**. When the theory seems sufficiently grounded and developed, then we review the literature in the field and relate the theory to it through integration of ideas" (p. 31).

When the grounded theory **study is complete**, the researcher formulates theories to explain what has been observed. Then, the researcher searches the literature to determine how his or her conclusions fit into the existing theories in the field.

Other fields of qualitative research may include **a brief review of related literature** at the beginning of a study to identify the theory that inspired the research. In the case of mixed methods research, the **literature review may take a more dynamic and flexible form**. It may be **exploratory** in the beginning stages of the study and **explanatory** at the end of the study. Or, it may take on both characteristics in iterative fashion as new research questions arise.

Efficient Location of Related Literature

Currently, most universities and colleges and many public and private libraries subscribe to indexing and abstracting periodicals that are incorporated into several **databases** that can be searched by computer. Computers can search for many topics simultaneously and combine them, using logical concepts known as **Boolean logic** (from the logic system developed by the 19th-century English mathematician George Boole).

Indexing and Abstracting Databases

Indexing and abstracting periodicals are vital for locating primary sources in your field. These publications subscribe to professional journals in a given discipline. Their staff then identifies the key terms for each article, indexes them, and typically provides an abstract for each article. Databases that combine several of these indexing and abstracting periodicals are very useful because you can ask for your key terms of interest and the database will identify the journal articles by journal, date, volume number, and pages that include your key terms.

Major Useful Databases for Educational Research

ERIC www.eric.ed.gov

Google Scholar http://scholar.google.com

WorldCat www.worldcat.org

JSTOR www.jstor.org

Internet

It is often more difficult to determine the worth of a website than that of a print source because many personal sites look as professional and authoritative as a governmental or educational site. One place to start is to consider the end of the address. Sites ending in *.edu* or *.gov* are education or government sites, which tend to have more credibility than sites ending in .com, .org, or .net. Many libraries and organizations provide lists of subject-specific websites for researchers.

Chapter 5: Hypothesis in Quantitative Research

After stating the research question and examining the literature, the quantitative researcher is ready to state a **hypothesis** based on the question. This should be done before beginning the research project. Recall that the **quantitative problem** asks about the relationship between two (or more) variables.

Although hypotheses serve several important purposes, some research studies may proceed without them. **Hypotheses are tools in the research process**, not ends in themselves. Studies are often undertaken in areas in which there is little accumulated background information.

Reasons for Making Hypotheses

- a well-grounded hypothesis indicates that the researcher has sufficient knowledge in the area to undertake the investigation;
- The hypothesis gives direction to the collection and interpretation of the data; it tells the researcher what procedure to follow and what type of data to gather and thus may prevent a great deal of wasted time and effort on the part of the researcher.

دانش کافی محقق از زمینه ی تحقیقاتی خود را دارد. در واقع یک فرضیه ی محکم راه و یک فرضیه ی خوب نشان از روش پژوهش را به محقق نشان می دهد.

Very simply, the hypothesis tells the researcher what to do. Facts must be selected and observations made because they have relevance to a particular question, and the hypothesis determines the relevance of these facts. The hypothesis provides a basis for **selecting** the **sampling**, **measurement**, and **research procedures to use**, as well as the **appropriate statistical analysis**. Furthermore, the hypothesis helps keep the study restricted in **scope**, preventing it from becoming too broad or unwieldy.

Deriving Hypotheses Inductively

In the inductive procedure, the researcher formulates an **inductive hypothesis** as a generalization from apparent observed relationships; that is, the researcher observes behavior, notices trends or probable relationships, and then hypothesizes an explanation for this observed behavior.

How about the reviewing literature?

This reasoning process **should be accompanied by an examination of previous research** to determine what findings other investigators have reported on the question.

The inductive procedure is a particularly **fruitful source of hypotheses** for classroom teachers. Why?

This sort of hypothesis derives inductively from teacher's observation in the classroom.

Summary

In the inductive process, the researcher makes observations, thinks about the problem, turns to the literature for clues, makes additional observations, and then formulates a hypothesis that seeks to account for the observed behavior. The researcher (or teacher) then tests the hypothesis **under controlled conditions** to examine scientifically the assumption concerning the relationship between the **specified variables**.

Deriving Hypothesis Deductively

In contrast to hypotheses formulated as generalizations from observed relationships, some others are derived by deduction from **theory.** These hypotheses have the advantage of **leading to a more general system of knowledge** because the framework for incorporating them meaningfully into the body of knowledge already exists within the theory. A science cannot develop efficiently if each study results in an isolated bit of knowledge. It becomes cumulative by building on the existing body of facts and theories. A hypothesis derived from a theory is known as a **deductive hypothesis.**

The Criteria for Hypothesis Evaluation

- A hypothesis states the **expected relationship** between variables
- A hypothesis must be testable

The most important characteristic of a "good" hypothesis is **testability**. A **testable hypothesis** is verifiable; that is, deductions, conclusions, or inferences can be drawn from the hypothesis in such a way that empirical observations either support or do not support the hypothesis.

The indicators of the variables are referred to as **operational definitions**.

Make sure the variables can be given operational definitions. **Avoid the use of constructs** for which it would be difficult or impossible to **find adequate measures**. Constructs such as *creativity*, *authoritarianism*, and *democracy* have acquired such diverse meanings that reaching agreement on operational definitions of such concepts would be difficult, if not impossible. Remember that the variables must be defined in terms of identifiable and observable behavior.

- A hypothesis should be consistent with the existing body of knowledge
- A hypothesis should be stated as simply and concisely as possible

If a researcher is exploring more than one relationship, he or she will need to state more than one hypothesis. The general rule is to state only one relationship in any one hypothesis.

TYPES OF HYPOTHESES

There are three categories of hypotheses: research, null, and alternate.

Research Hypothesis

The hypotheses we have discussed **thus far** are called **research hypotheses.** They are the hypotheses developed from observation, the related literature, and/or the theory described in the study. A research hypothesis states the relationship one expects to find as a result of the research. Research hypotheses may be stated in a **directional** or **non-directional** form. A directional hypothesis states **the direction of the predicted relationship or difference** between the variables.

Example: "There is a **positive** relationship between IQ and anxiety in elementary schoolchildren"

A **non-directional hypothesis**, in contrast, states that a relationship or difference exists but **without specifying the direction** or nature of the expected finding−for example, "There is a relationship between IQ and anxiety in children." The literature review generally provides the basis for stating a research hypothesis as directional or nondirectional.

The Null Hypothesis

It is **impossible** to test research hypotheses directly. آزمودن مستقیم فرضیه کاملا غیرممکن است.

You must first state a **null hypothesis** (symbolized *H*0) and assess the probability that this null hypothesis is true. The null hypothesis is a **statistical hypothesis**. It is called the null hypothesis because it states that there is no relationship between the variables in the population.

What is the point of the null hypothesis?

A null hypothesis lets researchers assess whether apparent relationships are genuine or are likely to be a function of chance alone. It states, "The **results** of this study could easily have happened **by chance**". Statistical tests are used to determine the probability that the null hypothesis is true. If the tests indicate that observed relationships had **only a slight probability of occurring** by **chance**, the null hypothesis becomes an unlikely explanation and the researcher rejects it.

Researchers aim to reject the null hypothesis as they try to show there *is* a relationship between the variables of the study. Testing a null hypothesis is analogous to the prosecutor's work in a criminal trial. To establish guilt, the prosecutor (in the U.S. legal system) must provide sufficient evidence to enable a jury to reject the presumption of innocence beyond reasonable doubt.

The Alternative Hypothesis

Note that the hypothesis "Children taught by individual instruction will exhibit less mastery of mathematical concepts than those taught by group instruction" **posits a relationship between variables** and therefore is *not* a null hypothesis. It is an example of an **alternative hypothesis.**

In the example, if the **sample mean** of the measure of mastery of mathematical concepts is higher for the **individual instruction** students than for the **group instruction** students, and inferential statistics indicate that the null hypothesis is unlikely to be true, you reject the null hypothesis and tentatively conclude that individual instruction results in greater mastery of mathematical concepts than does group instruction. If, in contrast, the **mean for the group instruction** students is higher than the mean for the individual instruction students, and inferential statistics indicate that this difference is not likely to be a function of chance, then you tentatively conclude that group instruction is superior.

If inferential statistics indicate that **observed differences** between the means of the two instructional groups **could easily be a function of chance**, the null hypothesis is retained, and you decide that insufficient evidence exists for concluding there is a relationship between the dependent and independent variables. The retention of a null hypothesis is ***not*** positive evidence that the null hypothesis is true. It indicates that the **evidence is insufficient** and that the null hypothesis, the research hypothesis, and the alternative hypothesis are all possible.

Testing the Hypothesis

If the null hypothesis is rejected, then the researcher often describes the results as being significant. In describing the importance of the results of the research study, however, there are two types of significance involved - **statistical significance** and **practical/educational significance**. **Rejecting a null hypothesis results in statistical significance**, but not necessarily practical significance

A statistically significant result is one that is likely to be due to a systematic (i.e., identifiable) difference or relationship, not one that is likely to occur due to chance. No matter how carefully designed the research project is, there is always the possibility that the result is due to something other than the hypothesized factor. The need to control all possible alternative explanations of the observed phenomenon cannot be emphasized enough. **Alternative explanations** can stem from an **unrepresentative sample**, some other type of **validity threat**, or an unknown, confounding factor. The ideal situation is one in which all other possible explanations are ruled out so that the only viable explanation is the research hypothesis.

The level that demarks statistical significance (called **alpha** and designated with the Greek letter, α) is completely under the control of the researcher. Norms for different fields exist. For example, α=.05 is generally used in educational research.

But, what does α=.05 actually mean?

The level of statistical significance is the level of risk that the researcher is willing to accept that the decision to reject the null hypothesis may be wrong by mis-attributing a difference to the hypothesized factor, when no difference actually exists. In other words, the level of statistical significance is the level of risk associated with rejecting a true null hypothesis. Selecting α=.05

indicates that the researcher is willing to risk being wrong in the decision to reject the null hypothesis 5 times out of 100, or 1 time out of 20. Referring back to the normal curve, α=.05 divides the area under the curve into two sections - one section where the null hypothesis is retained and another section where the null hypothesis is rejected. **Rejecting a true null hypothesis** is called committing a **Type I error.**

Another type of error that can be made is **retaining a false null hypothesis**. This is called **Type II error**. Like the chance of committing a Type I error, the chance of committing a Type II error is also under the control of the researcher. Unlike the Type I error level, which is set directly by the researcher, the Type II error level is determined by a combination of parameters, including the α level, sample size, and anticipated size of the results.

Decision	Null is true (not guilty)	**Null is false (guilty)**
Reject the null hypothesis (convict)	Type I error (convict the innocent) *level of statistical significance, α*	Correct decision (convict the guilty) power of the test, 1 - β
Retain the null hypothesis (acquit)	Correct decision (acquit the innocent)	Type II error (acquit the guilty) *chance of Type II error, β*

فرضیه حدس بخردانه ای درباره رابطه دو یا چند متغیر است که به صورت جمله ای خبری بیان شده و نشانگر نتایج مورد انتظار است (مقیمی ، 1383، 22). دربیانی دیگر، فرضیه حدسی است زیرکانه و علمی که باید به کمک واقعیات (داده ها) مورد بررسی قرار گرفته و سپس تایید یا رد گردد. :

فرضیه ها به دو نوع تقسیم میشود: فرضیه تحقیق (H1) و فرضیه صفر (H0)

فرضیه تحقیق از احتمال وجود رابطه یا اثر و یا تفاوت بین متغیرها خبر میدهد این فرضیه ها به دو نوع جهت دار و بدون جهت تقسیم می شود .

فرضیه صفر که به فرضیه آماری یا پوچ نیز موسوم است وجود رابطه ،اثر یا تفاوت بین متغیرها را رد کرده و انکار میکند .

مثال :

الف : فرضیه تحقیق(H1) دو نوع می باشد 1- فرضیه جهت دار و 2- فرضیه تحقیق بدون جهت

1- فرضیه تحقیق جهت دار : به نظر می رسد کارایی معلمان آموزش دیده بیشتر از معلمان آموزش ندیده است.

2- فرضیه تحقیق بدون جهت: به نظر می رسد بین آموزش معلمان و کارایی آنها رابطه وجود دارد.

ب : فرضیه صفر(H0) در همین مثال: به نظر می رسد بین کارآیی معلمان آموزش دیده و آموزش ندیده رابطه ای وجود ندارد.

اگرچه محقق فرضیه تحقیق را مطرح می کند و درصدد آزمایش آن است، پس از گرداوری اطلاعات و داده ها و طبقه بندی آن ها عملا فرضیه صفر را مورد آزمایش قرار می دهد. زیرا روش های آماری تجزیه و تحلیل داده-ها قادر به آزمون فرضیه صفر می‌باشند. درصورتی که فرض صفر رد شود آنگاه فرضیه تحقیق مورد تایید است . اما وقتی فرض صفر رد نشود گواه بر این است که داده های نمونه گواه کافی برای رد فرض صفر بدست نمی دهند و آزمون بی نتیجه می ماند . پیشنهاد می شود در صورت تاکید بر نتیجه گیری باید نمونه بیشتری مورد بررسی قرار بگیرد.

خطای نوع اول (مثبت کاذب) .... فرض صفر درست باشد و آزمون فرض آن را رد کند. **خطای نوع دوم (منفی کاذب)** .... فرض صفر درست نباشد و آزمون فرض آن را قبول کند.

Testing the Hypothesis

A quantitative study begins with a research hypothesis, which should be a simple, clear statement of the expected relationship between the variables. When researchers speak of testing a hypothesis, however, they are referring to the null hypothesis. Only the null hypothesis can be directly tested by statistical procedures.

Hypothesis testing involves the following steps:

1. State, in operational terms, the relationships that should be observed if then research hypothesis is true.

2. State the null hypothesis.

3. Select a **research method** that will enable the hypothesized relationship to be observed if it is there.

4. **Gather the empirical data** and select and calculate appropriate descriptive statistics for these data.

5. Calculate **inferential statistics** to determine the probability that your obtained results could have occurred by chance when the null hypothesis is true.

6. If the probability of the observed findings being due to chance is very small, one would have sufficient evidence to reject the null hypothesis.

The Main Point with Regard to Hypothesis

Although you may find support for a hypothesis, the hypothesis is not *proved* to be true. A hypothesis is never proved or disproved; it is only supported or not supported. Hypotheses are essentially probabilistic in nature; empirical evidence can lead you to conclude that the explanation is probably true or that it is reasonable to accept the hypothesis, but it never proves the hypothesis.

فرض تحقیق نه می تواند درست باشد، نه می تواند نادرست. آنچه که از اهمیت بالایی برخوردار است این است که فرض تحقیق ممکن است با استفاده از شواهد و داده ها مورد تایید قرار بگیرد یا اینکه قرار نگیرد.

Inferential statistics

Inferential statistics only address random error (**chance**). The reason for calculating an **inferential statistic** is to get a p value (p = probability). The p value is the probability that the samples are from the same population with regard to the dependent variable (outcome)

Research Plan

A research plan should include the following elements: (a) the problem, (b) the hypothesis, (c) the research methodology, and (d) proposed data analysis.

فرق بین نظریه و فرضیه در چیست؟

نظریه و قوانین عمدتا مشتمل بر قضایای کلی و عمومی هستند و به مورد خاصی تعلق ندارند و می توانند مصادیق زیادی داشته باشند . در حالی که فرضیه حالت کلی ندارد و مختص مساله تحقیق است که از قضایای کلی ناشی می شود ولی در یک قلمرو خاص شکل میگیرد.

Chapter 6: Descriptive Statistics

A fundamental step in the conduct of quantitative research is measurement - the process through which observations are translated into numbers.

Level of measurement

Psychologist Stanley Smith Stevens developed the best known classification with four levels, or scales, of measurement: nominal, ordinal, interval, and ratio. These scales are used for quantifying the observations.

Nominal Scale

The nominal type differentiates between items or subjects based only on their names or (meta-) categories and other qualitative classifications they belong to; thus **dichotomous data** involves the construction of classifications as well as the classification of items. Discovery of an exception to a classification can be viewed as progress. Numbers may be used to **represent the variables** but the numbers do **not have numerical value** or relationship.

Nominal measurement involves placing objects or individuals into mutually exclusive categories. Numbers are arbitrarily assigned to the categories for identification purposes. One can only count the number of observations in each category or **express the numbers in categories as a percentage** of the total number of observations.

Examples of these classifications include gender, nationality, ethnicity, language, genre, style, biological species, and form. In grammar, the parts of speech: noun, verb, preposition, article, pronoun, etc.

The nominal level is the lowest measurement level used from a statistical point of view.

The mode, i.e. the *most common* item, is allowed as the measure of central tendency for the nominal type. On the other hand, the median, i.e. the *middle-ranked* item, makes no sense for the nominal type of data since ranking is meaningless for the nominal type

Ordinal

The ordinal type allows for rank order (1st, 2nd, 3rd, etc.) by which data can be sorted, but still does not allow for relative *degree of difference* between them. Examples include, on one hand, **dichotomous** data with dichotomous values such as 'wrong/false' vs. 'right/true' when measuring truth value, and, on the other hand, **non-dichotomous** data consisting of a ***spectrum of values***, such as 'completely agree', 'mostly agree', 'mostly disagree', 'completely disagree' when measuring opinion.

The **lack of equal intervals** in ordinal scales limits the statistical procedures available for analyzing ordinal data.

Central tendency

The median, i.e. *middle-ranked*, item is allowed as the measure of central tendency; however, the mean (or average) as the measure of central tendency is not allowed. The mode is allowed.
Psychological measurement, such as measurement of opinions, usually operates on ordinal scales; thus means and standard deviations have no validity, but they can be used to get ideas for how to improve operationalization of variables used in questionnaires. Most psychological data collected by psychometric instruments and tests, measuring cognitive and other abilities, are ordinal, although some theoreticians have argued they can be treated as interval or ratio scales.

Interval

The interval type allows for the *degree of difference* between items, but not the ratio between them. Examples include *temperature* with the Celsius scale, which has two defined points (the freezing and boiling point of water at specific conditions) and then separated into 100 intervals.

Central tendency and statistical dispersion

The mode, median, and arithmetic mean are allowed to measure central tendency of interval variables, while measures of statistical dispersion include range and standard deviation. Since one can only divide by *differences*, one cannot define measures that require some ratios, such as the coefficient of variation. More subtly, while one can define moments about the origin, only central moments are meaningful, since the choice of origin is arbitrary. One can define standardized moments, since ratios of differences are meaningful, but one cannot define the coefficient of variation, since the mean is a moment about the origin, unlike the standard deviation, which is (the square root of) a central moment.

Numbers on an interval scale may be manipulated by addition and subtraction, but because the zero is arbitrary, multiplication and division of the numbers are not appropriate. Thus, ratios between the numbers on an interval scale are meaningless.

Ratio

in contrast to interval scales, ratios are now meaningful because having a non-arbitrary zero point makes it meaningful to say, for example, that one object has "twice the length" of another (= is "twice as long"). Very informally, many ratio scales can be described as specifying "how much" of something (i.e. an amount or magnitude) or "how many" (a count).
The Kelvin temperature scale is a ratio scale because it has a unique, non-arbitrary zero point called absolute zero.

Central tendency and statistical dispersion

The geometric mean and the harmonic mean are allowed to measure the central tendency, in addition to the mode, median, and arithmetic mean. The studentized range and the coefficient of variation are allowed to measure statistical dispersion. All statistical measures are allowed because all necessary mathematical operations are defined for the ratio scale

Organizing Research Data

Researchers typically collect a large amount of data. Before applying statistical procedures, the researcher must organize the data into a manageable form. The most familiar ways of organizing data are (1) arranging the measures into frequency distributions and (2) presenting them in graphic form. The first step in preparing a frequency distribution is to list the scores in a column from highest at top to lowest at bottom.

It is often helpful and convenient to present research data in graphic form. Among various types of graphs, the most widely used are the **histogram** and the **frequency polygon.**

Central Tendency

In statistics, a **central tendency** (or **measure of central tendency**) is a central or typical value for a probability distribution. It may also be called a **center** or **location** of the distribution. Colloquially, measures of central tendency are often called *averages.*

The central tendency of a distribution is typically **contrasted with** its *dispersion* or *variability*; dispersion and central tendency are the often characterized properties of distributions. Analysts may judge whether data has a strong or a weak central tendency based on its dispersion.

A convenient way of summarizing data is to find a single index that can represent a whole set of measures. Finding a single score that can give **an indication of the performance of a group of 300 individuals on an aptitude test** would be useful for comparative purposes. In statistics, three indexes are the arithmetic mean, the median and the mode. They are called **measures of central tendency,** or averages.

The most widely used measure of central tendency is the **mean,** or arithmetic average. It is the sum of all the scores in a distribution divided by the number of cases. The **median** is defined as that point in a distribution of measures below which 50 percent of the cases lie (which means that the other 50 percent will lie above this point). The **mode** is the value in a distribution that occurs most frequently. It is the simplest to find of the three measures of central tendency because it is determined by **inspection** rather than by **computation**.

Comparison of the Three Indexes of Central Tendency

Because the mean is an interval or ratio statistic, it is generally a more precise measure than the median (an ***ordinal statistic***) or the mode (a ***nominal statistic***). It takes into account the value of *every* score. It is also the most stable of the three measures of central tendency in that if a number

of samples are randomly drawn from a parent population, the means of these samples will vary less from one another than will their medians and their modes. For these reasons, the mean is more frequently used in research than the other two measures. The mean is the ***best indicator of the combined performance*** of an entire group. However, the median is the best indicator of *typical* performance.

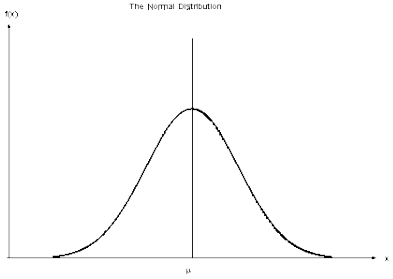
Shapes of Distributions

When graphed, the data in a set is arranged to show how the points are distributed throughout the set. These distributions show the spread (**dispersion**, **variability**, and **scatter**) of the data. The spread may be **stretched** (covering a wider range) or **squeezed** (covering a narrower range).

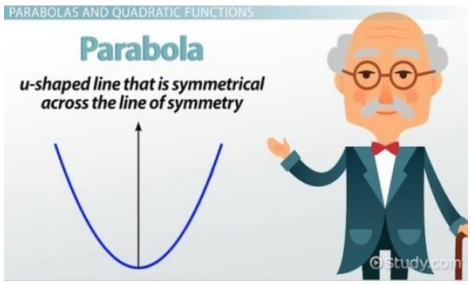
The shape of a distribution is described by **its number of peaks** and by its possession of symmetry, its tendency to skew, or its uniformity. (Distributions that are skewed have more points plotted on one side of the graph than on the other.)

1. One clear peak is called a *unimodal* distribution.
2. Two clear peaks are called a *bimodal* distribution.
3. Single peak at the center is called *bell shaped* distribution.

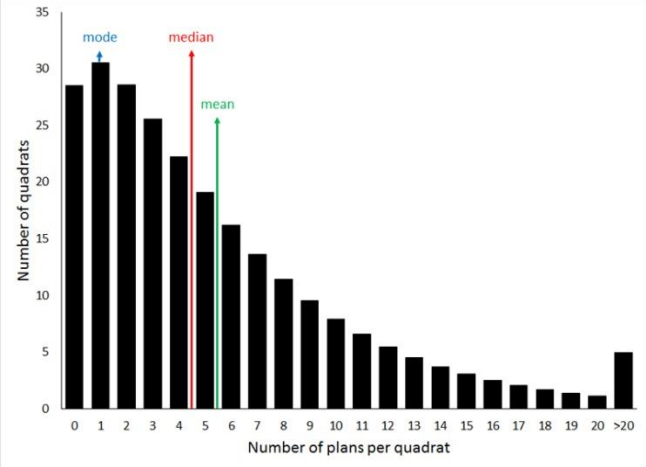
Symmetric (**bell shaped**) - when graphed, a vertical line drawn at the center will form ***mirror images***, with the left half of the graph being the mirror image of the right half of the graph. In the histogram and dot plot, this shape is referred to as being a "bell shape" or a "mound". This shape is often referred to as being a "normal curve" (or normal distribution).



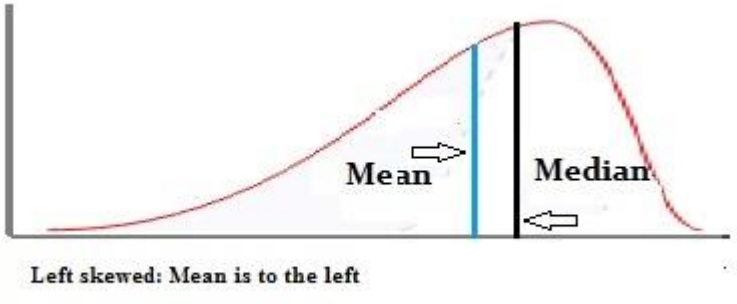
Symmetric (U-shaped) - as mentioned above, a symmetric graph forms a mirror image of itself when reflected in its vertical center line. Unlike the previous graphs, these histograms and dot plots have more of a U shape.



Skewed /skju:/**Right (positively skewed) -** fewer data plots are found to the right of the graph (toward the larger numeric values). The "tail" of the graph is pulled toward higher positive numbers, or to the right. The mean typically gets pulled toward the tail, and is greater than the median.



Skewed Left (negatively skewed) - fewer data plots are found to the left of the graph (toward the smaller numeric values). The "tail" of the graph is pulled toward the lower or negative numbers, or to the left. The mean typically gets pulled toward the tail, and is less than the median.



Uniform - The data is spread equally across the range. There are no clear peaks in these graphs, since each data entry appears the same number of times in the set.

Measures of Variability

Although indexes of central tendency help researchers describe data in terms of average value or typical measure, they ***do not give the total picture of a distribution***. The mean values of two distributions may be identical, whereas the degree of dispersion, or **variability,** of their scores might be different. In one distribution, the scores might cluster around the central value; in the other, they might be scattered.

In statistics, **dispersion** (also called **variability**, **scatter**, or **spread**) is the extent to which a distribution is stretched or squeezed.[1] Common examples of measures of statistical dispersion are the variance, standard deviation, and interquartile range.

Dispersion is contrasted with location or central tendency, and together they are the most used properties of distributions

Range

The simplest of all indexes of variability is the **range.** It is the difference between the upper real limit of the highest score and the lower real limit of the lowest score. The range is an ***unreliable index*** of variability because it is based on only two values, the highest and the lowest. It is not a stable indicator of the spread of the scores. For this reason, the use of the range is mainly limited to ***inspectional purposes***.

Variance and Standard Deviation

Variance and standard deviation are the most frequently used indexes of variability. They are both based on **deviation scores**—scores that show the difference between a raw score and the mean of the distribution.

Standard Deviation

In statistics, the **standard deviation** (**SD**) is a measure that is used to quantify the amount of variation or dispersion of a set of data values. A low standard deviation indicates that the data points tend to be close to the mean of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

The standard deviation of a random variable, statistical population, data set, or probability distribution is the square root of its variance.

Variance

In probability theory and statistics, **variance** is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of (random) numbers are spread out from their average value. Scores below the mean will have negative deviation scores, and scores above the mean will have positive deviation scores.

In most cases, educators prefer an index that summarizes the data in the same unit of measurement as the original data. **Standard deviation** (σ), the positive square root of variance, provides such an index. By definition, the standard deviation is the square root of the mean of the squared deviation scores. The standard deviation belongs to the same statistical family as the mean; that is, **like the mean, it is an interval or ratio statistic**, and its computation is based on the size of individual scores in the distribution. It is by far the most frequently used measure of variability and is used in conjunction with the mean.

Measures of Relative Position

Measures of relative position indicate where a score falls in relation to all other scores in the distribution. Statisticians often talk about the **position** of a value, relative to other values in a set of data. The most common measures of position are percentiles, quartiles, and standard scores (aka, z-scores).

Z Score

The *Z* score is defined as the distance of a score from the mean as measured by standard deviation units. A score exactly one **standard deviation above the mean** becomes a *z* of +1, a score exactly **one standard deviation below the mean** becomes a *z* of −1, and so on. A score equal to the mean will have a *z* score value of 0.

T Score

Scores can also be transformed into other standard score scales that do not involve negative numbers or decimals /ˈdɛsɪm(ə)lz/. One of the most common procedures is to convert to ***T* scores** by multiplying the *z* scores by 10 and adding 50. This results in a scale of positive whole numbers that has a mean of 50 and a standard deviation of 10.

Transforming a set of scores to standard scores does not alter the shape of the original distribution. If a distribution of scores is skewed, the derived standard scores also produce a skewed distribution. Only if the original distribution is normal do the standard scores produce a **normal distribution.**

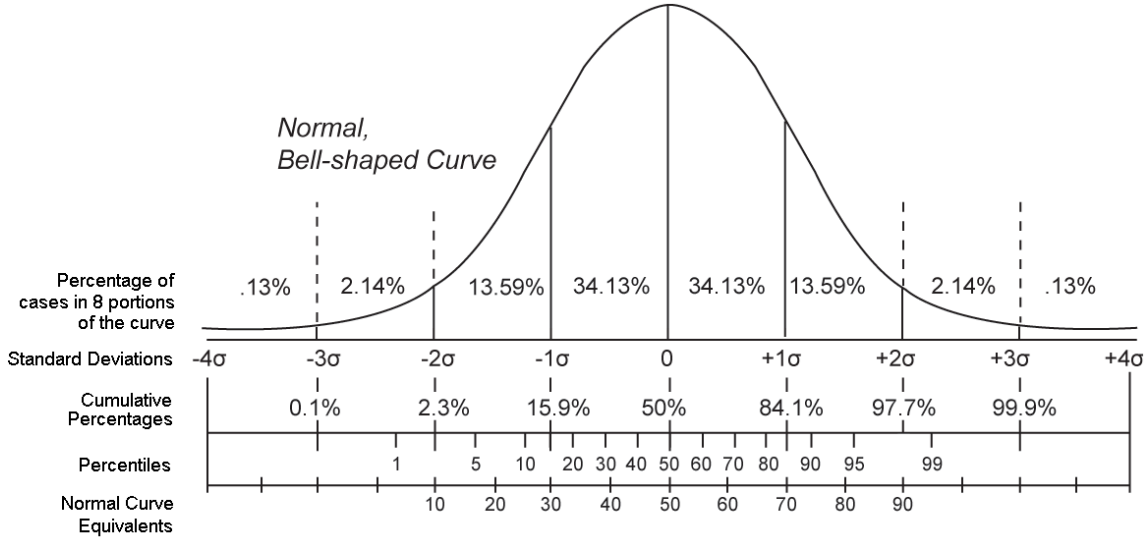
Percentile rank

A **percentile rank** (PR) indicates the percentage of scores in a distribution that fall below a given score point. It is easy to picture a score with a PR of 32 as having 32 percent of the scores in its distribution below it and a score with a PR of 89 as having 89 percent of the scores below it. For example, a test score that is greater than 75% of the scores of people taking the test is said to be at the 75th percentile, where 75 is the percentile rank. Percentile ranks are commonly used to clarify the **interpretation of scores** on standardized tests.

Normal Distribution

In a normal distribution, most of the cases concentrate **near** the **mean**. The frequency of cases decreases as you **proceed away from** the mean in either direction. Approximately 34 percent of the cases in a normal distribution fall between the mean and one standard deviation above the mean, and approximately 34 percent are between the mean and one standard deviation below the mean. Between one and two standard deviations from the mean on either side of the distribution are approximately 14 percent of the cases. Only approximately 2 percent of the cases fall

between two and three standard deviations from the mean, and only approximately one-tenth of 1 percent of the cases fall above or below three standard deviations from the mean.



Following the *z* score line are various standard scores transformed from *z* scores, including *T* scores, CEEB scores, stanines, percent in stanine, Wechsler subtest scores, and Weschsler deviation IQs. Note that 95 percent of the normal curve falls between plus and minus $z = 1.96$ and 99 percent falls between plus and minus $z = 2.58$. These boundaries become important when we discuss the use of the normal curve in inferential statistics

Among other applications, the **normal curve** can be used to help people who are unfamiliar with standard scores to interpret them. The most common use of the normal curve in descriptive statistics is going from a given *z* score to a percentile rank.

Correlations

Correlations indicate the relationship between paired scores. The correlation indicates whether the **relationship** between paired scores is positive or negative and the **strength of this relationship.** The pairs may be two scores for the same individual or two individuals matched on some measure such as reading test scores. In addition to looking at correlation through visual means, the researcher can calculate a **correlation coefficient** that represents the correlation.

Pearson Product Moment Correlation Coefficient

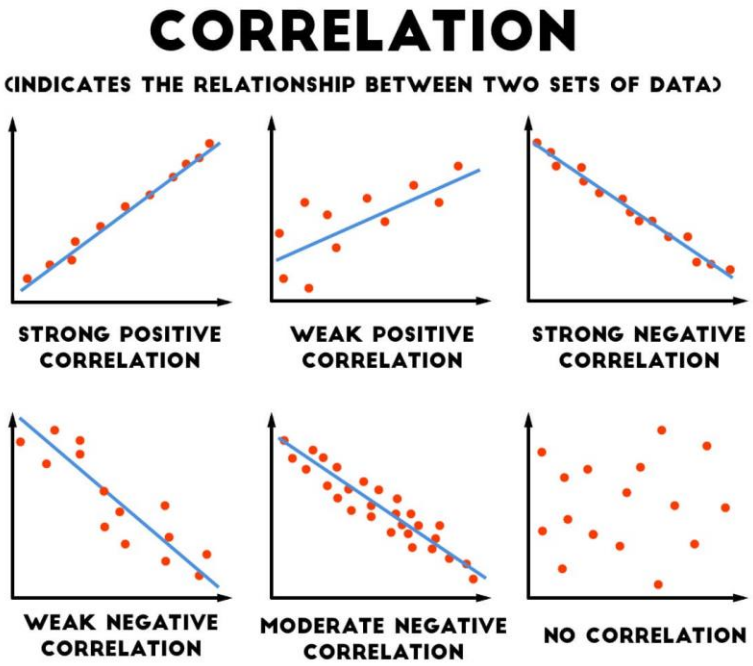
A very useful statistic, the **Pearson product moment correlation coefficient (**Pearson *r*), indicates both the **direction** and the **magnitude** of the relationship between two variables without needing a scatterplot to show it.

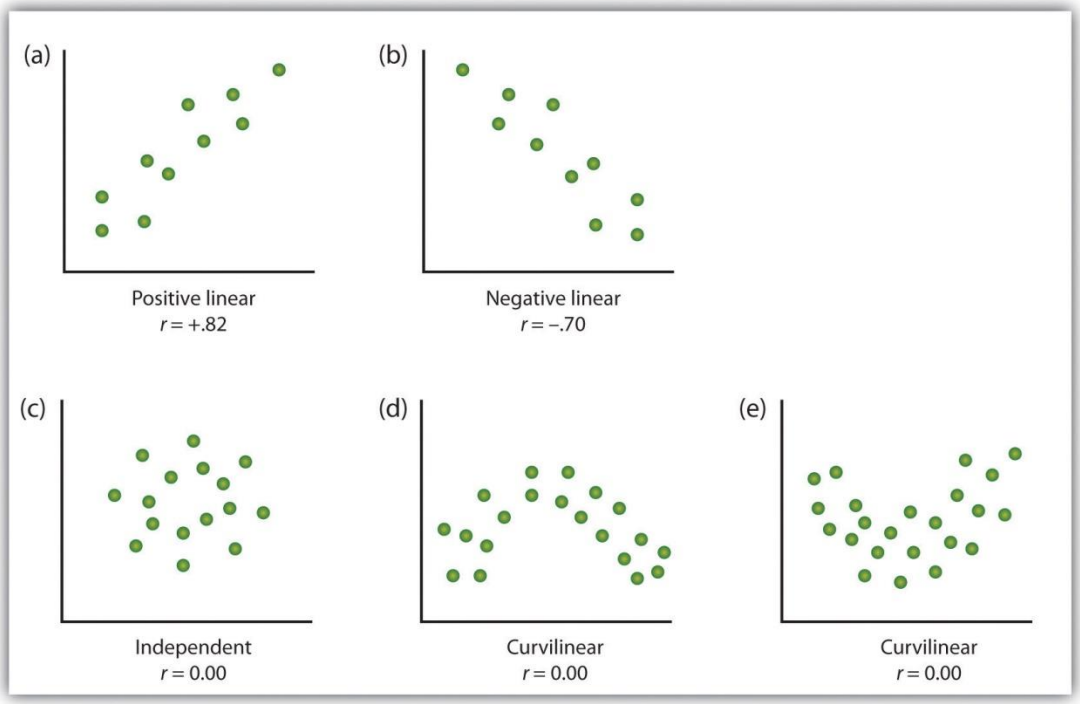
Scatterplot

A scatterplot illustrates the direction of the relationship between the variables. A scatterplot with dots going from **lower left** to **upper right** indicates a positive correlation (as variable *x* goes up, variable *y* also goes up). One with dots going from upper left to lower right indicates a negative correlation (as variable *x* goes up, variable *y* goes down).

A scatterplot of *z* scores also reveals the strength of the relationship between variables. If the dots in the scatterplot **form a narrow band** so that when a straight line is drawn through the band the dots will be near the line, there is a strong **linear relationship** between the variables. However, if the dots in the *z* score scatterplot scatter widely, the relationship between variables is relatively weak.

Like the **mean** and **standard deviation**, the **Pearson *r*** is an interval statistic that can also be used with **ratio data**. An assumption underlying the product moment coefficient of correlation is that the relationship between the two variables (*X* and *Y*) is linear—that is, that a straight line provides a reasonable expression of the relationship of one variable to the other. If a curved line is needed to express this relationship, it is said to be a **curvilinear relationship.** In a curvilinear relationship, as the values of *X* increase, the values of *Y* increase up **to a point**, at which further increases in *X* are associated with decreases in *Y*. An example is the relationship between anxiety and performance. As individuals' anxiety level increases, so does their performance, but only up to a point. With further increases in anxiety, performance decreases.





If the relationship between variables is curvilinear, the computation of the Pearson *r* will result in a misleading underestimation of the degree of relationship. In this case, another index, such as the **correlation ratio (Δ),** should be applied.

Interpretation of Pearson *r*

In interpreting the correlation coefficient, keep the following points in mind:

1. *Correlation does not necessarily indicate* ***causation***. When two variables are found to be correlated, this indicates that relative positions in one variable are *associated* with relative positions in the other variable. It does not necessarily mean that changes in one variable are *caused* by changes in the other variable.
2. *The* **size of a correlation** *is in part a function of the* **variability of the two distributions** *to be correlated*. Restricting the range of the scores to be correlated reduces the **observed degree of relationship** between two variables. For example, people have observed that success in playing basketball is related to height: The taller an individual is, the more probable that he or she will do well in this sport. ***This statement is true about the population at large*** (not restricted in range), where there is a wide range of heights. However, within a basketball team whose members are all tall, there may be little or no correlation between height and success because the range of heights is restricted. For a college that accepts students with a wide range of scores on a scholastic aptitude test, you would expect a correlation between the test scores and college grades. For a college that accepts only students with very high scholastic aptitude scores, you would expect very

little correlation between the test scores and grades because of the restricted range of the test scores in this situation.

3. **Correlation coefficients** should not be interpreted in terms of percentage of perfect correlations. Because correlation coefficients are expressed as decimal fractions, people who are not trained in statistics sometimes interpret correlation coefficients as a percentage of perfect correlation.

An **r** of 80 does not indicate **80** percent of a perfect relationship between two variables. This interpretation is erroneous because, for example, an *r* of .80 does not express a relationship that is twice as great as an *r* of .40. A way of determining the degree to which you can predict one variable from the other is to calculate an index called the **coefficient of determination.** The coefficient of determination is the square of the correlation coefficient. It gives the percentage of variance in one variable that is associated with the variance in the other.

Probably the best way to **give meaning to the size of the correlation coefficient** is to picture the degree of scatter implied by correlations of different sizes and to become familiar with the size of correlations commonly observed between variables of interest.

4. Avoid interpreting the coefficients of correlation in an absolute sense. In interpreting the **degree of correlation**, keep in mind the **purpose** for which it is being used. For example, it may not be wise to use a correlation of**.5** for predicting the future performance of an individual.

Effect Size

What are the main functions of Effect Size?

Effect sizes are interpreted in the same way that **z scores** are interpreted. Effect size can be used to ***compare the direction*** and the ***relative magnitude of the relationships*** that various independent variables have with a common dependent variable. In addition, it can be used to help decide whether the difference an independent variable makes on the dependent variable is strong enough to recommend its implementation in practice. Examples of effect sizes are the correlation between two variables, the regression coefficient in a regression, and the mean difference. Effect sizes complement statistical hypothesis testing, and play an important role in meta-analyses, where the purpose is to combine multiple effect sizes, the standard error (S.E.) of the effect size is of critical importance.

A note of caution: Effect size is independent of sample size. Therefore, large effect sizes can easily be observed through chance alone with very small samples. For example, an effect size of $d = .70$ between two samples of 4 each is essentially meaningless. A rule of thumb is that samples of less than 30 are considered small.

Meta-analysis

Meta-analysis is a statistical technique that combines the effect sizes reported in the results of studies with the same (or similar) independent and dependent variables. The result of a meta-analysis provides an overall summary of the outcomes of a number of studies by calculating a weighted average of their effect sizes. Meta-analysis gives a better estimate of the relationship among variables than do single studies alone.

Chapter 7: Sampling and Inferential Statistics

Inferential Statistics

An important characteristic of inferential statistics is the process of going from the part to the whole. Statistical inference is a procedure by means of which you estimate **parameters** (characteristics of populations) from **statistics** (characteristics of samples). Such estimations are based on the laws of probability and are best estimates rather than absolute facts. In making any such inferences, a **certain degree of error** is involved. Inferential statistics can be used to test hypotheses about populations on the basis of observations of a sample drawn from the population.

Rationale of Sampling

Inductive reasoning is an essential part of the scientific approach. The inductive method involves making observations and then drawing conclusions from these observations. Samples must be ***representative*** if you are to be able to generalize with reasonable confidence from the sample to the population. An unrepresentative sample is termed **a biased sample.** The findings on a biased sample in a research study cannot legitimately be generalized to the population from which it is taken.

Steps in Sampling

The first step in sampling is the identification of the **target population**. We make a distinction between the target population and the **accessible population**, which is the population of subjects accessible to the researcher for drawing a sample.

Sample Selection is the next step.
Two major types of sampling procedures are available to researchers: probability and nonprobability sampling. **Probability sampling** involves sample selection in which the elements are drawn by chance procedures. **Nonprobability sampling** includes methods of selection in which elements are not chosen by chance procedures.

Probability Sampling

The possible inclusion of each population element in this kind of sampling takes place by chance and is attained through random selection.

1.1. Types of probability sampling

1.1.1. Simple Random Sampling

The basic characteristic of simple random sampling is that all members of the population have an equal and independent chance of being included in the **random sample.** It has the following

three steps: (a) ***Define the population***, (b) ***list all members of the population***, and (c) ***select the sample*** by employing a procedure where sheer chance determines which members on the list are drawn for the sample.

A more ***systematic way*** to obtain a random sample is to use a ***table of random numbers***, which includes a series of numbers, typically four to six digits in length, arranged in columns and rows.

1.1.1.1. Advantage

When random sampling is used, the researcher can employ inferential statistics to estimate how much the population is likely to differ from the sample.

1.1.1.2. Disadvantage

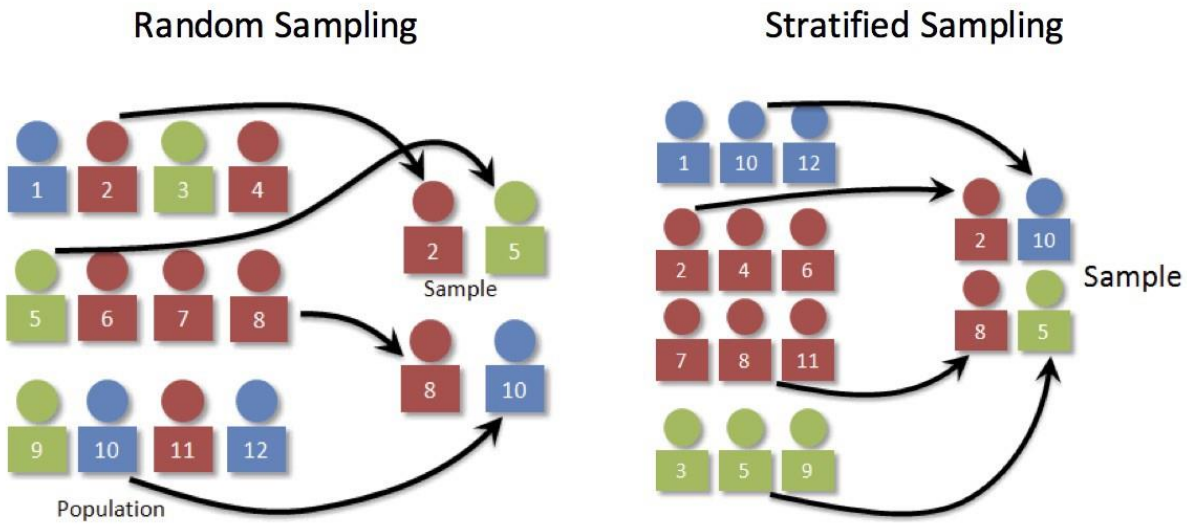
You would expect a random sample to be representative of the target population sampled. However, a random selection, especially with ***small samples***, does not absolutely guarantee a sample that will represent the population well. Unfortunately, simple random sampling requires enumeration of all individuals in a finite population before the sample can be drawn

1.2. Stratified Sampling (96 کنکور)

When the population consists of a number of subgroups, or strata, which may differ in the characteristics being studied, it is often desirable to use a form of probability sampling called stratified sampling. In stratified sampling, you first identify the strata (classes) of interest and then **randomly** draw a specified number of subjects from **each stratum**. The ***basis for stratification*** may be geographic or may involve characteristics of the population such as income, occupation, gender, age, year in college, or teaching level.

1.2.1. Proportional Stratified Sampling

The researcher studies the differences that might exist between various subgroups of a population. In this kind of sampling, you may either take equal numbers from each stratum or select in proportion to the size of the stratum in the population. The latter procedure is known as **proportional stratified sampling,** which is applied when the characteristics of the entire population are the main concern in the study.



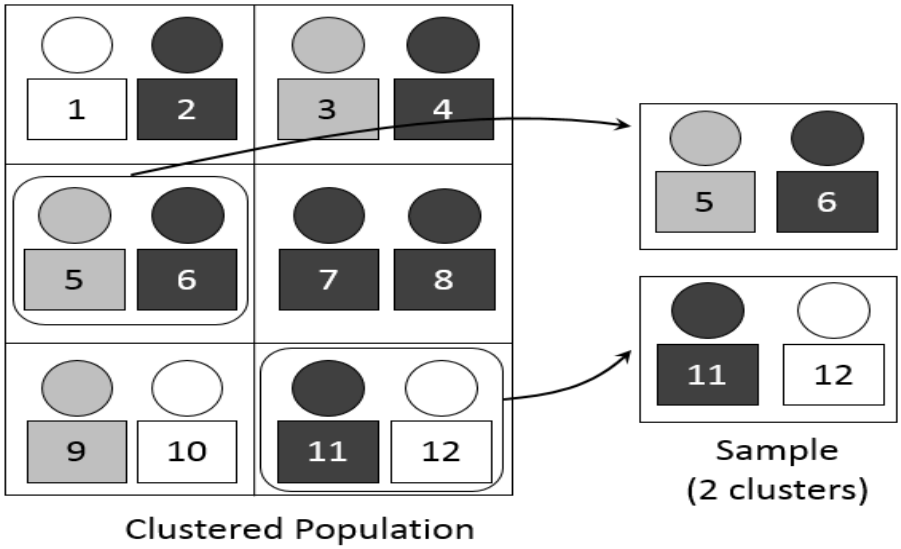
1.2.2. Advantage

- When the population to be sampled is ***not homogeneous*** (***there are some wild variables***) but consists of several subgroups, stratified sampling may give a more representative.

1.3. Cluster Sampling

This kind of probability sampling is referred to as cluster sampling because the unit chosen is not an individual but, rather, a group of individuals who are naturally together. These individuals constitute a cluster insofar as they are alike with respect to characteristics relevant to the variables of the study. A common application of cluster sampling in education is the use of ***intact classrooms*** as clusters.

To illustrate, let us assume a public opinion poll is being conducted in Atlanta. The investigator would probably not have access to a list of the entire adult population; thus, it would be impossible to draw a simple random sample. A more feasible approach would involve the ***selection of a random sample of, for example, 50 blocks*** from a city map and then the polling of all the adults living on those blocks. Each block represents a cluster of subjects, similar in certain characteristics associated with living in proximity.



1.3.1. Disadvantage

The ***sampling error*** in a cluster sample is much greater than in true random sampling.

1.3.2. Procedural Requirement

- It is essential that the clusters actually included in your study be chosen at random from a population of clusters, particularly when the number of clusters is small.
- Once a cluster is selected, all the members of the cluster must be included in the sample

Clustered Sampling

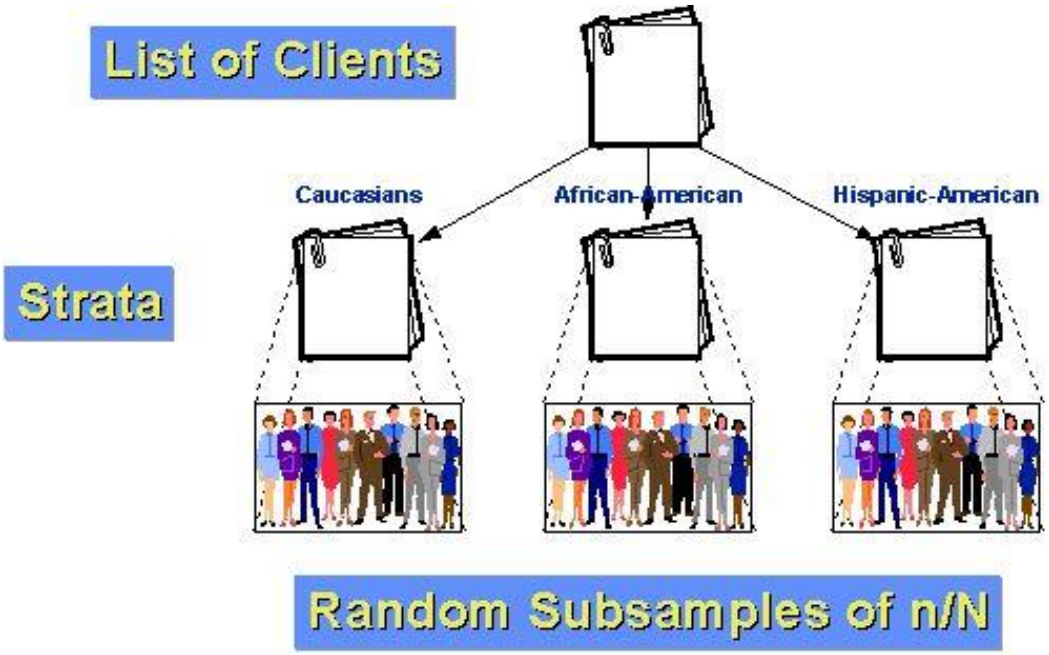
First, the researcher selects groups or clusters, and then from each cluster, the researcher selects the individual subjects by either simple random or systematic random sampling. The researcher can even choose to include the entire cluster and not just a subset from it. The most common cluster used in research is a **geographical cluster**. For example, a researcher wants to survey academic performance of high school students in Spain.

1. He can divide the entire population (population of Spain) into different clusters (cities).
2. Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.
3. Then, from the selected clusters (randomly selected cities) the researcher can either include all the high school students as subjects or he can select a number of subjects from each cluster through simple or systematic random sampling.

Stratified Sampling

Stratified Sampling is a method of sampling from a population. In statistical surveys, when subpopulations within an overall population vary, it is advantageous to sample each

subpopulation (stratum) independently. **Stratification** is the process of dividing members of the population into homogeneous subgroups before sampling. The strata should be **mutually exclusive**: every element in the population must be assigned to only one stratum. The strata should also be **collectively exhaustive**: no population element can be excluded.



1.4. Systematic Sampling

This procedure involves drawing a sample by taking ***every Kth case*** from a list of the population.

1.4.1. Procedure

First, you decide how many subjects you want in the sample (*n*). Because you know the total number of members in the population (*N*), you simply divide *N* by *n* and determine the ***sampling interval*** (*K*) to apply to the list. Select the first member randomly from the first *K* members of the list and then select every *K*th member of the population for the sample. For example, let us assume a total population of 500 subjects and a desired sample size of 50: $K = N/n = 500/50 = 10$.

1.4.2. Disadvantage

- Systematic sampling differs from simple random sampling in that the various ***choices are not independent***. Once the first case is chosen, all subsequent cases to be included in the sample are automatically determined.
- ***Biased sample***; if the original ***population list is*** in random order, systematic sampling would yield a sample that could be statistically considered a reasonable substitute for a random sample. However, if ***the population list is not random***, it is possible that every *K*th member of the population might have some unique characteristic that would affect the dependent variable of the study and thus yield a ***biased sample***.

Combination of Samplings

Note that the various types of probability sampling are not mutually exclusive. Various combinations may be used. For example, you could use cluster sampling if you were studying a very large and widely dispersed population. At the same time, you might be interested in stratifying the sample to answer questions regarding its different strata.

Nonprobability Sampling

In nonprobability sampling, there is no assurance /əˈʃʊər(ə)ns/ that every element in the population has a chance of being included. Its main advantages are convenience and economy.

Convenience (opportunity) Sampling

It is regarded as the ***weakest of all sampling*** procedures, involves using available cases for a study. Using a large undergraduate class, using the students in your own classroom as a sample, or taking volunteers to be interviewed in survey research are various examples of convenience sampling.

Error Estimation

There is no way (except by repeating the study using probability sampling) of estimating the error introduced by the convenience sampling procedures.

1.4.3. Finding Generalizability

If you do use convenience sampling, be extremely cautious in interpreting the findings and know that you cannot generalize the findings.

Purposive Sampling (Judgment Sampling)

In **purposive sampling** sample elements judged to be typical, or representative, are chosen from the population. The assumption is that errors of judgment in the selection will counterbalance one another. Researchers often use purposive sampling for forecasting national elections.

- **Disadvantage**

There is no reason to assume that the units judged to be ***typical of the population*** will continue to be typical over a ***period of time***. Consequently, the results of a study using purposive sampling may be ***misleading***. Because of its low cost and convenience, purposive sampling has been useful in **attitude** and **opinion surveys**.

Quota Sampling (dimensional Sampling)

Quota sampling involves selecting typical cases (selecting based on categories) from diverse strata of a population. It is similar to proportional stratified random sampling **without the random element**. That is, we start off with a sampling frame and then determine the main proportion of the subgroups defined by the parameters included in the frame.

The quotas are based on known characteristics of the population to which you wish to generalize.

Disadvantage
The major weakness of quota sampling lies in the selection of individuals from each stratum. You simply do not know whether the individuals chosen are representative of the given stratum. The selection of elements is likely to be based on accessibility and convenience.

Snowball Sampling
The researcher identifies a few people who meet the criteria of the particular study and then asks these participants to identify further appropriate members of the population. This is suitable for groups whose membership is not really identifiable or the group members which access to them is difficult.

Random Assignment
We distinguish random sampling from random assignment. **Random assignment** is a procedure used after we have a sample of participants and before we expose them to a treatment.

Random assignment or **random placement** is an experimental technique for assigning human participants or animal subjects to different groups in an experiment (e.g., a treatment group versus a control group) using randomization, such as by a chance procedure (e.g., flipping a coin) or a random number generator. This ensures that each participant or subject has an equal chance of being placed in any group. Random assignment of participants helps to ensure that any differences between and within the groups are **not systematic** at the outset of the experiment.

The Size of the Sample

A larger sample is more likely to be a good representative of the population than a smaller sample. However, the most important characteristic of a sample is its ***representativeness***, not its size. Size alone will not guarantee accuracy. A sample may be large and still contain a bias.

The researcher must recognize that sample size will not compensate for any bias that ***faulty sampling techniques*** may introduce. Representativeness must remain the prime goal in sample selection.

The Concept of Sampling Error
When an inference is made from a sample to a population, a certain amount of error is involved because even random samples can be expected to vary from one to another. The ***mean intelligence score*** of one random sample of fourth-graders will probably differ from the mean intelligence score of another random sample of fourth-graders from the same population. Such differences, called **sampling errors,** result from the fact that the researcher has observed only a sample and not the entire population.

Sampling error is calculated from the following formula:

$$e = X - \mu$$

where "e" is sampling error, X is the mean of the sample and μ is symbolized as the mean of the entire population. Put simply, sampling error is the ***difference*** between a population parameter and a sample statistic.

For example, if you know that the mean intelligence score for a population of 10,000 fourth-graders is $\mu = 100$ and a particular random sample of 200 has a mean of $X = 99$, then the sampling error is $X - \mu = 99 - 100 = -1$.

نمونه امار (sample statistics) در تخمین پارامترهای جمعیت (population parameters) بسیار مهم است. از این‌رو در آمار استنباط اختلاف بین نمونه و جمعیت بسیار مهم است.

The Strategy of Inferential Statistics

Inferential statistics is the science of making reasonable decisions with limited information. **Inferential statistics** are used to make generalizations from a sample to a population. There are two **sources of error** that may result in a sample's being *different* from the population from which it is drawn. These are (a) **sampling error** (random error due to chance) and (b) **sample bias** (constant error, due to inadequate design). Inferential statistics take into account **sampling error**. Inferential statistics do *not* correct for sample bias. Inferential statistics only address random error (chance).

p value

The reason for calculating an inferential statistic is to get a ***p* value** (*p* = probability). The *p* value is the probability that the samples are from the *same* population *with regard to the dependent variable* (outcome). Usually, the hypothesis we are testing is that the samples (groups) differ on the outcome. The *p* value is directly related to the **null hypothesis**. The *p* value determines whether or not we reject the null hypothesis. We use it to estimate whether or not we think the null hypothesis is true.

- **If the *p* value is small, reject the null hypothesis** and accept that the samples are truly different with regard to the outcome. It means the independent variable had been effective.
- **If the *p* value is large, accept the null hypothesis** and conclude that the treatment or the predictor variable had no effect on the outcome.

Decision rules - Levels of significance

How small is "small?" Once we get the p value (probability) for an inferential statistic, we need to make a decision. Do we accept or reject the null hypothesis? What *p* value should we use as a cutoff? In the behavioral and social sciences, a general pattern is to use either ***.05 or .01*** as the cutoff. The one chosen is called ***the level of significance.*** If the probability associated with an inferential statistic is equal to or less than .05, then the result is said to be significant at the .05 level. If the .01 cutoff is used, then the result is significant at the .01 level.

Rejecting or accepting the null hypothesis is a **gamble**. There is always a possibility that we are **making a mistake in rejecting the null hypothesis.** This is called a **Type I Error** - rejecting the null hypothesis when it is true. If we use a .01 cutoff, the chance of a Type I Error is 1 out of 100. With a .05 level of significance, we are taking a bigger gamble. There is a 1/20 (5 out of 100) chance that we are wrong, and that our treatment (or predictor variable) doesn't really matter.

Null hypothesis is a statement that there is *no* actual relationship between the variables and that any observed relationship is only a ***function of chance***.

Type I and Type II Errors

The investigator will either retain or reject the null hypothesis. Either decision may be correct or wrong. If the null hypothesis is true, the investigator is correct in retaining it and in error in rejecting it. The rejection of a ***true null hypothesis*** is labeled a **Type I error.** If the null hypothesis is false, the investigator is in error in retaining it and correct in rejecting it. The retention of a ***false null hypothesis*** labeled a **Type II error.**

Put simply, when the researcher concludes that there is no difference between the means – and in fact there is a difference – it is type I. However, when the researcher concludes that there is a difference between the means – and in fact there is no difference – it is type II error.

In interpreting the observed difference between the groups, the investigator must choose between the chance explanation (null hypothesis) and the explanation that states there is a relationship between variables (research hypothesis) and must do so without knowing the ultimate truth concerning the populations of interest.

Level of Significance

The predetermined level at which a null hypothesis would be rejected is called the **level of significance.** The probability of a Type I error is directly **under the control of the researcher**, who sets the level of significance according to the type of error he or she wishes to guard against.

Of course, a researcher could avoid Type I errors by always retaining the null hypothesis or avoid Type II errors by always rejecting it. Neither of these alternatives is productive. If the

consequences of a Type I error would be very serious but a Type II error would be of little consequence, the investigator might decide to risk the possibility of a Type I error only if the estimated probability of the observed relationship's being caused by mere luck is 1 chance in 1000 or less.

محقق تلاش می کند که امکان رخداد خطای نوع اول (به اشتباه فرض صفر را رد کردن) را به حداقل برساند. بنابراین، از احتمال رخداد 0.01 یا 0.05 استفاده می کند. به این معنی که امکان اشتباه رد کردن فرض صفر 0.01 یا 0.05 می باشد. درصورتیکه استفاده از این دو عدد از اهمیت بالایی برخوردار است که اشتباه دررد فرضیه صفر نتایج بد و ناخوشایندی را به همراه داشته باشد.

This is testing the hypothesis at the .001 level of significance, which is considered to be a quite conservative level. In this case, the investigator is being very careful not to declare that a relationship exists when there is no relationship. However, this decision means accepting a high probability of a Type II error, declaring there is no relationship when in fact a relationship does exist.

If the consequences of a Type I error are judged to be not serious, the investigator might decide to declare that a relationship exists if the probability of the observed relationship's being caused by mere luck is 1 chance in 10 or less. This is called "testing the hypothesis at the .10 level of significance'. Here, the investigator is taking only moderate precautions against a Type I error but is not taking a great risk of a Type II error.

The *level of significance* is the probability of a Type I error that an investigator is willing to risk in rejecting a null hypothesis. It is symbolized by the lowercase Greek alpha (α).

عدد 0.05 در رد فرضیه ی صفر یعنی اینکه محقق درنظر دارد بگوید که احتمال وجود رابطه بین متغییرها 99.95 درصداست و تنها 0.05 درصد احتمال دارد که بین دو متغییر رابطه ای وجود ندارد.

Traditionally, investigators determine the level of significance after weighing the relative seriousness of Type I and Type II errors but **before running the experiment**. If the data derived from the completed experiment indicate that the **probability of the null hypothesis** being true is **equal to or less than the predetermined acceptable probability**, the investigators reject the null hypothesis and declare the results to be statistically significant. If the probability is greater than the **predetermined acceptable probability**, the results are described as nonsignificant−that is, the null hypothesis is retained.

The familiar meaning of the word *significant* is "important" or "meaningful". In statistics, this word means "less likely to be a function of chance than some predetermined probability".

Directional and Non-directional Tests

In testing a null hypothesis, researchers **are not** usually concerned with the **direction of the differences**. Rather, they are interested in knowing about the **possible departure** of sample statistics from population parameters. When comparing the effectiveness of competing treatments, an investigator usually wants to learn if treatment A is better than treatment B or if treatment B is better than treatment A. This kind of test is called **non-directional** (two-tailed) because the investigator is interested in differences in either direction. The investigator states only that there will be a difference.

However, if only one alternative to the null hypothesis is of interest, a **directional test** (one-tailed) is used. For example, an investigator studying the effects of a specific diet among obese people would only be interested in assessing the probability that the diet reduces weight.

If on the basis of experience, previous research, or theory the researcher chooses to state the direction of possible differences, then he or she would perform a directional test. A directional hypothesis would state *either* that the parameter is greater than *or* that the parameter is less than the hypothesized value. Thus, in directional tests the critical region is located in only one of the two tails of the distribution.

One tailed or two tailed Hypothesis?

A **one-tailed** directional hypothesis **predicts** the nature of the **effect** of the independent variable on the dependent variable.

> Adults will correctly recall more words than children.

A **two-tailed** non-directional hypothesis predicts that the independent variable will have an effect on the dependent variable, but the direction of the effect is not specified.

• E.g.: There will be a difference in how many numbers are correctly recalled by children and adults.

Null Hypothesis

In research studies involving two groups of participants (e.g., experimental group vs. control group), the null hypothesis always predicts that there will be no differences between the groups being studied (Kazdin, 1992).

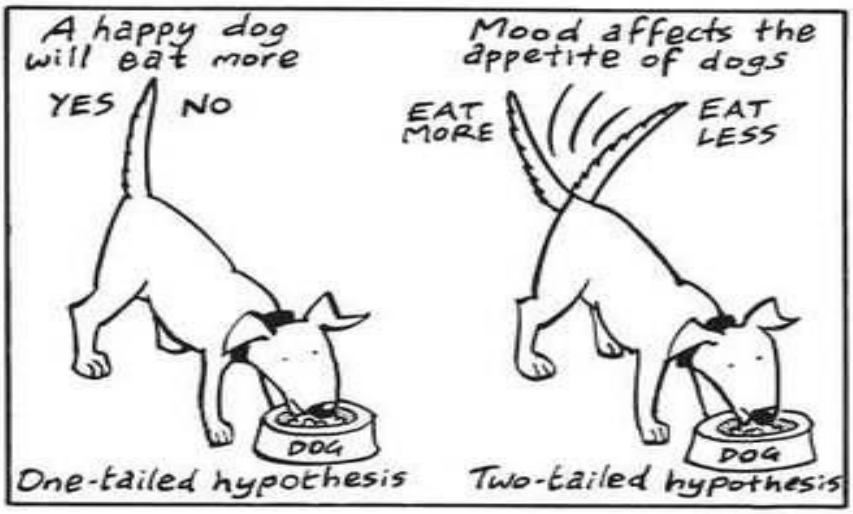
Non-directional Hypotheses

If the hypothesis simply predicts that there will be a difference between the two groups, then it is a nondirectional hypothesis (Marczyk, DeMatteo and Festinger, 2005). It is nondirectional because it predicts that there will be a difference but **does not specify** how the groups will differ.

Directional Hypotheses

If, however, the hypothesis uses so-called comparison terms, such as “greater,” “less,” “better,”

or "worse," then it is a directional hypothesis. It is directional because it predicts that there will be a difference between the two groups and it specifies how the two groups will differ (Marczyk, DeMatteo and Festinger, 2005).

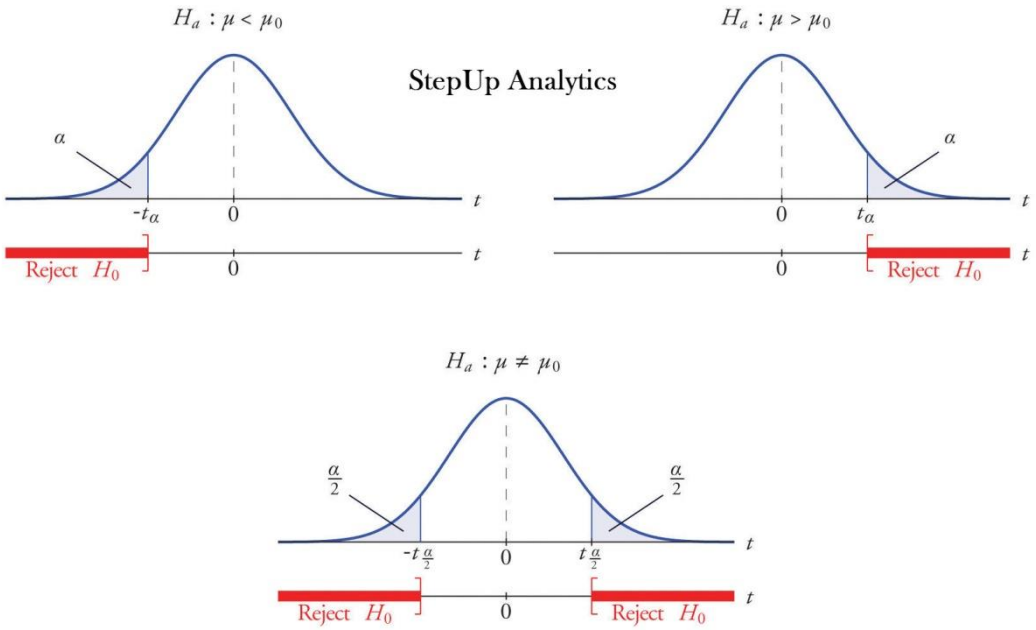


One-Tailed and Two Tailed Tests

- Where H_1 is Directional, ***One-Tailed Test***
- Where H_1 is Non-Directional, ***Two-Tailed Test***

TYPE OF TESTS	DIFFERENCE
One-Tailed Test	Region of Rejection lies entirely in one end of the distribution. Hypothesizing a Range of Values
Two Tailed Test	Involves a Critical Region which is split into two equal parts placed in each tail of the distribution. A value of the parameter is being hypothesized.

Mathematical Formulation of H_1	Region of Rejection
Greater Than (>)	Area of Rejection is placed entirely in the Right Tail of the Distribution
Less Than (<)	Region of Rejection is in the Left Tail
Not Equal To (≠)	Both Tails contain Equal areas serving as Critical Regions



Determining the Appropriate Sample Size

A scientific method of determining the sample size needed is to specify a meaningful **effect size** (Δ or *d*) and then determine the sample size needed to reach a desired probability of rejecting the null hypothesis at a given level of significance. Recall that effect size is the difference between experimental and control groups divided by the standard deviation of the control group (Δ) or the difference between two groups divided by the estimated population standard deviation (*d*).

'Effect size' is simply a way of **quantifying the size of the difference** between two groups. It is particularly valuable for quantifying the **effectiveness of a particular intervention,** relative to some comparison. It allows us to move beyond the simplistic, 'Does it work or not?' to the far more sophisticated, 'How well does it work in a range of contexts?' Moreover, by placing the emphasis on the **most important aspect of an intervention** - the size of the effect - rather than its statistical significance (which conflates effect size and sample size), it promotes a more scientific approach to the accumulation of knowledge. For these reasons, effect size is an important tool in reporting and interpreting effectiveness.

In statistics, an **effect size** is a quantitative measure of the strength of a phenomenon.[1] Examples of effect sizes are the correlation between two variables, the regression coefficient in a regression, the mean difference, or even the risk with which something happens, such as how many people survive after a heart attack for every one person that does not survive.

The specification of what is a meaningful effect size is a **judgment call**. However, professionals in their fields are usually able to specify an effect size that serves as a reasonable dividing line between meaningful and trivial differences (because there is not a general, acceptable rule or indices).

Determining the number needed in a sample is really a function of how precise you want to be– that is, how large or small an effect size you want to be statistically significant, how much chance of Type I error you are willing to live with, and how much probability of rejecting a false null hypothesis you want. These are all judgment calls, but they can all be made on a rational basis.

POWER

Power is the ability to reject a null hypothesis when it is false. The **power** of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis (H_0) when a specific alternative hypothesis (H_1) is true. The statistical power ranges from 0 to 1, and as statistical power increases, the probability of making a type 2 error decreases.

Power analysis can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size

Think about it 7.1.

Complete each line with either (a) goes up, (b) goes down, or (c) stays the same.

I. As you choose to put α up from .01 to .05,

1. Probability of Type I error **goes up.**

2. Probability of Type II error **goes down**.

3. Power **goes up**.

II. As you increase the number in the samples,

1. Probability of Type I error **stays the same**.

2. Probability of Type II error **goes down**.

3. Power **goes up**.

III. As the true difference between means goes up from 3 to 7,

1. Probability of Type I error stays the same,

2. Probability of Type II goes down

3. Power goes up.

IV. As effect size increases,

1. Probability of Type I error stays the same

2. Probability of Type II error goes down

3. Power goes up

V. As the heterogeneity (variance) within the samples increases,

1. Probability of Type I error stays same

2. Probability of Type II error goes up

3. Power goes down

VI. If you do a one-tailed test instead of a two-tailed test and if you correctly predicted the direction of the difference

1. Probability of Type I error stays the same

2. Probability of Type II error goes down

3. Power goes up.

The General Strategy of Statistical Tests

A statistical test **compares** what is observed (a statistic) with what we would expect to observe through chance alone. What we would expect through chance alone is called the **error term**. A ratio is formed:

$$\frac{Observation}{chance\ expectastion} = \frac{statistic}{error\ term}$$

When the observed statistic is equal to or less than the average value (mean) expected through chance alone (the **error term**), the most plausible explanation for the statistic is that it was due to chance alone. If the statistic is greater than the error term, then the chance explanation becomes less and less plausible as this ratio becomes greater and greater than 1.

In our math concepts example, the statistic is the difference between the mean of the group taught by method B and the group taught by method A $(\overline{XB} - \overline{XA})$.

What is a 'T-Test'

The **t test** is one type of inferential statistics. It is used to determine whether there is a significant difference between the means of two groups.

A t-test is an analysis of two populations' means through the use of statistical examination; a t-test with two samples is commonly used with small sample sizes, testing the difference between the samples when the variances of two normal distributions are not known.

The t-**test** for **independent means** is used when we want to know whether there is a difference between populations. For instance, we may want to know if college men and women differ on some **psychological** characteristic. ... The t-**test** for independent means is used only for **tests** of the sample means.

The *t* test is one type of inferential statistics. It is used to determine whether there is a significant difference between the means of two groups. With all inferential statistics, we assume the dependent variable fits a normal distribution. When we assume a normal distribution exists, we can identify the probability of a particular outcome. We specify the level of probability (alpha level, level of significance, *p*) we are willing to accept before we collect data ($p < .05$ is a common value that is used). After we collect data we calculate a test statistic with a formula. We compare our test statistic with a critical value found on a table to see if our results fall within the acceptable level of probability. Modern computer programs calculate the test statistic for us and also provide the exact probability of obtaining that test statistic with the number of subjects we have.

When the difference between two population averages is being investigated, a *t* test is used. In other words, a *t* test is used when we wish to compare two means (the scores must be measured on an **interval or ratio** measurement scale). We would use a *t* test if we wished to compare the reading achievement of boys and girls. With a *t*-test, we have **one independent** variable and **one dependent variable**. The independent variable (gender in this case) can only have two levels (male and female). The dependent variable would be reading achievement. If the independent had more than two levels, then we would use a one-way analysis of variance (ANOVA).

The test statistic that a *t* test produces is a *t*-value. Conceptually, *t*-values are an extension of *z*-scores. In a way, the *t*-value represents how many standard units the means of the two groups are apart.

With a *t* test, the researcher wants to state with some degree of confidence that the obtained difference between the means of the sample groups is too great to be a chance event and that some difference also exists in the population from which the sample was drawn. In other words, the difference that we might find between the boys' and girls' reading achievement in our sample might have occurred by chance, or it might exist in the population. If our *t* test produces a *t*-value that results in a probability of .01, we say that the likelihood of getting the difference we found by chance would be 1 in a 100 times. We could say that it is unlikely that our results occurred by chance and the difference we found in the sample probably exists in the populations from which it was drawn.

Five factors contribute to whether the difference between two groups' means can be considered significant:

1. How large is the **difference** between the means of the two groups? Other factors being equal, the greater the difference between the two means, the greater the likelihood that a statistically significant mean difference exists. If the means of the two groups are far apart, we can be fairly confident that there is a real difference between them.

2. How much **overlap** is there between the groups? This is a function of the variation within the groups. Other factors being equal, the smaller the variances of the two groups under consideration, the greater the likelihood that a statistically significant mean difference exists. We can be more confident that two groups differ when the scores within each group are close together.

3. How many **subjects** are in the two samples? The size of the sample is extremely important in determining the significance of the difference between means. With increased sample size, means tend to become more stable representations of group performance. If the difference we find remains constant as we collect more and more data, we become more confident that we can trust the difference we are finding.
4. What **alpha level** is being used to test the mean difference (how confident do you want to be about your statement that there is a mean difference). A larger alpha level requires less difference between the means. It is much harder to find differences between groups when you are only willing to have your results occur by chance 1 out of a 100 times ($p < .01$) as compared to 5 out of 100 times ($p < .05$).
5. Is a directional (one-tailed) or non-directional (two-tailed) hypothesis being tested? Other factors being equal, smaller mean differences result in statistical significance with a directional hypothesis. For our purposes we will use non-directional (two-tailed) hypotheses.

Assumptions Underlying the t Test

1. The samples have been **randomly** drawn from their respective populations
2. The scores in the population are **normally distributed**
3. The scores in the populations have the **same variance** (s1=s2)

Three Types of *t* tests

- **Pair-difference** *t* test (a.k.a. t-test for dependent groups, correlated *t* test) *df*= *n* (number of pairs) -1

This is concerned with the difference between the average scores of **a single sample of individuals** who are assessed at two different times (such as before treatment and after treatment). It can also compare average scores of samples of individuals who are paired in some

way (such as siblings, mothers, daughters, persons who are matched in terms of a particular characteristic).

- ***t* test for Independent Samples (with two options)**

This is concerned with the **difference between the averages** of two populations. Basically, the procedure compares the averages of two samples that were selected independently of each other, and asks whether those sample averages differ enough to believe that the populations from which they were selected also have different averages. An example would be comparing math achievement scores of an experimental group with a control group.

1. **Equal Variance** (Pooled-variance t-test) df=n (total of both groups) -2 Note: Used when both samples have the same number of subject or when s1=s2 (Levene or F-max tests have $p > .05$).

2. **Unequal Variance** (Separate-variance *t* test) df dependents on a formula, but a rough estimate is one less than the smallest group Note: Used when the samples have different numbers of subjects and they have different variances — s1<>s2 (Levene or F-max tests have $p < .05$).

With two independent samples when the dependent variable is ranked data, the **Mann–Whitney** (nonparametric test) test serves the same purpose as the t test for independent samples.

The *T* Test for Pearson *r* Correlation Coefficients

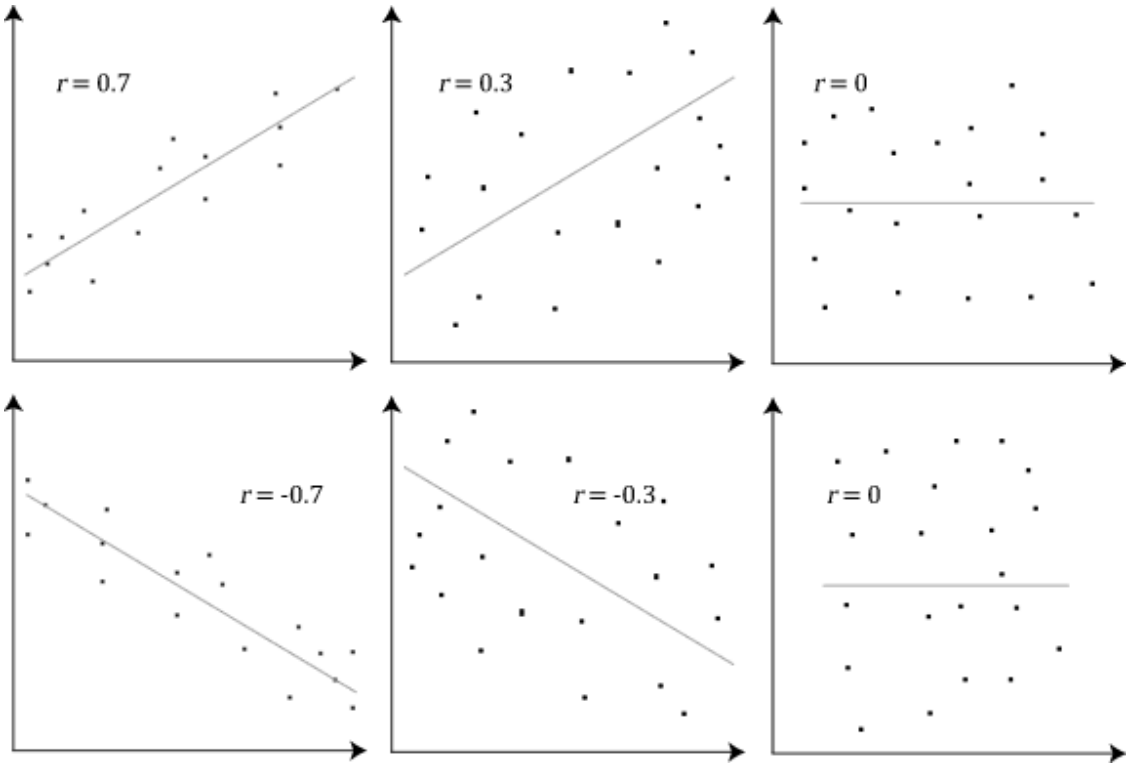
Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a **measure of the strength of the association** between the two variables.

The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

Values of Pearson's correlation coefficient

Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1



Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa.

Significance

The t-test is used to establish if the correlation coefficient is significantly different from zero, and, hence that there is evidence of an association between the two variables. There is then the underlying assumption that the data is from a **normal distribution** sampled **randomly**. If this is not true, the conclusions may well be invalidated. If this is the case, then it is better to use Spearman's coefficient of rank correlation (for non-parametric variables).

Analysis of Variance (ANOVA)

Simple or **one-way analysis of variance (ANOVA)** is a statistical procedure used to analyze the data from a study with more than two groups. The **null hypothesis** is that there is no difference among the group means. It is called one-way ANOVA because there is only one independent variable and one dependent variable. Because ANOVA can be used with more than two groups,

it is a more versatile /ˈvəːsətʌɪl/ technique than the *t* test. A *t* test can be used only to test a difference between *two* means. ANOVA can test the difference between *two or more* means.

The general rationale of ANOVA is that the *total variance* of all subjects in an experiment can be subdivided into two sources: *variance between groups* and *variance within groups*.

Assumptions

ANOVA models are parametric, relying on assumptions about the distribution of the dependent variables (DVs) for each level of the independent variable(s) (IVs).

In practice, the first two assumptions here are the main ones to check. Note that the larger the sample size, the more robust ANOVA is to violation of the first two assumptions: (a) normality and (b) homoscedasticity (homogeneity of variance).

1. **Normality** of the DV distribution[1]: The data in each cell should be approximately normally distributed.
2. **Homogeneity of variance**: The variance in each cell should be similar.
3. **Sample size**: per cell > 20 is preferred;
4. **Independent observations**:

Multi-factor Analysis of Variance

In the multi-factor model, there is a response (**dependent**) variable and one or more factor (independent) variables. This is a common model in designed experiments where the experimenter sets the values for each of the factor variables and then measures the response variable.

The Chi-Square Tests of Significance

The chi-squared test is used to determine whether there is a significant difference between the **expected frequencies** and the **observed frequencies** in one or more categories. Observed frequencies, as the name implies, are the actual frequencies obtained by observation. Expected frequencies are theoretical frequencies that would be observed when the null hypothesis is true.

Chi-Square Goodness of Fit Test (The One-Variable Chi Square)

This test is applied when you have one **categorical variable** from a single population. It is used to determine whether sample data are consistent with a hypothesized distribution.

[1] Distribution is related to dependent variable

For example, suppose a company printed baseball cards. It claimed that 30% of its cards were rookies; 60% were veterans but not All-Stars; and 10% were veteran All-Stars. We could gather a **random sample** of baseball cards and use a chi-square goodness of fit test to see whether our sample distribution differed significantly from the distribution claimed by the company.

When to Use the Chi-Square Goodness of Fit Test

The chi-square goodness of fit test is appropriate when the following conditions are met:

- The sampling method is **simple random sampling**.
- The variable under study is **categorical**.
- The expected value of the number of sample observations in each level of the variable is at least 5.

The Two-Variable Chi Square (Chi-Square Test of Independence)

The two-variable chi-square design uses two independent variables, each with two or more levels, and a dependent variable in the form of a frequency count. The purpose of the test is to determine whether or not the two variables in the design are independent of one another.

Chi-square Test

The Chi-square statistic is a non-parametric (distribution free) tool designed to analyze group differences when the dependent variable is measured at a nominal level. Like all non-parametric statistics, the Chi-square is robust with respect to the distribution of the data. Specifically, it does not require **equality of variances** among the study groups or **homoscedasticity** in the data. It permits evaluation of both dichotomous independent variables, and of multiple group studies. Unlike many other non-parametric and some parametric statistics, the calculations needed to compute the Chi-square provide considerable information about how each of the groups performed in the study. This richness of detail allows the researcher to understand the results and thus to derive more detailed information from this statistic than from many others.

The Chi-square test of independence (also known as the Pearson Chi-square test, or simply the Chi-square) is one of the most useful statistics for testing hypotheses when the **variables** are **nominal**. The Chi-square (χ^2) can provide information not only on the significance of any observed differences, but also provides detailed information on exactly which categories account for any differences found. Thus, the amount and detail of information this statistic can provide renders it one of the most useful tools in the researcher's array of available analysis tool.

The Chi-square test is a non-parametric statistic, also called a distribution free test. Non-parametric tests should be used when any one of the following conditions pertains to the data:

1. The level of measurement of all the variables is nominal or ordinal.

2. The sample sizes of the study groups are unequal; for the χ^2 the groups may be of equal size or unequal size whereas some parametric tests require groups of equal or approximately equal size.

Assumptions of the Chi-square

As with parametric tests, the non-parametric tests, including the $\chi 2$ assume the data were obtained through random selection. The assumptions of the Chi-square include:

1. Observations must be independent−that is, the subjects in each sample must be randomly and independently selected.
2. The categories must be mutually exclusive: Each observation can appear in one and only one of the categories in the table.
3. The observations are measured as frequencies.

Chapter 8: Research Tools

One must select or develop scales and instruments that can measure complex constructs such as intelligence, achievement, personality, motivation, attitudes, aptitudes, interests, and self-concept. There are two basic ways to obtain these measures for your study: Use one that has already been developed or construct your own.

Tests

Tests are valuable measuring instruments for educational research. The utility of these scores as indicators of the construct of interest is in large part a function of the **objectivity**, **validity**, and **reliability** of the tests. Objectivity is the extent of agreement among scorers.

Achievement Tests

Achievement tests are used to measure what individuals have learned. Achievement tests measure mastery and proficiency in different areas of knowledge by presenting subjects with a standard set of questions involving completion of cognitive tasks. Achievement tests are generally classified as either standardized or teacher/researcher made.

Standardized tests are published tests that have resulted from careful and skillful preparation by experts and cover broad academic objectives common to the majority of school systems. These are tests for which comparative norms have been derived, their validity and reliability established, and directions for administering and scoring prescribed.

In selecting an achievement test, **researchers** must be careful to choose one that is reliable and is appropriate (valid) for measuring the aspect of achievement in which they are interested. There should be a direct link between the test content and the curriculum to which students have been exposed. The test must also be valid and reliable for the type of subjects included in the study.

Researcher-Made Tests

It is much better to construct your own test than to use an inappropriate standardized one just because it is available. The advantage of a **researcher-made test** is that it will match more closely the content that was covered in the classroom or in the research study.

Norm-Referenced and Criterion-Referenced Tests

On the basis of the type of interpretation made, standardized and **teacher-made tests** may be further classified as **norm-referenced** or **criterion-referenced.** Norm-referenced tests permit researchers to compare individuals' performance on the test to the performance of other individuals. Typically, standardized tests are norm referenced, reporting performance in terms of **percentiles**, standard scores, and similar measures.

In contrast, criterion-referenced tests enable researchers to describe what a specific individual can do, **without reference to the performance of others**. Performance is reported in terms of the level of mastery of some well-defined content or skill domain. Typically, the level of mastery is indicated by **the percentage of items** answered correctly.

Test Performance Range

The range of performance that an achievement test permits is important. Researchers want a test designed so that the subjects can perform fully to their ability level without being restricted by the test. Two types of testing effects may occur: (a) **ceiling effect** and (b) floor effect.

A **ceiling effect** occurs when many of the scores on a measure are at or near the maximum possible score. Tests with a ceiling effect are **too easy** for many of the examinees, and we do not know what their scores might have been if there had been a higher ceiling. For example, if we gave a 60-item test and most of the scores fell between 55 and 60, we would have a ceiling effect. A graph of the frequency distribution of scores would be **negatively skewed**.

Likewise, test performance may be restricted at the lower end of the range, resulting in a **floor effect.** A floor effect occurs when a test is **too difficult** and many scores are near the minimum possible score. A graph of the frequency distribution of scores would be **positively skewed**.

Performance Assessments

Another way to classify achievement tests is whether they are verbal or **performance tests.** Performance assessment, usually administered individually, is **a popular alternative** to traditional **paper-and-pencil tests** among educators. A performance test is a technique in which a researcher directly observes and assesses an individual's performance of a certain task and/or judges the finished product of that performance. The test taker is asked to carry out a *process* such as playing a musical instrument or tuning a car engine or to produce a *product* such as a written essay. The performance or product is judged against established criteria.

Performance assessments provide an opportunity for teachers and researchers to gain a more holistic view of changes in students' performance over time.

Aptitude Test

Aptitude tests differ from achievement tests in that aptitude tests attempt to measure **general ability** or **potential for learning** a body of knowledge and skills, whereas achievement tests attempt to measure the actual extent of acquired knowledge and skills in specific areas. Aptitude tests measure a subject's ability to perceive relationships, solve problems, and apply knowledge in a variety of contexts.

Educators have found aptitude tests useful and generally valid for the purpose of predicting school success. Many of the tests are referred to as **scholastic aptitude tests,** a term pointing out

specifically that the main function of these tests is to predict school performance. Well-known aptitude tests are the ACT (American College Testing Assessment) and the SAT (Scholastic Assessment Test) for high school students and the GRE (Graduate Record Exam) and MAT (Miller Analogies Test) for college seniors. Researchers often use aptitude tests. Aptitude or intelligence is frequently a variable that needs to be controlled in educational experiments. To control this variable, the researcher may use the scores from a scholastic aptitude test.

Measures of Personality

There are several different types of personality measures, each **reflecting a different theoretical point of view**. Some reflect trait and type theories, whereas others have their origins in psychoanalytic and motivational theories. Researchers must know precisely what they wish to measure and then select the instrument, paying particular attention to the evidence of **its validity**. Two approaches are used to measure personality: objective personality assessment and projective personality assessment.

Objective Personality Assessment

Self-report inventories present subjects with an extensive collection of statements describing behavior patterns and ask them to indicate whether or not each statement is characteristic of their behavior by checking *yes*, *no*, or *uncertain*. Other formats use multiple choice and true‑false items. The score is computed by counting the number of responses that agree with a trait the examiner is attempting to measure. For example, someone with paranoid tendencies would be expected to answer *yes* to the statement "People are always talking behind my back" and *no* to the statement "I expect the police to be fair and reasonable". Of course, similar responses to only two items would not indicate paranoid tendencies. However, such responses to a large proportion of items could be considered an indicator of paranoia.

Inventories have the advantages of economy, simplicity, and objectivity. They can be administered to groups and do not require trained psychometricians. Most of the disadvantages are related to the **problem of validity**. The validity of self-report inventories depends in part on the respondents' **being able to read and understand the items**, their understanding of themselves, and especially their willingness to give **frank and honest answers**. As a result, the information obtained from inventories may be superficial or biased.

Projective Personality Assessment

Projective techniques are measures in which an individual is asked to respond to an ambiguous or unstructured stimulus. They are called *projective* because a person is expected to project into the stimulus his or her own needs, wants, fears, beliefs, anxieties, and experiences. On the basis of the subject's interpretation of the stimuli and his or her responses, the examiner attempts to

construct a comprehensive picture of the individual's **personality structure**. Projective methods are used mainly by **clinical psychologists** for studying and diagnosing people with emotional problems. They are **not frequently used in educational research** because of the necessity of specialized training for administration and scoring and the expense involved in individual administration.

Scales

Scales are used to measure attitudes, values, opinions, and other characteristics that are **not easily measured by tests** or other measuring instruments. A **scale** is a set of categories or numeric values assigned to individuals, objects, or behaviors for the purpose of measuring variables. The process of assigning scores to those objects in order to obtain a measure of a construct is called *scaling*. **Scales differ from tests** in that the results of these instruments, unlike those of tests, do not indicate success or failure, strength or weakness. T**hey measure the degree** to which an individual exhibits the characteristic of interest. For example, a researcher may use a scale to measure the attitude of college students toward religion or any other topic. A number of scaling techniques have been developed throughout the years.

Attitude Scales

Attitude scales use multiple responses—usually responses to statements—and combine the responses into a single scale score. Rating scales use judgments—made by the individual under study or by an observer—to assign scores to individuals or other objects to measure the underlying constructs. The measurement of attitudes presumes the ability to place individuals **along a continuum** of favorableness-unfavorableness toward the object.

We discuss two types of attitude scales: summated or **Likert scales** and **bipolar adjective scales**.

A Likert scale is constructed by assembling a large number of statements about an object, approximately half of which express a clearly favorable attitude and half of which are clearly unfavorable. **Neutral items** are not used in a Likert scale. It is important that these statements constitute a representative sample of all the possible opinions or attitudes about the object. It may be helpful to think of all the subtopics relating to the attitude object and then write items on each subtopic.

Item Analysis

After administering the attitude scale to a preliminary group of respondents, the researcher does an **item analysis** to identify the best functioning items. **Item analysis** is a process which examines student responses to individual test **items** (questions) in order to assess the **quality of those items** and of the test as a whole. The item analysis typically yields three statistics for each

item: (1) an item **discrimination index,** (2) the percentage of respondents marking each choice to each item, and (3) the item mean and standard deviation.

Validity

Validity concerns the extent to which the scale really measures the attitude construct of interest.

Reliability

The reliability of the new scale must also be determined. Reliability is concerned with the extent to which the measure would yield consistent results each time it is used. The first step in ensuring reliability is to make sure that the scale is long enough.

Bipolar Adjective Scales

The **bipolar adjective scale** presents a respondent with a list of adjectives that have bipolar or opposite meanings. Respondents are asked to place a check mark at one of the seven points in the scale between the two opposite adjectives to indicate the degree to which the adjective represents their attitude toward an object, group, or concept.

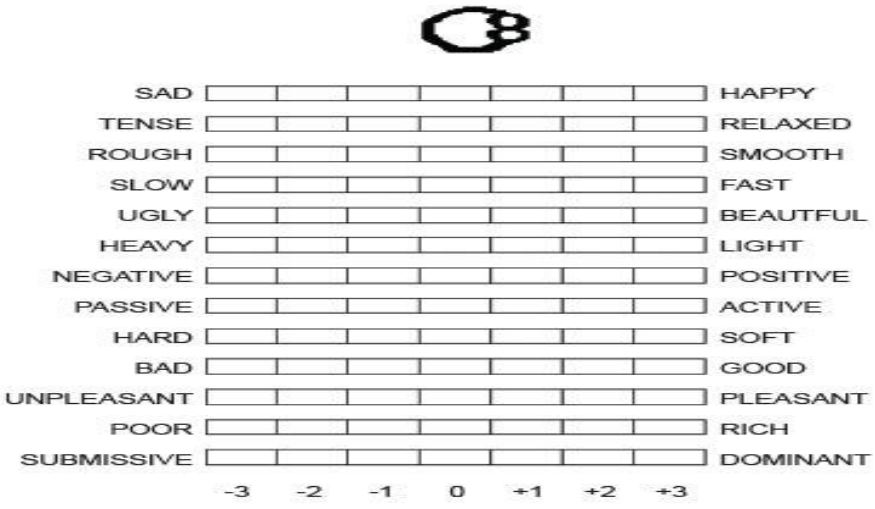


Figure 2. Example of a "D" stimulus above the set of bipolar scales.

The bipolar adjective scale is a very flexible approach to measuring attitudes. A researcher can use it to investigate attitudes toward any concept, person, or activity in any setting. It is much easier and **less time-consuming** to construct than a Likert scale. Instead of having to come up with approximately 20 statements, you need only select **four** to **eight** adjective pairs. It requires very little reading time by participants. The main difficulty is the selection of the adjectives to use.

Rating Scales

There are several rating scales: (a) Graphic scales, (b) Category scales, and (c) Comparative rating scales.

Errors in Rating

Because ratings depend on the perceptions of human observers, who are susceptible to various influences, rating scales are subject to considerable error. Among the most frequent systematic errors in rating people is the **halo effect** which occurs when raters allow a generalized impression of the subject to influence the rating given on very specific aspects of behavior.

The halo effect is a type of cognitive bias in which our overall impression of a person influences how we feel and think about his or her character. Essentially, your overall impression of a person ("He is nice!") impacts your evaluations of that person's specific traits ("He is also smart!"). One great example of the halo effect in action is our overall impression of celebrities. Since we perceive them as attractive, successful, and often likable, we also tend to see them as intelligent, kind, and funny.

Another type of error is the **generosity error,** which refers to the tendency for raters to give subjects the benefit of any doubt. When raters are not sure, they tend to rate people favorably. In contrast, the **error of severity** is a tendency to rate all individuals too low on all characteristics. Another source of error is the **error of central tendency,** which refers to the tendency to avoid either extreme and to rate all individuals in the middle of the scale.

How could we reduce errors?

One way of reducing such errors is to **train the raters** thoroughly before they are asked to make ratings. They should be informed about the possibility of making these "personal bias" types of errors and how to avoid them. It is absolutely essential that raters have adequate time to observe the individual and his or her behavior before making a rating.

Another way to minimize error is to make certain that the **behavior** to be rated and the **points** on the rating scale are clearly **defined**. The **points** on the scale **should be described** in terms of overt behaviors that can be observed, rather than in terms of behaviors that require inference on the part of the rater.

The accuracy or reliability of ratings is usually increased by having two (or more) trained raters make independent ratings of an individual. These independent ratings are pooled, or averaged, to obtain a final score. A researcher may also correlate the ratings of the two separate raters in order to obtain a coefficient of interrater reliability. The size of the coefficient indicates the extent to which the raters agree. An **interrater reliability** coefficient of .70 or higher is considered acceptable for rating scales.

Direct Observation

In many cases, systematic or **direct observation** of behavior is the most desirable measurement method. Observation is used in both quantitative and qualitative research. When observations are made in an attempt to obtain a comprehensive picture of a situation, and the product of those observations is notes or narratives, the research is qualitative.

Researchers use **checklists**, rating scales, and coding sheets to record the data collected in direct observation.

Advantages of Direct Observation

It provides a record of the actual behavior that occurs. It is appropriate to be used with young children. It is used extensively in research on infants and on preschool children who have difficulty communicating through language and may be uncomfortable with strangers. Another advantage is that systematic observation can be used in natural settings.

The main **disadvantage** of systematic observation is the expense. Observations are more costly because of the time required of trained observers. Subjects may be observed for a number of sessions, requiring extended hours.

Validity and Reliability of Direct Observation

The best way to enhance validity is to **carefully define the behavior** to be observed and to **train the people** who will be making the observations. Observers must be aware of two sources of bias that affect validity: (a) observer bias and (b) observer effect.

Observer bias occurs when the observer's own perceptions, beliefs, and biases influence the way he or she observes and interprets the situation. Having more than one person make independent observations helps to detect the presence of bias.

Observer effect occurs when people being observed behave differently just because they are being observed. In many cases, however, after an initial reaction the subjects being observed come to pay little attention to the observer, especially one who operates unobtrusively. Some studies have used interactive television to observe classrooms unobtrusively. Videotaping for later review and coding may also be useful.

The accuracy or reliability of direct observation is usually investigated by having at least two observers independently observe the behavior and then determining the extent to which the observers' records agree.

Contrived (adjective): deliberately created rather than arising naturally or spontaneously.

- Created or arranged in a way that seems artificial and unrealistic.

"The ending of the novel is too pat and contrived"

Contrived Observation

In **contrived observations,** the researcher arranges for the observation of subjects in simulations of real-life situations. The circumstances have been arranged so that the desired behaviors are elicited. One form of contrived observation is the **situational test.**

Data Collection in Qualitative Research

The most widely used tools in qualitative research are interviews, document analysis, and observation.

Validity and Reliability

Historically, **validity** was defined as the extent to which an instrument measured what it claimed to measure. The focus of recent views of validity is not on the instrument itself but on the interpretation and meaning of the scores derived from the instrument.

Constructs such as intelligence, creativity, anxiety, critical thinking, motivation, self-esteem, and attitudes represent **abstract variables** derived from theory or observation. Researchers have no direct means of measuring these constructs. To measure these hypothetical constructs, you must move from the **theoretical domain** surrounding the construct to an **empirical level** that operationalizes the construct. That is, we use an operational definition to measure the construct. We do this by selecting specific sets of observable tasks believed to serve as indicators of the particular theoretical construct. Then we assume that performance (scores) on the tasks reflects the particular construct of interest as distinguished from other constructs. Essentially, validity deals with how well the operational definition fits with the conceptual definition.

Why should we operationalize the construct?

Messick (1995) identified two problems that **threaten** the interpretation (validity) of test scores: construct underrepresentation and construct-irrelevant variance. The term **construct underrepresentation** refers to assessment that is too narrow and fails to include important dimensions of the construct. The test may not adequately sample some kinds of content or some types of responses or psychological processes and thus fails to adequately represent the theoretical domain of the construct.

The term **construct-irrelevant variance** refers to the extent to which test scores are affected by variables that are extraneous to the construct.

Validation

The process of **gathering evidence** to support (or fail to support) a particular interpretation of test scores is referred to as validation. We need evidence to establish that the **inferences**, which are made on the basis of the test results, are **appropriate**.

Evidence Based on Test Content

Evidence based on test content involves the test's content and its relationship to the construct it is intended to measure. Content-related evidence is the degree to which the samples of items, tasks, or questions on a test are representative of some defined universe or domain of content". That is, the researcher must seek evidence that the test to be used represents an adequate sampling of all the relevant knowledge, skills, and dimensions making up the content domain.

Evidence based on test content is especially important in evaluating **achievement tests**. In this age of educational accountability, content validity is receiving renewed attention.

Accountability is considered as being able to demonstrate the extent to which we have effectively and efficiently discharged responsibility and point out that without accountability in language teaching, students can pass several semesters of language courses with high grads and still be unable to use the language for reading or for conversing with speakers of that language.

To ensure content validity in a classroom test, a teacher should prepare a "**blueprint**" showing the content domain covered and the relative emphasis given to each aspect of the domain.

Although **content-related validity evidence** is especially important for achievement tests, it is also a concern for other types of measuring instruments, such as personality and aptitude measures.

Face validity is a term sometimes used in connection with a test's content. Face validity refers to the extent to which **examinees** believe the instrument is measuring what it is supposed to measure. Although it is not a technical form of validity, face validity can be important to ensure acceptance of the test and cooperation **on the part of the examinees**.

Evidence Based on Relations to a Criterion

Criterion-related validity evidence refers to the extent to which test scores are systematically related to one or more outcome criteria. The emphasis is on the criterion because one will use the test scores to infer performance on the criterion. Historically, two types of criterion-related validity evidence have been distinguished: concurrent and predictive.

Criterion validity (or criterion-related validity) measures how well one measure predicts an outcome for another measure. A test has this type of validity if it is useful for predicting performance or behavior in another situation (past, present, or future). For example:

- A job applicant takes a performance test during the interview process. If this test accurately predicts how well the employee will perform on the job, the test is said to have criterion validity.
- A graduate student takes the GRE. The GRE has been shown as an effective tool (i.e. it has criterion validity) for predicting how well a student will perform in graduate studies.

The first measure (in the above examples, the job performance test and the GRE) is sometimes called the predictor variable or the estimator. The second measure is called the criterion variable as long as the measure is known to be a valid tool for predicting outcomes.

Validity Coefficient

The coefficient of correlation between test scores and criterion is called a **validity coefficient** (*rxy*). Like any correlation coefficient, the size of a validity coefficient is influenced by the strength of the relationship between test and criterion and the range of individual differences in the group.

Validity coefficients indicate whether the test will be useful as a predictor or as a substitute measure. If it has been shown that a test has a high correlation with a future criterion, then that test can later be used to predict that criterion. Accumulating **predictive evidence** requires time and patience. In some cases, researchers must wait for several years to determine whether performance on a measure is useful for predicting success on a criterion. Concurrent criterion-related validity evidence is important in tests used for classification, certification, or diagnosis.

Construct-Related Evidence of Validity

Construct-related evidence of validity focuses on test scores as a measure of a psychological construct. To what extent do the **test scores** reflect the **theory** behind the psychological construct being measured? Recall that psychological constructs such as intelligence, motivation, anxiety, or critical thinking are **hypothetical qualities** or characteristics that have been "constructed" to account for observed behavior. They cannot be seen or touched or much less measured directly.

Validity

Historically, **validity** was defined as the extent to which an instrument measured what it claimed to measure. The focus of recent views of validity is not on the instrument itself but on the interpretation and meaning of the scores derived from the instrument.

Assessing the validity of **score-based interpretations** is important to the researcher because most instruments used in educational and psychological investigations are designed for measuring hypothetical constructs. Recall that constructs such as intelligence, creativity, anxiety, critical thinking, motivation, self-esteem, and attitudes represent abstract variables derived from theory or observation. Researchers have no direct means of measuring these constructs such as exist in the physical sciences for the measurement of characteristics such as length, volume, and weight. To measure these hypothetical constructs, you must move from the theoretical domain surrounding the construct to an empirical level that operationalizes the construct. That is, we use an operational definition to measure the construct. We do this by **selecting specific sets of observable tasks** believed to serve as indicators of the particular theoretical construct. Then we assume that performance (scores) on the tasks reflects the particular construct of interest as distinguished from other constructs. Essentially, validity deals with how well the operational definition fits with the conceptual definition.

Messick (1995) identified two problems that threaten the interpretation (validity) of test scores: construct underrepresentation and construct-irrelevant variance. The term **construct underrepresentation** refers to assessment that is too narrow and fails to include important dimensions of the construct. The **test may not adequately sample some kinds of content** or some types of responses or psychological processes and thus fails to adequately represent the theoretical domain of the construct.

The term **construct-irrelevant variance** refers to the extent to which test scores are affected by variables that are extraneous to the construct. Low scores should not occur because the test contains something irrelevant that interferes with people's demonstration of their competence. Construct-irrelevant variance could lower scores on a science achievement test for individuals with limited reading skills or limited English skills.

Validity of Criterion-Referenced Tests

Recall that **criterion-referenced tests** are designed to measure a rather narrow body of knowledge or skills. Thus, the main concern in assessing the validity of criterion-referenced tests is *content validity*. The basic approach to determining content validity is to have teachers or subject matter experts examine the test and judge whether it is an adequate sample of the content and objectives to be measured.

Application of the Validity Concept

Validity is always specific to the particular purpose for which the instrument is being used." No test is valid for all purposes or in all situations" (*Standards*, 1999, p. 17). Validity should be viewed as a **characteristic of the interpretation** and **use** of test scores and not of the test itself. A test that has validity in one situation and for one purpose may not be valid in a different situation or for a different purpose. A German proficiency test might be appropriate for placing undergraduates in German classes at a university but not be a valid exit exam for German majors. Thus, validation is always a responsibility of the test user as well as of the test developer.

We have viewed "test validation" as a process of gathering different types of evidence (content, criterion-related, and construct) in support of score-based interpretations and inferences. The goal of the process is to derive the best possible case for the inferences we want to make.

In **quantitative research**, there are three types of validity: **construct**, **external and internal** validity.

Construct validity deals with the degree to which the instruments used in the study measures the construct that is being examined. **External validity** deals with the extent to which the findings of a study **can be generalized** to a wider population. In **quantitative studies**, generalizability is often achieved by using **a random sample** of a representative group of the target population.

Internal validity deals with the degree to which the research deign is such that it has controlled **for variables that could influence the outcome** of the study. The extent to which the outcome of the research is due to the manipulation imposed by the research not other factors. In order to achieve internal validity the researcher tries to **control as many variables as possible**.

There are two kinds of reliability: internal and external reliability. **Internal reliability** deals with the extent to which someone else analyzing the same data would come up with the same results. Internal reliability can be judged through inter-rater reliability or intra-rater reliability. **External reliability** deals with whether or not another researcher, undertaking a similar study, would come to the same conclusions.

On a theoretical level, reliability is concerned with the **effect of error** on the consistency of scores. In this world measurement always involves some error. There are two kinds of errors: **random errors of measurement** and **systematic errors of measurement.** Random error is error that is a result of pure chance. Random errors of measurement may inflate or depress any subject's score in an unpredictable manner. Systematic errors, on the other hand, inflate or depress scores of identifiable groups in a predictable way. Systematic errors are the root of validity problems; random errors are the root of reliability problems.

Chapter 10: Experimental Research

An experiment has three characteristics: (1) An independent variable is manipulated; (2) all other variables that might affect the dependent variable are held constant; and (3) the effect of the manipulation of the independent variable on the dependent variable is observed. Thus, in an experiment the two variables of major interest are the independent variable and the dependent variable. The *independent variable* is manipulated (changed) by the experimenter. The variable on which the effects of the changes are observed is called the *dependent variable*, which is observed but not manipulated by the experimenter.

The essential requirements for experimental research are (a) **control**, (b) **manipulation** of the independent variable, and (c) **observation** and measurement.

Control

The purpose of control in an experiment is to arrange a situation in which the effect of a **manipulated variable** on a dependent variable can be investigated. The conditions for applying the law of the **single variable** are more likely to be fulfilled in the physical sciences than in the behavioral sciences.

Because educational research is concerned with human beings, many variables are always present. To attempt to reduce educational problems to the operation of a single variable is not only unrealistic but also perhaps even impossible. Fortunately, such rigorous control is not absolutely essential because many aspects in which situations differ are irrelevant to the purpose of the study and thus can be ignored. It is sufficient to apply the law of the single *significant* independent variable.

Although the law of the **single independent variable** cannot be followed absolutely, educational experimenters approximate it as closely as possible. Therefore, in experimental studies in education **you need procedures** that permit you to compare groups on the basis of significant variables. A number of methods of control have been devised to make such comparisons possible.

Example

Assume that you wish to test the hypothesis that children taught by the inductive method (group A) show greater gains in learning scientific concepts than children taught by the deductive method (group B). To draw a conclusion concerning the relationship between teaching method (independent variable) and the learning of scientific concepts (dependent variable), you must rule out the possibility that the outcome is due to some extraneous, usually unmeasured variable(s).

An **extraneous variable** is a variable that is not related to the purpose of the study but may affect the dependent variable. In this experiment, aptitude is a factor that certainly affects the

learning of scientific concepts; therefore, it would be considered a relevant extraneous variable that you must control. Otherwise, if the children in group A had more aptitude than those in group B, the greater gains in learning by group A could be attributed to aptitude and therefore you could not properly evaluate the effects of the teaching method on learning. Aptitude has **confounded** the relationship between the variables in which you are interested. The term **confounding** refers to the "mixing" of the variables extraneous to the research problem with the independent variable(s) in such a way that their effects cannot be separated. It could not be determined whether the relation found is (1) between the independent variable and the dependent variable of the study, (2) between the extraneous variables and the dependent variable, or (3) a combination of (1) and (2). **Eliminating confounding** by controlling for the effect of extraneous variables enables the experimenter to rule out other possible explanations of any observed changes. In the preceding experiment, the best way to control for aptitude is to **randomly assign** subjects to the two groups.

Manipulation

The **manipulation of an independent variable** is a deliberate operation performed by the experimenter. In educational research and other behavioral sciences, the manipulation of an independent variable involves setting up different *treatment* conditions. Treatment is another word for the experimental manipulation of the independent variable. The different treatment conditions administered to the subjects in the experiment are the *levels* of the independent variable.

Observation and Measurement

After applying the experimental treatment, the researcher observes to determine if the hypothesized change has occurred. Some changes can be observed directly, whereas other changes are measured indirectly. Learning, for example, is often the dependent variable in educational research. Researchers cannot measure learning directly. They can only **estimate learning through scores** on an achievement test or other measures chosen according to the operational definition. Therefore, strictly speaking, the dependent variable is observed scores rather than learning per se.

Experimental Design

The term **experimental design** refers to the conceptual framework within which the experiment is conducted. The experimental design sets up the conditions required for demonstrating cause-and-effect relationships.

An experimental design serves two functions: (1) It **establishes the conditions** for the comparisons required to test the hypotheses of the experiment, and (2) it enables the experimenter, through statistical analysis of the data, to **make a meaningful interpretation** of the results of the study.

The most important requirement is that the design *must be appropriate* for testing the previously stated hypotheses of the study. A second requirement is that the design must *provide adequate control* so that the effects of the independent variable can be evaluated as unambiguously as possible. **Randomization** is the single best way to achieve the necessary control. Therefore, the best advice is to select a design that uses randomization in as many aspects as possible.

Validity of Research Designs

Researchers must ask if the inferences drawn about the relationship between the variables of a study are valid or not. A very significant contribution to an understanding of the validity of experimental research designs was made by Campbell and Stanley (1963). They defined two general categories of validity of research designs: *internal validity* and *external validity*. Validity is not a property of an experimental design but, rather, refers to the **validity of the inferences**.

Four Types of Validity of Research Designs

Internal validity: The validity of the inferences about whether the effect of variable A (the treatment) on variable B (the outcome) reflects a causal relationship

Statistical conclusion validity: The validity of the inferences about the covariation between treatment and outcome

Construct validity: The validity of the inferences about psychological constructs involved in the subjects, settings, treatments, and observations used in the experiment

External validity: The validity of the inference about whether the cause–effect relationship holds up with other subjects, settings, and measurements

Internal Validity

Internal validity refers to the inferences about whether the changes observed in a dependent variable are, in fact, caused by the independent variable(s) in a particular research study rather than by some extraneous factors. Internal validity is concerned with such questions as Did the experimental treatment cause the observed change in the dependent variable or was some spurious (false) factor working? and Are the findings accurate?

Threats to Internal Validity

1. *History*. Specific events or conditions, other than the experimental treatment, may occur between the beginning of the treatment and the posttest measurement and may produce changes in the dependent variable. Such events are referred to as the **history effect.**

History doesn't refer to past events but to extraneous events occurring at the same time that the experimental treatment is being applied and that could produce the observed outcome even without any treatment.

2. *Maturation*. The term **maturation** refers to changes (biological or psychological) that may occur **within the subjects** simply as a function of the passage of time. These changes threaten internal validity because they may produce effects that could mistakenly be attributed to the experimental treatment. Subjects may perform differently on the dependent variable measure simply because they are older, wiser, hungrier, more fatigued, or less motivated than they were at the time of the first measurements. Maturation is especially a threat in **research on children** because they are naturally changing so quickly.
3. *Testing*. Taking a test once may affect the subjects' performance when the test is taken again, regardless of any treatment. This is called the **testing effect**. In designs using a pretest, subjects may do better on the posttest because they have learned subject matter from a pretest, have become familiar with the format of the test and the testing environment, have developed a strategy for doing well on the test, or are less anxious about the test the second time.

Pretest sensitization refers to the potential or actuality of a pretreatment assessment's effect on subjects in an experiment.

4. *Instrumentation*. The **instrumentation** threat to internal validity is a result of a change in the instruments used during the study.
5. *Statistical regression*. The term statistical regression refers to the well-known tendency for subjects **who score extremely high or extremely low on a pretest** to score closer to the mean (regression toward the mean) on a posttest. Statistical regression is a threat to internal validity when a subgroup is selected from a larger group on the basis of the subgroup's extreme scores (high or low) on a measure. When tested on subsequent measures, the subgroup will show a tendency to score less extremely on another measure, even a retest on the original measure. The subgroup will have a mean score closer to the mean of the original group.
6. *Selection bias*. Selection is a threat when there are important differences between the experimental and control groups even before the experiment begins. A selection bias is a **nonrandom factor** that might influence the selection of subjects into the experimental or the control group.

In a learning experiment, for example, if more capable students are in the experimental group than in the control group, the former would be expected to perform better on the dependent variable measure even without the experimental treatment. The best way to control selection bias is to use **random assignment of subjects to groups.** With random assignment, you cannot determine in advance who will be in each group; randomly assigned groups differ only by chance.

Selection bias is most likely to occur when the researcher cannot assign subjects randomly but must use **intact groups** (quasi-experiment). An intact group is a preexisting group such as a class or a group set up independently of the planned experiment.

Selection bias is also a threat when **volunteers** are used. People who volunteer for a study may differ in some important respects from non-volunteers.

7. *Experimental mortality (attrition)*. The **experimental mortality (attrition)** threat occurs when there is differential **loss of participants** from the comparison groups. This differential loss may result in differences on the outcome measure even in the absence of treatment.

Attrition is **not** usually a **serious threat** unless the study goes on for a long time or unless the treatment is so demanding that it results in low-performing participants dropping out.

8. **Selection – maturation Interaction**: Some of these threats may interact to affect internal validity. For example, selection and maturation may interact in such a way that the combination results in an effect on the dependent variable that is mistakenly attributed to the effect of the experimental treatment. Such interaction may occur in a **quasi-experimental design** in which the experimental and control groups are not randomly selected but instead are preexisting intact groups, such as classrooms. Although a pretest may indicate that the groups are equivalent at the beginning of the experiment, the experimental group may have a higher rate of maturation than the control group, and the increased rate of maturation accounts for the observed effect. If more rapidly maturing students are "selected" into the experimental group, the **selection–maturation interaction** may be mistaken for the effect of the experimental variable.
9. *Experimenter effect*. Experimenter effect refers to unintentional effects that the **researcher has on the study**. Personal characteristics of the researcher, such as gender, race, age, and position, can affect the performance of subjects.
10. *Subject effects*. Subjects' attitudes developed in response to the research situation called subject effects can be a threat to internal validity.

The tendency for subjects to change their behavior just because of the attention gained from participating in an experiment has subsequently been referred to as the **Hawthorne effect**. This effect can be a problem in educational research that compares exciting new teaching methods with conventional methods. Sometimes subjects may react to what they perceive to be the special ***demands*** of an experimental situation. That is, subjects react not as they normally might but as they think the more "**important**" researcher wants them to act. Research has shown, for instance, that subjects who know they are in an experiment tolerate more stress or administer more stress to others than they normally would.

The opposite of the Hawthorne effect is the **John Henry effect.*** This effect, also called **compensatory rivalry,** refers to the tendency of control group subjects who know they are in an

experiment to exert **extra effort** and hence to perform above their typical or expected average. They may perceive that they are in competition with the experimental group and they want to do just as well or better. Thus, the difference (or lack of difference) between the groups may be caused by the **control subjects' increased motivation** rather than by the experimental treatment. This effect is likely to occur in classroom research in which a new teaching technique is being compared to a conventional method that may be replaced by the new method. The students in the conventional classroom may want to show that they can do just as well as the students being taught by the new method.

Another subject effect, called **compensatory demoralization,** occurs when subjects believe they are receiving less desirable treatment or are being neglected. Consequently, they may become resentful or demoralized and put forth **less effort** than the members of the other group.

11. *Diffusion*. **Diffusion** occurs when participants in one group (typically the experimental group) communicate information about the treatment to subjects in the control group in such a way as to influence the latter's behavior on the dependent variable. Also, teachers involved with the experimental group may share information about methods and materials with teachers of the control group.

Threats to Internal Validity

History: Unrelated events that occur during the study affect the dependent variable.

Maturation Changes: occur within the participants just as a function of time.

Testing effect: Exposure to prior test affects posttest.

Instrumentation: Unreliability or a change in the measuring instrument affects result.

Regression: Extremely high or low scorers on a pretest regress toward mean on a posttest.

Selection bias: Because of selection methods, subjects in the comparison groups are not equivalent prior to study.

Mortality: A differential loss of participants from the groups affects dependent variable.

Selection–maturation Interaction: Subjects with different maturation rates are selected into treatment groups.

Experimenter effect: Unintentional bias or behavior of experimenter affects results.

Subject effect: Attitudes developed during the study affect performance on dependent variable.

Diffusion: Participants in experimental group communicate information about treatment to control group, which may affect the latter's performance.

Dealing with Threats to Internal Validity

Six basic procedures are commonly used to control inter-subject differences and increase equivalence among the groups that are to be exposed to the various experimental situations: (1) random assignment, (2) randomized matching, (3) homogeneous selection, (4) building variables into the design, (5) statistical control (analysis of covariance), and (6) use of subjects as their own controls.

Random Assignment (Randomization)

Randomization is the **most powerful method** of control because only chance would cause the groups to be unequal with respect to any potential extraneous variables.

Note that random assignment is **not** the same thing as random selection. Random selection is the use of a chance procedure to select a sample from a population. Random assignment is the use of a chance procedure to assign subjects to treatments.

When subjects have been randomly assigned to groups, the groups can be considered statistically equivalent. **Statistical equivalence** does not mean the groups are absolutely equal, but it does mean that any difference between the groups is a function of chance alone and not a function of experimenter bias, subject's choices, or any other factor. When random assignment has been employed, any pretreatment differences between groups are **nonsystematic**—that is, a function of chance alone.

Randomized Matching

When random assignment is not feasible, researchers sometimes select pairs of individuals with identical or almost identical characteristics and randomly assign one member of the matched pair to treatment A and the other to treatment B. This procedure is called **randomized matching.** Note that randomized matching requires that the subjects be matched on relevant variables first and then randomly assigned to treatments. The researcher first decides what variables to use for matching. The major limitation of matching is that it is almost impossible to find subjects who match on more than one variable. Subjects are lost to the experiment when no match can be found for them. This loss, of course, reduces the sample size and introduces **sampling bias** into the study.

Homogeneous Selection

Another method that can make groups reasonably comparable on an extraneous variable is to select samples that are as **homogeneous as possible on that variable**. This is called **homogeneous selection.** If the experimenter suspects that age is a variable that might affect the dependent variable, he or she would select only children of a particular age. By selecting only 6-year-old children, the experimenter would control for the effects of age as an extraneous independent variable.

Although homogeneous selection is an effective way of controlling extraneous variables, it has the disadvantage of decreasing the extent to which the findings can be generalized to other populations. If a researcher investigates the effectiveness of a particular method with such a homogeneous sample, such as children with average IQs, the results could not be generalized to children in other IQ ranges.

Building Variables into the Design

Some variables **associated with the subjects** (such as gender) can be built into the experimental design and thus controlled. For example, if you want to control gender in an experiment and you choose not to use the homogeneous selection technique just discussed, you could add gender as another independent variable. You would include both males and females in the study and then use **analysis of variance** to determine the effects of both gender and the main independent variable on the dependent variable. This method not only controls the extraneous gender variable but also yields information about its effect on the dependent variable, as well as its possible interaction with the other independent variable(s).

Statistical Control

Analysis of covariance (ANCOVA) is a statistical technique used to control for the effect of an extraneous variable known to be correlated with the dependent variable.

Using Subjects as Their Own Controls

Still another procedure involves **using subjects as their own controls**—assigning the same subjects to all experimental conditions and then obtaining measurements of the subjects, first under one experimental treatment and then under another.

Controlling Situational Differences

Extraneous variables may operate in the experimental setting to create **situational differences** that can also threaten internal validity. Three methods are commonly used to control potentially contaminating situational variables: (1) hold them constant, (2) randomize them, or (3) manipulate them systematically and separately from the main independent variable.

Single or Double-blind Experimental Procedures

The use of a placebo as just described illustrates what is called a **single-blind experiment.** The subjects are unaware of the treatment condition they are in, although the researcher knows. Sometimes, however, it is necessary to hold the attitudes of the researcher constant for different independent variable levels. This is done by using a **double-blind experimental** procedure in

which neither the experimenter nor the subjects know which kind of treatment the subjects are getting. In a double-blind situation, the experimenter must depend on other people to set up the groups, administer the treatment, and record results**.**

Another way to control extraneous situational variables is by manipulating them systematically. Many educational experiments must use a sequence of experimental and control conditions to control progressive effects, such as practice and fatigue effects.

Statistical Conclusion Validity

Statistical conclusion validity refers to the appropriate use of statistics to infer whether an observed relationship between the independent and dependent variables in a study is a **true cause–effect relationship** or whether it is just due to chance. Any inappropriate use of statistics is thus a threat because it may result in an erroneous conclusion about the effect of the independent variable on the dependent variable. Threats to statistical conclusion validity include using tests with low power, which may fail to detect a relationship between variables;

Construct Validity of Experiments

Specifically, construct validity of experiments is defined as the **validity of the inferences** made about a construct based on the measures, treatment, subjects, and settings used in an experimental study.

Threats on Construct validity

1. *Measure of the construct*. The measures used were not appropriate (poor operational definition), so the construct was not accurately measured.

2. *Manipulation of the construct*. The construct was not properly manipulated in the study; faulty manipulation may lead to incorrect inferences.

3. *Reactivity to the experimental situation*. Subjects' perceptions of the experimental situation become part of the treatment construct actually being tested. Recall the Hawthorne effect from the discussion of internal validity.

4. *Experimenter effect*. The experimenter can convey expectations about desirable responses, and those expectations become part of the treatment construct being studied.

Promoting Construct Validity

Shadish et al. (2002) suggested the following ways to improve construct validity of experiments: (1) Start with a clear explanation of the persons, setting, treatment, and outcome constructs of interest; (2) carefully select instances that match those constructs; (3) assess the match between instances and constructs to determine if any slippage between the two occurred; and (4) revise construct descriptions accordingly.

External validity

External validity refers to the extent to which the findings of a study can be generalized to other subjects, settings, and treatments.

Threats to External Validity

1. *Selection–treatment interaction* (non-representativeness). A major threat to external validity of experiments is the possibility of interaction between **subject characteristics and treatment** so that the results found for certain kinds of subjects may not hold for different subjects. This interaction occurs when the subjects in a study are not representative of the larger population to which one may want to generalize.

2. *Setting–treatment interaction* (**artificiality**). Artificiality in the setting may limit the generalizability of the results. The findings of a contrived lab study of motivation may not be the same as one would obtain in a study conducted in a public school setting.

3. *Pretest–treatment interaction*. Using a pretest may increase or decrease the experimental subjects' sensitivity or responsiveness to the experimental variable and thus **make** the **results** obtained for this pretested population **unrepresentative** of effects of the experimental variable on the unprotested population from which the experimental subjects are selected.

4. *Subject effects*. Attitudes and feelings of the participants that develop during a study may influence the generalizability of the findings to other settings. This threat is also called the **reactive effect** because subjects are reacting to the experience of participating in an experiment.

For example, **Hawthorne effect** as an internal validity problem can also be an external validity problem. Subjects' knowledge that they have been selected for an experiment and are being treated in a special way may affect the way they respond to the treatment.

Likewise, the **John Henry effect** may occur when subjects in the untreated control group are determined to do as well as or better than the subjects in the experimental group.

5. *Experimenter effects*. Another threat to external validity is the experimenter effect, which occurs when the experimenter consciously or unconsciously provides cues to subjects that influence their performance. For example, researcher's personality or characters, the presence of the researcher all could create experimenter effects.

Threats to External Validity

Selection–treatment interaction: An effect found with certain kinds of subjects might not apply if other kinds of subjects were used. Researcher should use a large, random sample of participants.

Setting–treatment interaction: An effect found in one kind of setting may not hold if other kinds of settings were used.

Pretest–treatment interaction: Pretest may sensitize subjects to treatment to produce an effect not generalizable to an unpretested population.

Subject effects: Subjects' attitudes developed during study may affect the generalizability of the results. Examples are the Hawthorne and the John Henry effects.

Experimenter effects: Characteristics unique to a specific experimenter may limit generalizability to situations with a different experimenter.

Experimentation is the most rigorous and the most desirable form of scientific inquiry.

Chapter 11: Experimental Research Design

An **experimental design** is the general plan for carrying out a study with an **active independent variable**. The design is important because it determines the **study's internal validity**, which is the ability to reach valid conclusions about the effect of the experimental treatment on the dependent variable. Designs differ in their efficiency and their demands in terms of time and resources, but the major difference is in how effectively they rule out **threats to internal validity**.

Experimental designs may be classified according to the number of independent variables: **single-variable designs** and **factorial designs**. A *single-variable design* has one manipulated independent variable; *factorial designs* have two or more independent variables, at least one of which is manipulated. Experimental designs may also be classified according to how well they provide control of the threats to internal validity: **pre-experimental**, **true experimental**, and **quasi-experimental designs**. *Pre-experimental designs* do **not** have random assignment of subjects to groups or other strategies to control extraneous variables. *True experimental designs* (also called randomized designs) use randomization and provide maximum control of extraneous variables. *Quasi-experimental designs* **lack randomization but employ other strategies** to provide some control over extraneous variables. They are used, for instance, when intact classrooms are used as the experimental and control groups. Thus, true experimental designs have the **greatest** internal validity, quasi-experimental designs have somewhat **less** internal validity, and the pre-experimental designs have **the least** internal validity.

Pre-experimental Designs

Pre-experimental design provides little or no control of extraneous variables. We include these weak designs in our discussion simply because they illustrate quite well the way that **extraneous** variables may operate to **jeopardize the internal validity** of a design.

Design 1: One-Group Pretest–Posttest Design

The **one-group pretest–posttest design** usually involves three steps: (1) administering a pretest measuring the dependent variable; (2) applying the experimental treatment *X* to the subjects; and (3) administering a posttest, again measuring the dependent variable. There is **no** control group! There is no way to assess the effect of the pretest.

The best advice is to avoid using Design 1. Without a control group to make a comparison possible, the results obtained in a one-group design are basically uninterpretable.

Design 2: Static Group Comparison

The **static group comparison** uses two or more preexisting or intact (static) groups, only one of which is exposed to the experimental treatment. Although this design uses two groups for

comparison, it is flawed because the **subjects are not randomly assigned** to the groups and no pretest is used. The researcher makes the assumption that the groups are equivalent in all relevant aspects before the study begins and that they differ only in their exposure to *X*.

True Experimental Designs

The designs in this category are called *true experiments* because subjects are randomly assigned to groups. Because of the control they provide, they are the most highly recommended designs for experimentation in education.

Design 3: Randomized Subjects, Posttest-Only Control Group Design

Randomized subjects, posttest-only control group design is one of the simplest yet one of the most powerful of all experimental designs. It has the two essential elements necessary for maximum control of the threats to internal validity: **randomization** and a **control group**. No pretest is used;

After the subjects are randomly assigned to groups, only the experimental group is exposed to the treatment. In all other respects, the two groups are treated alike. Members of both groups are then measured on the dependent variable Because of the lack of a pretest, **mortality** could be a threat. Without having pretest information the researcher has no way of knowing if those who dropped out of the study were different from those who continued.

Design 4: Randomized Matched Subjects, Posttest-Only Control Group Design

Randomized matched subjects, posttest-only control group design is similar to Design 3, except that it uses a matching technique to form equivalent groups. Subjects are matched on one or more variables that can be measured conveniently, such as IQ or reading score. The flip of a coin can be used to assign one member of each pair to the treatment group and the other to the control group.

Matching is most useful in studies in which **small samples** are to be used and where Design 3 is not appropriate. Design 3 depends completely on random assignment to obtain equivalent groups. With small samples the influence of chance alone may result in a situation in which random groups are initially very different from each other. Design 3 provides no assurance that small groups are really comparable before the treatments are applied. The matched-subjects design controls preexisting intersubject differences on variables highly related to the dependent variable that the experiment is designed to affect. The random procedure used to assign the matched pairs to groups adds to the strength of this design.

If one or more subjects were excluded because an appropriate match could not be found, this would bias the sample. When using Design 4, it is essential to match every subject, even if only approximately, before random assignment.

Design 5: Randomized Subjects, Pretest–Posttest Control Group Design

Design 5 is one of the most widely used true (randomized) experiments. In the **randomized subjects, pretest–posttest control group design,** one randomly assigns subjects to the experimental and control groups and administers a pretest on the dependent variable *Y*. The treatment is introduced only to the experimental subjects. The recommended statistical procedure to use with Design 5 is an analysis of covariance (ANCOVA) with posttest scores as the dependent variable and pretest scores as the covariate to control for initial differences on the pretest. The main strength of this design is the **initial randomization**, which ensures statistical equivalence between the groups prior to experimentation.

Design 5 thus controls most of the extraneous variables that pose a threat to internal validity. For example, the effects of **history** and **maturation** are experienced in both groups; therefore, any difference between the groups on the posttest measure could probably not be attributed to these factors. **Differential selection** of subjects and **statistical regression** are also controlled through the randomization procedure. There is one internal validity issue, however. Although both E and C groups take the pretest and may experience the sensitizing effect, the pretest can cause the experimental subjects to respond to the *X* treatment in a particular way just because of their increased **sensitivity**. The result is a difference on the posttest that could mistakenly be attributed to the effect of the treatment alone.

The main concern in using Design 5 is **external validity**. Ironically, the problem stems from the use of the pretest, an essential feature of the design. As mentioned previously, there may be an **interaction** between the **pretest** and the **treatment** so that the results are generalizable only to other pretested groups. The responses to the posttest may not be representative of how individuals would respond if they had not been given a pretest.

Design 6: Solomon Three-Group Design

The first of the Solomon designs uses three groups, with random assignment of subjects to groups. Note that the first two lines of this design are identical to Design 5. However, the **Solomon three-group design** has the advantage of employing a second control group labeled C2 that is ***not*** **pretested** but is exposed to the treatment *X*. This group, despite receiving the experimental treatment, is functioning as a control and is thus labeled control group.

This design overcomes the difficulty inherent in Design 5—namely, the interactive effect of pretesting and the experimental treatment. The posttest scores for the three groups are compared to assess the interaction effect.

Design 7: Solomon Four-Group Design

The Solomon four-group design provides still more rigorous control by extending Design 6 to include one more control group that receives neither pretest nor treatment.

Factorial Design

The designs presented thus far have been the classical single-variable designs in which the experimenter manipulates **one independent variable** *X* to determine its effect on a dependent variable *Y*. However, in complex social phenomena several variables often interact simultaneously, and restricting a study to one independent variable may impose an artificial simplicity on a complex situation.

A **factorial design** is one in which the researcher manipulates two or more variables simultaneously in order to study the independent effect of each variable on the dependent variable, as well as the effects caused by interactions among the several variables.

The **independent variables** in factorial designs are referred to as ***factors***. Factors might be categorical variables such as gender, ethnicity, social class, and type of school, or they might be continuous variables such as aptitude or achievement.

Design 8: Simple Factorial Design

Factorial designs have been developed at varying levels of complexity. The simplest factorial design is the 2 × 2, which is read as "2 by 2." This design has two factors, and each factor has two levels.

Other Randomized Experimental Designs

The experimental designs we have discussed so far use at least two groups of subjects, one of which is exposed to the treatment (independent variable) and the other that does not receive the treatment or is exposed to another level of the treatment. The researcher then compares the dependent variable scores for the different treatment groups. The essential feature of these designs is that they compare **separate groups of subjects** in order to determine the effect of the treatment. When the independent variable is manipulated in this way, we have what is called a **between-subjects design.** For example, a researcher who compares reading achievement scores for students taught by one method with scores for an equivalent group of students taught by a different method is using a between subjects design.

However, the manipulation of an independent variable does not have to involve different groups of subjects. It is possible to use experimental designs in which the same participants are exposed to **different levels of the independent variable** at different times.

For example, a researcher might measure the learning of nonsense syllables by one group of students under different levels of anxiety or the math performance scores of a group of students when music is played in the classroom versus no music. This type of design in which a researcher observes each individual in all of the different treatments is called a **within- subjects design.**

It is also called a **repeated-measures design** because the research repeats measurements of the same individuals under different treatment conditions. The main advantage of a within-subjects design is that it **eliminates the problem of differences in the groups** that can confound the findings in between subjects research. Remember that one is not comparing one group of subjects to another; one is comparing each individual's score under one treatment with the same individual's score under another treatment. Each subject serves as his or her own control.

Another advantage of within-subjects designs is that they **can be conducted with fewer subjects.** The disadvantage of these designs is the **carryover effect** that may occur from one treatment to another. To deal with this problem, researchers typically arrange for the participants to experience the different treatments in random or counterbalanced order.

Quasi-Experimental Designs

Quasi-experimental designs are similar to randomized experimental designs in that they involve manipulation of an independent variable but differ in that subjects are not randomly assigned to treatment groups. Because the quasi-experimental design does not provide full control, it is extremely important that researchers be aware of the threats to both internal and external validity and consider these factors in their interpretation.

Design 9: Nonrandomized Control Group, Pretest–Posttest Design

The **nonrandomized control group, pretest–posttest design** is one of the most widely used quasi-experimental designs in educational research. You can see that it is similar to **Design 5** but with one important difference: Design 9 does *not* permit random assignment of subjects to the experimental and control groups.

Without random assignment of subjects, you do not know if the groups were equivalent before the study began. Perhaps the class designated the experimental group would have done better on the posttest without the experimental treatment. Thus, there is an initial *selection bias* that can seriously threaten the internal validity of this design. The pretest, the design's most important feature, provides a way to deal with this threat. The **pretest enables you to check on the equivalence of the groups on the dependent variable before the experiment begins.** If there are no significant differences on the pretest, you can discount selection bias as a serious threat to internal validity and proceed with the study. If there are some differences, the investigator can use ANCOVA to statistically adjust the posttest scores for the pretest differences.

Because both experimental and control groups take the same pretest and posttest, and the study occupies the same period of time, other threats to internal validity, such as **maturation**, **instrumentation**, **pretesting**, **history**, and **regression**) should not be serious threats to internal validity. Having the same person teach both English classes would be recommended.

Design 10: Counterbalanced Design

A counterbalanced design, another design that can be used with **intact class groups**, rotates (alternates) the groups at intervals during the experimentation. For example, groups 1 and 2 might use methods A and B, respectively, for the first half of the experiment and then exchange methods for the second half. The distinctive feature of Design 10 is that all groups receive all experimental treatments but in a different order. In effect, **this design involves a series of replications;** in each replication the groups are shifted so that at the end of the experiment each group has been exposed to each *X*. The order of exposure to the experimental situation differs for each group.

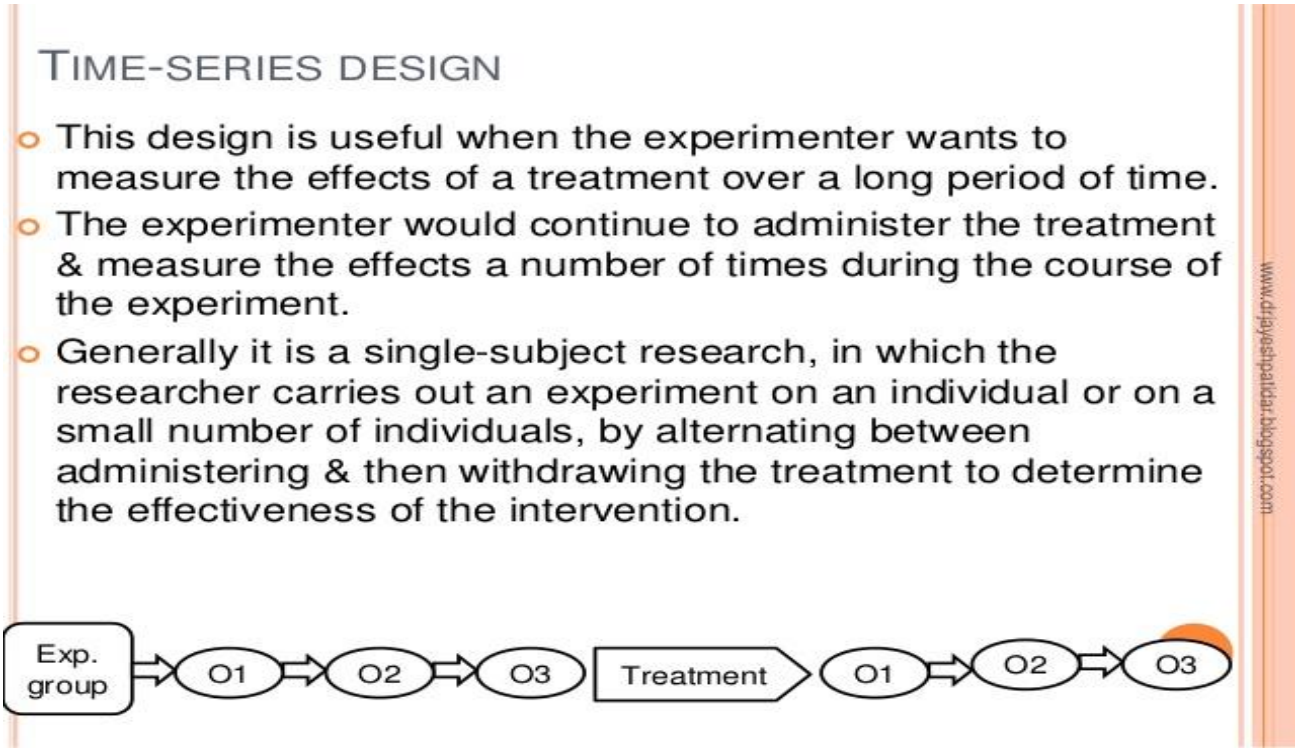
Design 10 overcomes some of the weaknesses of Design 9; that is, when intact classes must be used, **counterbalancing** provides an opportunity to rotate out any differences that might exist between the groups. Because all treatments are administered to all groups, the results obtained for each *X* cannot be attributed to preexisting differences in the subjects.

The main shortcoming of Design 10 is that there may be a **carryover effect** from one *X* to the next. Therefore, this design should be used only when the experimental treatments are such that exposure to one treatment will have no effect on subsequent treatments. Another weakness of the counterbalanced design is the possibility of boring students with the repeated testings this method requires.

What are Time Series Research Designs?

The defining feature of time series research designs is that each participant or sample is observed multiple times, and its performance is compared to its own prior performance. In other words, each participant or population serves as its own control. The outcome is measured repeatedly for the same subject or population during one or more baseline and treatment conditions.

As the design indicates, a number of measurements on a dependent variable are taken, X is introduced, and additional measurements of *Y* are made. By comparing the measurements before and after, you can assess the effect of *X* on the performance of the group on *Y*.



Design 11 is similar to Design 1 in that it uses before-and-after measures and lacks a control group. However, it has certain advantages over Design 1 that make it more useful in educational research. The repeated testing provides a check on some common threats to internal validity. The major weakness of Design 11 is its **failure to control history**; that is, you cannot rule out the possibility that it is not *X* but, rather, some simultaneous event that produces the observed change. The extent to which history (uncontrolled contemporary events) is a plausible explanatory factor must be taken into account by the experimenters as they attempt to interpret their findings. You must also consider the **external validity** of the time design. Because there are repeated tests, perhaps there is a kind of **interaction effect of testing** that would restrict the findings to those populations subject to repeated testing.

Design 12: Control Group Time-Series Design

The **control group time-series design** is an extension of Design 11 to include a control group. The control group, again representing an intact class, would be measured at the same time as the experimental group but would not experience the *X* treatment. This design overcomes the weakness of Design 11, that is, **failure to control history** as a source of extraneous variance. The control group permits the necessary comparison.

Single-Subject Experimental Designs

The single-subject experimental designs are a type of experimental design with a unique feature: the sample size is just one or is composed of a few participants who are treated as one unit. Obviously, there can be **no random assignment** or use of **control groups**. In single-subject

experimental designs (also called single-case experimental designs), the participant serves as both the treatment and the control group. The researcher measures participant behavior repeatedly. The periods during which the treatment is given are called *treatment periods*, and the periods during which the treatment is not present are called *baseline periods*.

The data for the baseline period would serve as the control group data and would be compared with the data during obtained the treatment and after the treatment period. **Single-subject research** has become popular during the past 30 years as proponents of this particular methodology have demonstrated that experimental control can be effectively achieved in other than the traditional ways. Study of the individual has always had a place in educational and psychological research. Freud's case studies and Piaget's observations of individual children are notable examples.

Single-Subject Research versus Case Study

Although case studies and single-subject experiments both study the individual, in a single-subject experiment, the investigator deliberately **manipulates** one or more independent variables, whereas in a case study the observer studies the subjects' interaction with events that occur naturally. Single-case designs have been particularly useful in clinical applications in which the focus is on the therapeutic value of an intervention for the client.

Chapter 12: Ex Post Facto Research

Active vs. Attribute Variables

An **active independent variable** is one that is designed, imposed, controlled by the investigators. This is the highest level of independent variables, met by true experimental studies. It has the advantage of having a consistent intervention. An **attribute independent variable** occurs when groups are compared, but the grouping variable cannot be chosen and manipulated by the investigators because it is a characteristic of the subjects themselves.

Ex post facto research is conducted after **variation** in the variable of interest has **already** been **determined** in the natural course of events. This method is sometimes called *causal comparative* because its purpose is to investigate cause-and-effect relationships between independent and dependent variables. Researchers use it in situations that do not permit the randomization and manipulation of variables characteristic of experimental research. Thus, much of the basic rationale for experimental and ex post facto is the same. They both investigate relationships among variables and test hypotheses.

The effects of extraneous variables in an experiment are controlled by the experimental conditions, and the antecedent independent variable is directly manipulated to assess its effect on the dependent variable.

Ex post facto research, unlike experimental research, does not provide the safeguards that are necessary for making strong inferences about causal relationships. Mistakenly attributing causation based on a relationship between two variables is called the **post hoc fallacy.** An investigator who finds a relationship between the variables in an ex post facto study has secured evidence only of some concomitant variation.

Post-hoc Analysis

In the design and analysis of experiments, **post hoc analysis** consists of looking at the data—after the experiment has concluded—for patterns that were not specified *a priori*.

In practice, post hoc analyses are usually concerned with finding patterns and/or relationships between subgroups of sampled populations that would otherwise remain undetected and undiscovered were a scientific community to rely strictly upon ***prior statistical*** methods. Post hoc tests—also known as ***a posteriori*** tests. Because the investigator has not controlled *X* or other possible variables that may have determined *Y*, there is **less basis for inferring** a causal relationship between *X* and *Y*.

Spurious Relationship

A **spurious relationship** is one in which the two variables really have no effect on each other but are related because some other variable influences both.

In statistics, a **spurious relationship** or **spurious correlation** is a mathematical relationship in which two or more events or variables are not causally related to each other, yet it may be wrongly inferred that they are, due to either coincidence or the presence of a certain third, unseen factor (referred to as a "common response variable", "confounding factor", or "lurking variable").

Planning an Ex Post Facto Research Study

1. *The first step in an ex post facto study is to state the research problem, usually in the form of a question.*

2. *Next, select two or more groups to be compared.*

Recall that investigators doing ex post facto research achieve the variation they want not by directly manipulating the variable but by selecting individuals in whom the variable is present or absent, strong or weak. Thus, these two groups should differ on the variable of interest in that one group should possess the characteristic and the other group should not, but they should be similar on any relevant extraneous variables. **Differential (subject) selections** pose a major threat to the internal validity of ex post facto investigations because you have no control over the selection of subjects into the two groups. They are selected because they already possess the variable of interest, for example, smoker/nonsmoker and retained/not retained. Whenever assignment is not random, there is always an opening for other variables to enter to explain the observed difference between the groups. The way to deal with this threat is to collect data to show that the groups are similar on other extraneous variables that might affect the variable of interest.

For example, if you were studying the effect of preschool attendance on the social maturity of kindergarteners, you would have to control any other factors that might have been shown to influence social maturity. Some of these might be age, gender, socioeconomic status, and aptitude. You use logic and previous research to determine what factors need to be controlled in an ex post facto study.

3. *Determine whether your question requires a proactive or a retroactive design.*

Alternative Explanations in Ex Post Facto Research

When investigators can control the treatment (*X*) and then observe the dependent variable (*Y)* as in experimental research, they have reasonable evidence that *X* influences *Y*. Ex post facto research, on the other hand, lacks control of the independent variable and thus has **lower internal validity**. If researchers cannot control (*X*), they may be led to inappropriate conclusions.

When interpreting ex post facto research, one should consider alternative explanations, such as common cause, reverse causality, and the presence of other independent variables.

Common Cause

In an ex post facto investigation, you must consider the possibility that both the independent variable and the dependent variable of the study are merely two separate results of a third variable—that they have a **common cause.** For example, if you use a school's total budget as an independent variable and cases of diagnosed learning disability as a dependent variable, you might find a positive correlation between the two variables. Does this mean that an increase in total school budget leads to an increase in cases of learning disability? A more plausible explanation is that the **relationship is spurious.** An increase in school size/number of children attending could account for both the budget and the cases of diagnosed learning disability because funding is tied to the number of students. It is well established that the average income of private high school graduates is much higher than the average income of public and parochial high school graduates. Does this mean that private schools better prepare students for financial success?

When doing ex post facto research, you must always consider the possibilities of common cause or causes accounting for an observed relationship.

Reverse Casualty

In interpreting an observed relationship in an ex post facto study, the researcher must consider the possibility of **reverse causality**—that the reverse of the suggested hypothesis could also account for the finding. Instead of saying that *X* causes *Y*, perhaps it is the case that *Y* causes *X*.

The hypothesis of reverse causality is easier to deal with than the hypothesis of common cause. With the latter, numerous **common causes** in each case could **produce a spurious relationship**. With reverse causality, there is only one possibility in each case: *Y* caused *X*, instead of *X* caused *Y*.

Other Possible Independent Variables

Independent variables other than the one considered in the ex post facto study may bring about the observed effect on the *Y* variable; that is, in addition to *X*1, other variables, *X*2 and *X*3, may also be antecedent factors for the variation in the dependent variable.

An obvious first task for investigators is to attempt to **list all the possible alternative independent variables.** Then by holding the others constant, you can test in turn each variable to determine if it is related to *Y*. If you can eliminate the alternative independent variables by showing that they are not related to *Y*, you gain support for the original hypothesis of a relationship between *X* and *Y*.

Partial Control in Ex Post Facto Research

There are strategies for improving the credibility of ex post facto research, although none can adequately compensate for the inherent weakness of such research—namely, lack of control over the independent variable. These strategies provide **partial control** of the internal validity problems of common cause and other possible independent variables. Among these strategies are matching, homogeneous groups, building extraneous variables into the design, analysis of covariance, and partial correlation.

Purpose of Ex Post Facto Design (Online Search)

The main purpose of using an ex post facto is to determine the cause and effect relationship between the dependent and the independent variable. If for instance a researcher wishes to do a research on the effects of anxiety on student's performance, he will measure and use the anxiety levels of the students and later compare them with the overall performance once the results are released (Ary, & Sorensen, 2009). The variable in such a research are a "matter of fact" and cannot be manipulated so as to determine whether they will be variation of results. The fact that the independent variable in this type of research cannot be manipulated put the research design as not credible.

Strategies Used In Ex Post Facto Research

The first strategy involves **matching**. Matching is done on a subject to subject basis so as to create matched pairs. The matching criteria will be based on criteria that enables the researcher to class similar subjects together.

The second strategy involves the use of **homogenous groups**. Selecting subjects that are homogenous gives the researcher some degree of control over the variable and the ability to get the desired results. If a researcher needs to determine the academic performance of female students between 13-15 years, then he can specifically look for female students in that age bracket.

The third strategy involves **building extra venous variables** in the design. This involves selecting subject who specifically fit the type of data that the researcher needs. If for instance the researcher identifies social economic status as a potential variable he will select to work with a sample that involve student from low socio economic status.

Analysis of covariance is another strategy that can be used (Ary, & Sorensen, 2009). It is technique that is used to equate (compare) groups depending on specified variable. However, because the adjustment is only partial, ANCOVA does **not** "solve" the problem of initial differences between groups but only reduces it. When interpreting ex post facto research, it is inappropriate to assume ANCOVA has satisfactorily adjusted for initial differences

Types of ex post Facto research design

There are two main types of ex post facto research design. The first is the **proactive** design. The subjects of this designed are grouped depending on an already pre existing independent variable for example subject studies on scholarship and subject privately pays for his or her studies. Once the pre existing groups are determined, measurements are made based on dependent variable such as level of confidence, class performance etc. The second category is the **retroactive** ex post fact where the researcher seeks that causes of a pre existing situation. The cause in this case will be the independent variable whereas the pre-existing situation is the dependent variable.

Importance of Ex post facto research

If properly implemented ex post facto research can help researchers to clear the air on some research studies whose findings are inconclusive. This is especially so for results gathered through experimental studies. By using the suggested **partial control strategies** and determining **alternative hypothesis** a researcher who had first used the experimental research may find more conclusive findings if he uses ex post facto research (Ary, & Sorensen, 2009).

Ex post facto research is also important when conducting research studies that involve human beings as the subjects. Many researches involving human subjects are prone to **ethical issues.** This is especially so if the research involves some form of manipulation that will negatively affect the participants. In such a situation the use of ex post facto research will focus on cause and effect relations rather than intrusive manipulations that may cross the ethical boundaries of using human participants as subjects (Ary, & Sorensen, 2009).

Ex post facto research is important in **verifying already made conclusions** on various researches. This form of research design can therefore be used to authenticate the conclusions sand findings made in an in other experimental studies. If the findings coincide then the findings are deemed to be true. However, if the findings contradict, then it create room for more research studies so as to finding the ultimate truths.

Ex post facto research in education has permitted investigations of the effects of variables such as home background, father absence, early experiences, disabilities, teacher competence, and others that are **beyond the control of educators**. In some instances, ex post facto research has discovered relationships or raised questions that can later be investigated more systematically in well-controlled experimental studies. Appropriately used and cautiously interpreted, ex post facto research will continue to provide a valuable methodology for the acquisition of knowledge.

Although there are many disadvantages of ex post facto design, it nevertheless is frequently the only method by which educational researchers can obtain necessary information about **characteristics of defined groups of students** or information needed for the intelligent **formulation of programs** in the school.

It permits researchers to investigate situations in which **controlled variation is impossible** to introduce. Attributes such as academic aptitude, creativity, self-esteem, socioeconomic status, and teacher personality cannot be manipulated and hence must be investigated through ex post facto research rather than through the more rigorous experimental approach.

The possibility of **spurious relationships** is always present in ex post facto research. Considering the possibilities of common cause, reversed causality, and possible alternate independent variables can help educators evaluate such research more **realistically**. Several **partial control strategies** can help researchers avoid **gross (obvious) errors** in ex post facto designs, but **none** can entirely **solve** the **problems** inherent in those designs. Always exercise caution when interpreting ex post facto results.

Chapter 13: Correlational Research

Correlational research is nonexperimental research that is similar to ex post facto research in that they both employ data derived from preexisting variables. There is no manipulation of the variables in either type of research. They differ in that in ex post facto research, selected variables are used to **make comparisons** between two or more existing groups, whereas correlational research **assesses the relationships** among two or more variables in a single group. Ex post facto research investigates possible cause-**and-effect relationships**; correlational research typically does not. An advantage of correlational research is that it provides information about the **strength of relationships** between variables.

Correlational research produces indexes that show both the **direction** and the **strength** of **relationships** among variables, taking into account the entire range of these variables. This index is called a **correlation coefficient**.

Uses of Correlational Research

Correlational research is useful in a wide variety of studies. The most useful applications of correlation are (1) assessing relationships, (2) assessing consistency, and (3) prediction.

In Chapter 9, we noted that the reliability (consistency) of a test can be assessed through correlating test–retest, equivalent-forms, or split-half scores. Correlation can be used to measure consistency (or lack thereof) in a wide variety of cases.

If you find that two variables are correlated, then you can use one variable to predict the other. The higher the correlation, the more accurate the prediction. Prediction studies are frequently used in education.

Design of Correlational Studies

The basic design for correlational research is **straightforward**. First, the researcher specifies the problem by **asking a question** about the relationship between the variables of interest. The variables selected for investigation are generally based on a theory, previous research, or the researcher's observations. Because of the potential for spurious results, we do not recommend the **"shotgun" approach** in which one correlates a number of variables just to see what might show up. The population of interest is also identified at this time. In **simple correlational studies**, the researcher focuses on gathering data on two (or more) measures from a single group of subjects. For example, you might correlate vocabulary and reading comprehension scores for a group of middle school students. Occasionally, correlational studies investigate relationships between scores on one measure for logically paired groups such as twins, siblings, or husbands and wives. For instance, a researcher might want to study the correlation between the SAT scores of identical twins.

It is important to select or develop measures that are appropriate indicators of the constructs to be investigated, and that it is especially important that these instruments have **satisfactory reliability and are valid for measuring the constructs** under consideration. In correlation research, the **size of a coefficient of correlation** is influenced by the **adequacy of the measuring instruments** for their intended purpose. Instruments that are too easy or too difficult for the participants in a study would not discriminate among them and would result in a smaller correlation coefficient than instruments with appropriate difficulty levels.

Following the selection or development of instruments, the researcher specifies his or her population of interest and draws a **random sample** from that population. Finally, the researcher **collects the quantitative data** on the two or more variables for each of the students in the sample and then **calculates the coefficient(s) of correlation** between the paired scores. Before calculating the coefficient, the researcher should **look at a scatterplot or a graph** of the relationship between the variables.

Spearman's Rank Correlation Coefficient

In statistics, **Spearman's rank correlation coefficient** is a nonparametric measure of **rank correlation** (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function.

The sign of the Spearman correlation indicates the direction of association between *X* (the independent variable) and *Y* (the dependent variable). If *Y* tends to increase when *X* increases, the Spearman correlation coefficient is positive. If *Y* tends to decrease when *X* increases, the Spearman correlation coefficient is negative. A Spearman correlation of zero indicates that there is no tendency for *Y* to either increase or decrease when *X* increases.

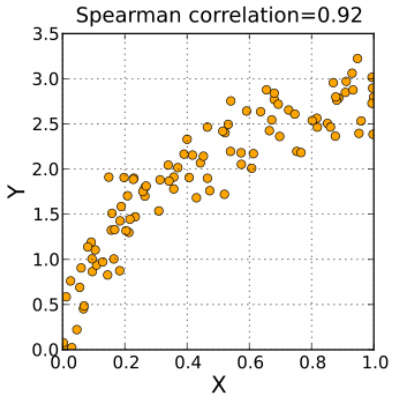


Figure. A positive Spearman correlation coefficient corresponds to an increasing monotonic trend between *X* and *Y*

Phi Coefficient

In statistics, the **phi coefficient** (or **mean square contingency coefficient**) is a measure of association for two binary variables. Introduced by Karl Pearson,[1] this measure is similar to the Pearson correlation coefficient in its interpretation.

Comparison to Other Correlations

Practical Utility

Always consider the practical significance of the correlation coefficient. Although a correlation coefficient may be statistically significant, it may have little practical utility. With a sample of 1000, a very small coefficient such as .08 would be statistically significant at the .01 level. But of what practical importance would this correlation be?

Failure to find a statistically significant relationship between two variables in one study does not necessarily mean there is *no* relationship between the variables. It only means that in that particular study, **sufficient evidence** for a relationship was not found. Recall from Chapter 6 that other factors, such as reliability of the measures used and range of possible values on the measures, influence the size of a correlation coefficient.

Statistical Significance

In evaluating the size of a correlation, it is important to consider the **size of the sample** on which the correlation is based. Without knowing the sample size, you do not know if the correlation could easily have occurred merely as a result of **chance** or is likely to be an indication of a **genuine relationship**. If there were fewer than 20 cases in the sample (which we would not recommend), then a "high" *r* of .50 could easily occur by chance. You should be very careful in attaching too much importance to large correlations when small sample sizes are involved; an *r* found in a small sample does not necessarily mean that a correlation exists in the population.

To avoid the error of inferring a relationship in the population that does not really exist, the researcher should state the null hypothesis that the population correlation equals 0 ($H0$: $\rho xy = 0$) and then determine whether the obtained sample correlation departs sufficiently from 0 to justify the rejection of the null hypothesis.

Determining Sample Size

The Pearson product moment correlation is a form of effect size. Therefore, Table A.3 in the Appendix can be used to determine the needed sample size for a predetermined level of significance and predetermined tolerable probability of Type I error.

Correlation and Causation

In evaluating a correlational study, one of the most frequent errors is to interpret a correlation as indicating a **cause-and-effect relationship**. Correlation is a necessary but never a sufficient condition for causation.

Partial Correlation

Partial correlation is a technique used to determine what correlation remains between two variables when the effect of another variable is eliminated. We know that correlation between two variables may occur because both of them are correlated with a third variable. Partial correlation controls for this third variable.

Partial correlation is a measure of the strength and direction of a linear relationship between two continuous variables whilst controlling for the effect of one or more other continuous variables (also known as 'covariates' or 'control' variables). Although partial correlation does not make the distinction between independent and dependent variables, the two variables are often considered in such a manner (i.e., you have one continuous dependent variable and one continuous independent variable, as well as one or more continuous control variables).

Multiple Regression

Multiple regression is a correlational procedure that examines the relationships among several variables. Specifically, this technique enables researchers to find the best possible weighting of two or more independent variables to yield a maximum correlation with a single dependent variable.

Match the procedure listed in the left column with the defi nition in the right column:

1. Spearman rho	a. Shows sign and magnitude of correlation between two nominal variables
2. Pearson *r*	b. Shows sign and magnitude of correlation between two ordinal variables
3. Multiple regression	c. Shows sign and magnitude of correlation between two interval variables
4. Phi coefficient	d. Uses a number of independent variables to predict a single dependent variable
5**. Eta correlation**	e. Used when the relationship between two variables is curvilinear
Answers	1. b; 2. c; 3. d; 4. a; 5. E

Factor Analysis

Factor analysis, or exploratory factor analysis, is a family of techniques used to detect patterns in a set of interval-level variables (Spicer, 2005). The purpose of the analysis is to try to reduce the **set of measured variables to a smaller set of underlying factors** that account for the pattern of relationships. The search follows the law of parsimony, which means that the data should be accounted for with the smallest number of factors. This reduction of the number of variables serves to make the data more manageable and interpretable.

There are two types of situations in which factor analysis is typically used. In the first, a researcher is interested in reducing a set of variables to a smaller set. The second type of situation is when researchers use factor analysis to determine the characteristics or underlying structure of a measuring instrument such as a measure of intelligence, personality, or attitudes.

Factor analysis is a technique that is used to reduce a large number of variables into fewer numbers of factors. Factor analysis assumes several assumptions: there is **linear relationship**, there is no **multicollinearity**, it includes relevant variables into analysis, and there is **true correlation** between variables and factors.

Confirmatory Factor Analysis

Confirmatory factor analysis, like exploratory factor analysis, is used to examine the relationships between a set of measured variables and a smaller set of factors that might account for them. Confirmatory factor analysis, however, assumes relatively precise advance knowledge and allows a researcher to specify a priori what these relationships might look like and then to test the accuracy of these formal hypotheses.

Other Complex Correlational Procedures

Several more complex techniques are available to investigate correlation of more than two variables.

Canonical correlation is a generalization of multiple regression that adds more than one dependent variable (criterion) to the prediction equation.

Discriminant analysis is a statistical procedure related to multiple regression, but it differs in that the criterion is a *categorical* variable rather than a *continuous* one.

Structural equation modeling (SEM) is a popular technique used in the analysis of causality. SEM combines confirmatory factor analysis and **path analysis** to test both a measurement model and a structural model.

Pedhazur (2006) defines **path analysis** as a method for studying direct and indirect effects of variables hypothesized as causes of variables treated as effects.

Chapter 14: Survey Research

In **survey research,** investigators ask questions about peoples' beliefs, opinions, characteristics, and behavior.

Types of Surveys

Before initiating survey research, the investigator must determine the **format** that is most appropriate for the proposed investigation. Surveys are classified according to their focus and **scope** (census and sample surveys) or according to the time **frame** for data collection (longitudinal and cross-sectional surveys).

Surveys Classified According to Focus and Scope

A survey that covers the entire population of interest is referred to as a **census,** an example of which is the U.S. Census, undertaken by the government every 10 years. In research, the term **population** is used to refer to the entire group of individuals to whom the findings of a study apply. The researcher defines the specific population of interest. It is often difficult or even impossible for researchers to study very large populations. Hence, they select a smaller portion, a **sample,** of the population for study. A survey that studies only a portion of the population is known as a **sample survey.**

Surveys may be confined to simple tabulations of **tangibles,** such as what proportion of children rides school buses and the average class enrollment. The most challenging type of survey is one that seeks to measure **intangibles,** such as attitudes, opinions, values, or other psychological and sociological constructs.

If you classify surveys on the basis of their scope (census versus sample) and their focus (tangibles versus intangibles), four categories emerge: (1) a **census of tangibles,** (2) a **census of intangibles,** (3) a **sample survey of tangibles,** and (4) a **sample survey of intangibles.** Each type has its own contributions to make and its own inherent problems.

A Census of Tangibles

When you seek information about a small population, such as a single school, and when the variables involved are concrete, there is little challenge in finding the required answers. The strength of a census of this type lies in its **irrefutability**. Its weakness lies in its confinement to a single limited population at a single point in time. The information provided by such a census may be of immediate importance to a limited group, but typically such surveys add little to the general body of knowledge in education.

A Census of Intangibles

The task will be more difficult because this census deals with **constructs** that are **not** directly **observable** but must be inferred from indirect measures. Test scores and responses to questionnaires serve to approximate constructs such as knowledge and attitudes. The value of a census of intangibles is largely a question of the extent to which the instruments used actually measure the constructs of interest. Reasonably **good instruments** are available for measuring **aptitude** and **achievement** in a variety of academic areas.

A Sample Survey of Tangibles

When investigators seek information about large groups, the expense involved in carrying out a census is often **prohibitive**. Therefore, researchers use sampling techniques and use the information they collect from the sample to make inferences about the population as a whole.

A Sample Survey of Intangibles

The public opinion polls are examples of studies measuring intangible constructs. Opinion is not directly observable but must be inferred from responses made by the subjects to questionnaires or interviews. Opinion polling began in the 1930s and has grown tremendously.

Surveys Classified According to the Time Dimension

Longitudinal Surveys

Longitudinal surveys gather information at different points in time in order to study the changes over extended periods of time. Three different designs are used in longitudinal survey research: panel studies, trend studies, and cohort research.

Panel Studies In a **panel study,** the *same* subjects are surveyed several times over an extended period of time. Because the same subjects are studied over time, researchers can see the changes in the individuals' behavior and investigate the reasons for the changes.

Trend Studies A **trend study** differs from a panel study in that *different* individuals randomly drawn from the same general population are surveyed at intervals over a period of time. In fact, each time different students are selected.

Cohort Studies In a **cohort study,** a *specific* population is followed over a length of time with different random samples studied at various points. Whereas trend studies sample a general population that changes in membership over time, a cohort study samples a specific population whose members do not change over the duration of the survey. Typically, a cohort group has age in common. For example, a school system might follow the high school graduating class(es) of 2004 over time and ask them questions about higher education, work experiences, attitudes, and so on. **From a list of all the graduates, a random sample is drawn at different points in time**,

and data are collected from that sample. Thus, the population remains the same during the study, but the individuals surveyed are different each time.

Cross-Sectional Surveys

Cross-sectional surveys study a cross section (sample) of a population at a single point in time. The cross-sectional survey is the method of choice if you want to gather the data at one point in time.

How would you administer a questionnaire to assess changes in students' political attitudes during college with a (a) cross-sectional approach, (b) panel study, (c) trend study, and (d) cohort study?

Answers

a. In the cross-sectional study, you would draw a **random sample** from each of the four levels and administer the questionnaire to them at the **same time**.

b. Panel, trend, and cohort studies are all longitudinal. In all three, you first **randomly** draw a sample of **freshmen** from your population of interest. In a panel study, you assess your original sample and study **the same individuals** again when they are sophomores, juniors, and seniors.

c. In the trend study, you draw a **random sample of sophomores** from the population. A year later, you draw a **random sample of juniors**, and then in the final year you **draw a random sample of seniors**.

d. The cohort study would differ from the trend study in that the **subsequent samples** are drawn only from the population who were enrolled as **freshmen** when the study began and does not include students who transferred in later.

Longitudinal surveys are more time-consuming and expensive to conduct because the researcher must keep up with the subjects and maintain their cooperation over a long period of time. Cross-sectional surveys, in contrast, do not require years to complete. Hence, they are less expensive. A major disadvantage of the cross-sectional method is that **chance differences** between samples may seriously **bias the results**. You may by chance draw a sample of first-graders who are more mature than average and a sample of fourth graders who are less mature than average, with the result that the difference between the groups appears much smaller than it really is. However, researchers can usually obtain **larger samples for cross-sectional studies** than for longitudinal studies, and the larger samples mitigate the problem of chance differences.

Six Basic Steps Involved in Survey Research

1. *Planning.* Survey research begins with a question that the researcher believes can be answered most appropriately by means of the survey method. A research question in survey research

typically concerns the beliefs, preferences, attitudes, or other self-reported behaviors of the people (respondents) in the study.

2. *Defining the population.* One of the first important steps is to define the population under study. Defining the population is essential for identifying the appropriate subjects to select and for knowing to whom the results can be generalized. Once the population has been defined, the researcher must obtain or construct **a complete list of all individuals** in the population. This list, called the **sampling frame,** can be very difficult and time-consuming to construct if such a list is not already available.

3. *Sampling.* Because researchers generally cannot survey an entire population, they select a *sample* from that population. It is very important to select a sample that will provide results similar to those that would have been obtained if the entire population had been surveyed. In other words, the sample must be representative of the population. The sampling procedure that is most likely to yield a representative sample is some form of **probability sampling**. Probability sampling permits you to estimate how far sample results are likely to deviate from the population values.

4. *Constructing the instrument.* A major task in survey research is constructing the instrument that will be used to gather the data from the sample. The two basic types of data-gathering instruments are **interviews** and **questionnaires**.

5. *Conducting the survey.* Once the data-gathering instrument is prepared, it must be **field tested** to determine if it will provide the desired data. Also included in this step are **training** the users of the instrument, interviewing subjects or distributing questionnaires to them, and verifying the accuracy of the data gathered.

6. *Processing the data.* The last step includes coding the data, statistical analysis, interpreting the results, and reporting the findings.

Data-Gathering Techniques

There are two basic data-gathering techniques in survey research: interviews and questionnaires

Personal Interview

This technique has two advantages: (a) **flexibility** and (b) **response rate,** and (c) having **control** over the order of questions. The flexibility means the interviewer has the opportunity to observe the subject and the total situation in which he or she is responding. Questions can be repeated or their meanings explained in case they are not understood by the respondents. The interviewer can also press for additional information when a response seems incomplete or not entirely relevant. The term **response rate** refers to the proportion of the selected sample that agrees to be interviewed or returns a completed questionnaire. With interviews, **response rates are very**

high—perhaps 90 percent or better. Personal contact increases the likelihood that the individual will participate and will provide the desired information.

Another advantage is the control that the interviewer has over the order with which questions are considered. In some cases, it is very important that respondents not know the nature of later questions because their responses to these questions may influence earlier responses.

The main disadvantage of the personal interview is that it **is more expensive** than other survey methods. Another disadvantage is the possibility of **interviewer bias,** which occurs when the interviewer's own feelings and attitudes or the interviewer's gender, race, age, and other characteristics influence the way questions are asked or interpreted. As a general rule, interviewers of the same ethnic/racial group get the most accurate answers to race-related questions. Still **social desirability bias** is another problem in which respondents want to please the interviewer by giving socially acceptable responses that they would not necessarily give on an anonymous questionnaire.

Focus Groups

A specific category of interviews is the **focus group.** Several subjects are interviewed at the same time. An advantage of a focus group is that participants respond not only to the researcher but also to other participants and their responses. This method can provide the researcher with insight into how disagreements are or are not resolved. Sometimes the researcher can report a final consensus. Focus groups are often used in **qualitative research**. The researcher invites people who are interested in the same general topic to assemble to discuss it. They are assured that they will be **free to express themselves** in their own words and to respond not only to the researcher but also to other participants and their responses.

Telephone Interviews

The telephone interview is popular, and studies show that it compares quite favorably with face-to-face interviewing. Its major advantages are lower cost and faster completion, with relatively high response rates. The main disadvantage of the telephone interview is that there is less opportunity for **establishing rapport** with the respondent than in a face-to-face situation. Another limitation of telephone interviews is that **complex questions** are sometimes difficult for respondents to follow. If they misunderstand the questions, the interviewer may not know.

Computer-Assisted Telephone Interviewing (CATI)

Computer and telecommunications technology has been applied to telephone surveys. Wearing earphones, the interviewer sits at a computer while it randomly selects a telephone number and dials.

Conducting the Interview

Whether the interview is conducted in person or by telephone, the interviewers' main job is to ask the questions in such a way as to obtain **valid responses** and to record the responses accurately and completely. The initial task for the interviewer is to create an atmosphere that will put the **respondent at ease**. After introducing yourself in a **friendly way**, briefly state the purpose of the interview but **avoid giving too much information** about the study, which could **bias** the respondent. It is well to begin the interview with fairly simple, nonthreatening questions.

The interviewer also has the responsibility of keeping the respondent's attention focused on the task and for keeping the interview moving along smoothly.

Less or More Structured Interview?

Interviews can be **more or less structured**. In a less structured interview, the same questions are asked of all respondents, but the interview is more conversational and the interviewer has more freedom to arrange the order of the questions or to rephrase the questions. If comparable data are to be obtained, however, the interviewer must standardize the procedure by using a structured interview schedule. A structured interview schedule contains specific questions in a fixed order, to be asked of all respondents, along with transition phrases and **probes** (questions used to clarify a response or that push a little further into a topic). For example, if the respondent starts to hedge, digress, or give irrelevant responses, or if he or she has obviously misinterpreted the question, then the interviewer may use a fixed probe such as "***Explain your answer a little further***" or "***Can you tell me a little more about that***?"

Another important technique besides the probe is the **pause.** A good interviewer needs skill in listening and is quiet at times until the respondent answers. In less structured interviews, any marked deviations from the protocol should be documented so that the information can be taken into account when analyzing the interviewee's response. In using probes, take care not to suggest or give hints about possible responses. Interviewer trainees should be provided with written manuals on interviewing procedures.

Mailed Questionnaires

A **mailed questionnaire** has the advantage of guaranteeing **confidentiality** or **anonymity**, thus perhaps eliciting more truthful responses than would be obtained with a personal interview. In an interview, subjects may be reluctant to express unpopular or politically incorrect points of view or to give information they think might be used against them at a later time. The mailed questionnaire also eliminates the problem of interviewer bias.

A disadvantage of the mailed questionnaire is the possibility of respondents misinterpreting the questions. Another important limitation of mailed questionnaires is the low return rate. A low response rate limits the generalizability of the results of a questionnaire study.

A number of factors have been found to influence the rate of returns for a mailed questionnaire, including (1) length of the questionnaire, (2) cover letter, (3) sponsorship of the questionnaire, (4) attractiveness of the questionnaire, (5) ease of completing it and mailing it back, (6) interest aroused by the content, (7) use of a monetary incentive, and (8) follow-up procedures used.

Directly Administered Questionnaires

A **directly administered questionnaire** is given to a group of people assembled at a certain place for a specific purpose. The main advantage of directly administering questionnaires is the **high response rate**, which typically is close to 100 percent. Other advantages are the **low cost** and the **fact** that the researcher is present to provide assistance or answer questions. The disadvantage is that the researcher is usually restricted in terms of where and when the questionnaire can be administered. Also, when a population is limited, the results of the survey will be equally limited in terms of generalizability.

Standard Error of the Sampling Proportion

Even with random sampling there will always be some error in estimating a population parameter from sample statistics. The statistic most commonly reported in a sample survey is a proportion or a percentage of the sample that gives a particular response. The discrepancy between the **known sample proportion** and the **unknown population value** is referred to as **sampling error.** The first step in assessing how much sample results are likely to deviate from the population values is to calculate the standard error of the sampling proportion.

Constructing the Instrument: Format of Questions

Two basic types of questions are used in survey instruments: closed-ended or fixed alternative and open-ended or free-response questions. Use **closed-ended questions** when all the possible, relevant responses to a question can be specified, and the number of possible responses is limited. **Open-ended questions** are used when there are a great number of possible answers or when the researcher cannot predict all the possible answers.

A limitation of the closed-ended question is that it does **not provide much insight** into whether respondents really have any information or any clearly formulated opinions about an issue.

Matrix Sampling

A procedure called **matrix sampling** is sometimes used when the survey is long and the accessible population is large. This technique involves **randomly** selecting respondents, each of whom is administered a subset of questions, randomly selected from the total set of items.

Field Testing

Before the final printing, the researcher must **field test** the instrument to identify ambiguities, misunderstandings, or other inadequacies. First, it is a good idea to **ask colleagues** to examine a

draft of the questionnaire and give their opinions. Next, **administer the questionnaire personally** and one at a time to a small group drawn from the population to be considered in the study. Respondents answer the questions and provide feedback to the researcher on any difficulties they have with the items. The results of field tests can be used to clarify the items or perhaps to eliminate some.

Validity in Interview and Questionnaire

Attention should be given to the validity of interviews and questionnaires— that is, whether they are really measuring what they are supposed to measure. *Face validity* can be important in survey research. Subjects should perceive questions to be relevant.

Construct validity can be assessed by having some colleagues who are familiar with the purpose of the survey examine the items to judge whether they are appropriate for measuring what they are supposed to measure and whether they are a representative sample of the behavior domain under investigation.

Criterion-related validity can be based on the relationship of survey responses to other variables. **Direct observation of behavior**, for example, has been a criterion used to validate surveys. After responses were obtained, observations were made to determine whether the actual behavior of the subjects agreed with their expressed attitudes, opinions, or other answers. If you find agreement between ***survey responses*** and ***actual behavior***, you have some evidence for the criterion-related validity of the survey.

Reliability

A procedure for assessing the reliability of an interview procedure is to have two or more interviewers ask the same subjects identical questions and then assess the consistency of the responses that the interviewers report. With questionnaires, internal consistency may be checked by building some **redundancy** into the instrument – items on the same topic may be rephrased and repeated in the questionnaire or interview. The more consistent the responses, the higher the reliability.

Cross Tabulations

Cross tabulations provide an excellent way to show the relationship existing among the variables in a survey.

Chapter 15: Defining and Designing Qualitative Research

Chapter 15 discusses qualitative research in terms of how it differs from quantitative research, and how to carry it out.

Quantitative approaches in the human sciences rely on a **hypothetico-deductive model** of explanation. Inquiry begins with a theory of the phenomenon to be investigated. From that theory any number of hypotheses are deduced that, in turn, are tested using a predetermined procedure such as an experimental, ex post facto, or correlational design. The ultimate goal of researchers using this hypothetico-deductive model is to **revise** and **support theories** of social and behavioral phenomena based on the results of hypothesis testing. One goal of quantitative approaches is to generalize findings from a randomized sample to a larger population.

The ultimate goal of qualitative inquiry is to portray the complex pattern of what is being studied in sufficient depth and detail so that someone who has not experienced it can understand it. When qualitative inquirers interpret or explain the meaning of events, actions, and so forth, they generally use one of the following types of interpretation: (1) construction of patterns through analysis and resynthesis of constituent parts, (2) interpretation of the social meaning of events, or (3) analysis of relationships between events and external factors. These interpretations may lead to the generation of theories.

Methods

Quantitative methods use empirical approaches, experimental designs, and often statistical testing compared to the more **naturalistic**, **emergent**, and **field-based methods** typical of qualitative research. The primary instrument used for data collection in qualitative research is the researcher him- or herself, often collecting data through **direct observation** or **interviews**. Quantitative research more typically relies on measurement tools such as scales, tests, **observation checklists**, and questionnaires. The selection of subjects for study also differs. The ideal selection in quantitative research is random sampling, which allows for control of variables that may influence findings. Qualitative studies more typically use **nonrandom** or **purposive** selection techniques based on particular criteria.

Values

Quantitative inquirers admit that the inquirer's values may play a role in deciding what topic or problem to investigate but maintain that the actual investigation should aim to be as value free as possible; that is, the inquirer must follow procedures specifically designed to isolate and remove subjective elements to the extent possible. The goal is to **control or remove personal value** from the inquiry situation so that what remains are just the "objective facts".

Qualitative inquirers argue that inquiry is **value bound** in the choice of a problem to investigate, in the choice of whether to adopt a quantitative or qualitative approach to a problem, in the

choice of methods used to investigate that problem, in the choice of a way to interpret results or findings, and by the values inherent in the context where the study takes place. Qualitative inquirers believe that it is impossible to develop a meaningful understanding of human experience without taking into account the **interaction** of both the inquirers and the participants' values and beliefs. They believe that rather than try to eliminate bias, it is important to **identify** and **monitor biases** and how they may affect data collection and interpretation.

Major Characteristics of Qualitative Research

Concern for Context and Meaning

Qualitative inquiry shows **concern for context and meaning.** It assumes that human behavior is context bound—that human experience takes its meaning from and, therefore, is inseparable from social, historical, political, and cultural influences. Thus, inquiry is always bounded by a particular context or setting. Qualitative researchers focus on **how people make sense** of or **interpret their experience**. Qualitative inquiry aims to understand intention. There is **no attempt** to **predict** what will happen in the future but, rather, to understand a unique and particular context. Proponents of qualitative inquiry argue that the quantitative approach to the study of human experience seeks to isolate human behavior from its context; it engages in **context stripping.**

Naturally Occurring Settings

Qualitative research studies behavior as it occurs naturally in a classroom, or an entire school. Qualitative inquiry takes place in the field, in settings as they are found. It is not a setting contrived specifically for research, and there is no attempt to manipulate behavior. In addition, qualitative inquiry places **no prior constraints** on what is to be studied. It does **not** identify, define, and investigate or test the relationship between independent and dependent variables in a particular setting.

Human as Instrument

In qualitative studies, the human investigator is the primary instrument for the gathering and analyzing of data.

Descriptive Data

The qualitative inquirer deals with data that are in the form of words or pictures rather than numbers and statistics.

Emergent Design

In quantitative studies, researchers carefully design all aspects of a study *before* they actually collect any data; they specify variables, measures for those variables, statistics to be used to analyze data, and so forth. In contrast, while qualitative inquirers broadly specify aspects of a

design before beginning a study, the design continues to *emerge* as the study unfolds, hence the term **emergent design**. They adjust their methods and way of proceeding (design) to the subject matter at hand. This is necessary because the qualitative inquirer is never quite sure just what will be learned in a particular setting because what can be learned in a particular setting depends on the **nature and types of interactions between the inquirer and the people and setting**, and those interactions are not fully predictable, and also because important features in need of investigation cannot always be known until they are actually witnessed by the investigator.

Inductive Analysis

It is a process of **inductive data analysis;** it proceeds from data to theory or interpretation. As the inquirer reduces and reconstructs the data through the processes of coding and categorization, he or she aims at interpreting the phenomena being observed.

Designing Qualitative Research

The qualitative researcher begins from a **conceptual framework**—a "system of concepts, assumptions, expectations, beliefs, and theories" (Maxwell, 2005) that informs the design. The design begins with a general statement of a **research problem** or topic. This initial topic that a qualitative researcher chooses for investigation is referred to as the **focus of inquiry.**

To develop the focus of inquiry, the beginning researcher needs to think about some topic in which he or she has an interest and wants to know more about. The research question may be one that comes from the investigator's **observations** and **experiences** with particular **topics**, **settings**, or **groups**. Qualitative problems examine the context of events, real-world setting, subjects' perspectives, unfolding and uncontrolled events, reasons for the events, and phenomena needing exploration and explanation.

Research Question Choice

The choice of the research question is crucial because the question (what you really want to understand) determines the design. Maxwell (2005) describes types of research questions posed in qualitative research. **Particularizing questions** ask about a specific **context**—*what is happening in this particular school?*—and are less concerned about generalizing but, rather, focus on *developing rich descriptions and interpretations*. Case studies typically use particularizing questions. **Generic questions** about a broader population are more typically used in ***quantitative*** research with samples selected as representative in an attempt to generalize. Generic questions can be used in qualitative research, such as with multisite studies, but must be used with caution.

Process questions examine how things happen—the process by which a phenomenon takes place. Questions asking about *meaning*, *influences*, and *context* are process oriented. **Variance questions** (questions that ask *to what extent* or about differences) are best answered by

quantitative studies rather than by qualitative studies. **Instrumentalist questions** are formulated in terms of observable, measurable data and are the norm in quantitative studies.

Realist questions treat unobserved phenomena (*feelings*, *beliefs*, *intentions*, etc.) as real and are common is qualitative studies. The difference between these might be seen in a study involving interviews in which the research question is posed as "*What is the effect on other students when a classmate is a victim of gang violence?*" (a realist question) or as "*What effects do students self-report when a classmate is a victim of gang violence?"* (an instrumentalist approach).

Tunnel Vision

Qualitative researchers **do not begin** a study with **no questions**; they begin with a base of experience, theoretical knowledge, and certain goals that drive **provisional questions** that may evolve with time. If the initial research questions are too diffuse, the researcher may have difficulty in the design phase or in connecting to research goals. If the research questions are too focused, it may create **"tunnel vision**".

After deciding on the problem and questions and determining that qualitative methodology is indeed appropriate, next you need to make decisions about the particular qualitative approach, the main data collection tools, the setting for the study, the participants, the size of sample, and the behaviors to study. A qualitative design, however, is flexible and may be changed as the researcher gets into the setting.

Several criteria are available for evaluating the qualitative design to be used to answer the research question: One criterion is **informational adequacy.** That is, does the research plan maximize the possibility that the researcher will understand the setting thoroughly and accurately? A second criterion is **efficiency.** Does the plan allow adequate data to be collected in a cost and time-effective manner? A third criterion to use is **ethical considerations.**

Sampling

Qualitative researchers are purposeful in selecting participants and settings. They select **purposive samples** believed to be sufficient to provide maximum insight and understanding of what they are studying.

Because of the depth and extent of the information sought in qualitative studies, purposive samples are typically small. How large should the sample be? **There is no general rule** about the number of participants to include in a qualitative study. Of course, practical considerations such as time, money, and availability of participants influence the size of the sample. However, the primary criterion of sample size is **redundancy** of **information**. Sampling should be terminated when no new information is forthcoming from new units. A unit is an individual participant, group, organization, event, setting, document, or artifact selected as part of the qualitative study. This point is referred to as **data saturation.**

1. **Comprehensive sampling**. In comprehensive sampling, every unit is included in the sample. For example, a study of physically disabled students in a high school would include all such students in the school. Comprehensive sampling is used when the number of units is small.

2. **Critical case sampling.** Critical case sampling involves the selection of a single unit that provides a crucial test of a theory or program. An example is selecting a single school that has decided to adopt a well-known character education program in order to change the culture of the school and observing the school during a year-long implementation to determine the impact on behaviors and interactions in the school. Examination of critical cases can enhance the ability to generalize or apply findings to other cases.

3. **Maximum variation sampling.** In maximum variation sampling, units are included that maximize differences on specified characteristics. For example, a study of U.S. high school students might include students from schools that differ in location, student characteristics, parental involvement, and other factors. This type of sampling reveals differences but may also identify commonalities across the units.

4. **Extreme, deviant, or unique case sampling.** Extreme case sampling selects units that are atypical, special, or unusual. For example, you might choose to study a high-poverty, inner-city elementary school that has achieved exemplary reading and mathematics test scores.

5. **Typical case sampling.** Typical case sampling selects units that are considered typical of the phenomenon to be studied. In a study of an elementary school, you would select a school considered typical rather than a very high achieving school or a very low achieving school. This approach highlights what is normal or average.

6. **Negative or discrepant case sampling.** This method of sampling selects units that are examples of exceptions to expectations. The researcher would intentionally look for examples that appear not to confirm the theory being developed. This strategy is also called confirming and disconfirming sampling.

7. **Homogeneous sampling.** Homogeneous sampling selects a subgroup that is considered homogeneous in attitudes, experiences, and so on. For example, you might choose only a sample of special education teachers from a population of teachers. This approach may be used with focus group interviewing.

8. **Snowball, chain, or network sampling.** Snowball, chain, or network sampling occurs when the initially selected subjects suggest the names of others who would be appropriate for the sample. These next subjects might then suggest others and so on.

9. **Intensity sampling.** Intensity sampling involves selecting participants who exhibit different levels of the phenomenon of interest to the researcher. The researcher would select several cases at each of several levels of variation of the phenomenon. For example, the researcher may select

some high achieving, average-achieving, and low-achieving students or in a study of bullying, may select students who have different levels of aggressive tendencies.

10. **Stratified purposeful sampling.** Stratified purposeful sampling attempts to ensure that subgroups are represented so that **comparisons** can be facilitated. For example, in a study of teaching practices, experienced and inexperienced teachers would be included for observation.

11. **Random purposeful sampling.** When the potential purposeful sample is too large (e.g., when resources are limited), the credibility of the study can be enhanced by randomly selecting participants or sites from the larger group.

12. **Theoretical** or theory-based **sampling.** In theoretical sampling, the researcher begins by selecting a person or site that exemplifies the theoretical construct and continues to select new cases that reflect the developing theory to include as the research unfolds and the theory emerges.

13. **Criterion sampling.** In this type of sampling, the researcher sets the criterion and includes all cases that meet that criterion.

14. **Opportunistic sampling.** Opportunistic sampling takes advantage of new leads or unexpected opportunities.

15. **Convenience sampling.** Convenience sampling is choosing a sample based on availability, time, location, or ease of access. Convenience sampling is not recommended because it may produce evidence that is not credible. Studies of your children or your workplace are examples of convenience sampling.

Data Collection

The most common data collection methods used in qualitative research are (1) *observation*, (2) *interviewing*, and (3) *document* or *artifact analysis*. Artifacts may include audio and video recordings, photographs, games, artwork, or other items that provide insight related to the context or participants.

Observation

Observation is a basic method for obtaining data in qualitative research. It is a more global type of observation than the systematic, structured observation used in quantitative research. The qualitative researcher's goal is a **complete description of behavior** in a specific setting rather than a numeric summary of occurrence or duration of observed behaviors. Qualitative observation usually takes place over a **more extended period of time** than quantitative observation. Also, qualitative observation is more likely to proceed without any prior hypotheses. Quantitative observations often use checklists and behavior observation tools developed prior to the observation to record or document observed behaviors. Qualitative

observations rely on narrative or words to describe the setting, the behaviors, and the interactions. The goal is to understand **complex interactions** in natural settings.

There are some specialized approaches to observation, such as **interaction analysis** (sometimes used in small group or classroom settings). Two types of interaction analysis are ***kinesics*** (the study of body movements and how those motions communicate messages) and ***proxemics*** (the study of how people use space). In both kinesics and proxemics, there are limitations related to cultural awareness because gestures.

Stance toward Observation

Five stances toward observation have been identified: (1) complete participant, (2) participant as observer, (3) observer as participant, (4) complete observer, and (5) collaborative partner.

A **complete** or **covert participant** is a member of the group or context under study and focuses on the natural activity of the group without informing the group that it is under study. The ethics of the covert approach, however, may be questionable.

In the **participant as observer** stance, the observer actively participates and becomes an insider in the event being observed so that he or she experiences events in the same way as the participants. The researcher's role is known to the people being observed. Anthropologists often are participant observers when they conduct a study of a particular culture.

In the **observer as participant** stance, researchers may interact with subjects enough to establish ***rapport*** but do not really become involved in the behaviors and activities of the group. Their status as observer/researcher is known to those under study. Their role is more peripheral rather than the active role played by the participant observer.

The **complete observer** is typically hidden from the group or may be simply in a public setting observing public behavior. The qualitative researcher simply observes and records events as they occur. No attempt is made to alter the situation in any way. These are considered naturalistic observations.

Simple **naturalistic observation** can take a great deal of time because you must wait for the behavior to occur naturally. For this reason, some researchers set up contrived naturalistic situations to elicit the behavior to be observed. Although the setup is contrived, the researcher tries to maintain the naturalness of the situation and makes the observations in a way not noticeable to the subjects.

The **collaborative partner** stance described in action research and feminist research has as a defining characteristic an equal partnership in the research process between the researcher and participants.

The degree of participation in an observation study is thus a continuum ranging from a complete participant at one end to a complete observer at the other. It is easier to ask questions and record observations if members of the group know your purpose; furthermore, it may be more ethical to make people aware of what is going on. **Being open, however, may present problems**. Knowing they are being observed, group members may behave differently from the way they usually do, or they may not be truthful when answering questions. This impact of the observer on the participants being studied is called **observer effect** and can result in an inaccurate picture of the group and its interactions. There is a risk that the observer will destroy the very naturalness of the setting that he or she wants.

Observer expectation may occur when the researcher knows the participants are associated with certain characteristics and may expect certain behaviors. In other words, expectations may cause you to see or interpret actions or events in a particular way.

Another problem with observation is the possible effect that the observer him- or herself might have on the results. **Observer bias** occurs when the observer's personal attitudes and values affect the observation and/or the interpretation of the observation.

The most common method of recording the data collected during observation is **field notes.** The researcher may make brief notes during the observation but then later expands his or her account of the observation as field notes. Field notes have two components: (1) the ***descriptive*** part, which includes a complete description of the setting, the people and their reactions and interpersonal relationships, and accounts of events (who, when, and what was done); and (2) the ***reflective*** part, which includes the observer's personal feelings or impressions about the events, comments on the research method, decisions and problems, records of ethical issues, and speculations about data analysis. The researcher's reflections are identified as **observer comments** to distinguish them from the descriptive information.

Although field notes are the most common data collection technique used in observations, other techniques may include **audio or video recordings** or **photographs**. A disadvantage of some recording methods is that participants may be conscious of the camera or other recording device and behave differently or may try to avoid being filmed or photographed.

Interviews

The **interview** is one of the most widely used and basic methods for obtaining qualitative data. Interviews are used to gather data from people about opinions, beliefs, and feelings about situations in their own words. They are used to help understand the experiences people have and the meaning they make of them rather than to test hypotheses. Interviews may provide information that cannot be obtained through observation, or they can be used to verify observations.

For example, observing a teacher in a classroom tells us something about the behavior, but interviewing helps us to **put the behavior in context** and helps us understand actions and choices. The qualitative interview is typically **more probing** and **open ended and less structured** than the interview used in quantitative research but varies considerably in the way it is conducted.

At one extreme is the **unstructured interview** in which the questions arise from the situation. It is sometimes described as "*a conversation with a purpose*". The most *data-dense interviews* may be of this form. The interview is not planned in detail ahead of time; the researcher asks questions as the opportunity arises and then listens closely and uses the subjects' responses to decide on the next question. The subjects in the setting may not even realize they are being interviewed. Using the *who, what, when, where, why, and how categories* is generally enough guidance for the researcher to follow in asking questions.

At the other end of the continuum lies the more **structured interview**, scheduled for the specific purpose of getting **certain information** from the subjects. Each respondent is asked the same set of questions, but **with some latitude (freedom)** in the sequence. Although the questions are structured, qualitative structured interviews differ from quantitative structured interviews. In the qualitative approach, the list of questions is generally more limited in length and most questions cannot be answered with yes or no or limited word responses.

In between the unstructured and structured interview is the **semi-** or **partially structured interview,** in which the area of interest is chosen and questions are formulated but the interviewer may modify the format or questions during the interview process. One characteristic that all qualitative interview formats share is that the ***questions are typically open ended*** (cannot be answered with a yes or no or simple response) and the questions are designed to reveal what is important to understand about the phenomenon under study.

An interview has the advantage of supplying large volumes of **in-depth data rather quickly.** Interviews provide insight on participants' perspectives, the meaning of events for the people involved, information about the site, and perhaps information on unanticipated issues. Interviews allow **immediate follow-up** and clarification of participants' responses. One disadvantage of the interview as a data-gathering tool is that interviewees may not be willing to share information or may even offer false information.

One of the most efficient ways to collect interview data is to use an **audio recorder**. This is much less distracting than taking notes, and it also provides a verbatim (precise; exact) record of the responses.

Qualitative interviews might involve **one-time interviews** with a subject or subjects, **multiple interviews** with the same subject or subjects, or group interviews or **focus groups**. A **focus group,** which is like a group interview, typically centers on a *particular issue*; the trained interviewer elicits the views of the group members while *noting interactions* within the group.

The assumption is that individual attitudes, beliefs, and choices of action do not form in a vacuum. Listening to others helps people form their own opinions. Focus groups are helpful because they bring several different perspectives into contact. The researcher gains insight into how the participants are thinking and why they are thinking as they do.

A **focused interview** is much more flexible and open in form than the survey interview discussed in Chapter 14. The respondents are **free to answer** in their own words and can answer either briefly or at length. The questions asked may even vary from individual to individual. The responses are recorded by taking notes or with an audiotape. Focus groups are *more socially oriented* than individual interviews and can increase the sample size in the study, but they allow less control than individual interviews and data can be more **difficult to analyze**.

Focus groups typically consist of 6 to 12 people. The group should be small enough that everyone can take part in the discussion but large enough to provide diversity in perspective. Focus group discussions usually need to last at least1 hour and possibly 2 hours. Groups should be homogeneous in terms of prestige and status to ensure comfort in expressing opinions.

Focus group interviewing is a specific approach used in qualitative research, but there are other approaches as well that are related to particular types of qualitative research. **Ethnographic interviewing,** grounded in anthropology, attempts to understand the participants' worldviews through gathering cultural knowledge and includes descriptive questions, structural questions, and contrast questions. **Phenomenological interviewing,** grounded in philosophy, attempts to examine lived experience through three in-depth interviews. **Elite interviewing** selects individuals based on their expertise—those who are considered particularly influential or well informed.

Documents and Artifacts

Qualitative researchers may use written documents or other artifacts to gain an understanding of the phenomenon under study. The term ***documents*** here refers to a wide range of written, physical, and *visual materials*, including what other authors may term artifacts. Documents may be personal, such as *autobiographies*, diaries, and letters; official, such as fi les, reports, memoranda, or minutes; or documents of popular culture, such as books, films, and videos**.**

Documents can be classified into four categories: (1) *public records*, (2) *personal documents*, (3) *physical materials*, and (4) *researcher-generated documents*.

Ethical Considerations in Qualitative Research

1. *Researcher's relationship to participant.* After spending a great amount of time observing or interviewing, the **researcher's relationship to participants** may gradually become less that of researcher and researched and more like friendship. Because the researcher is regarded as a friend, the participants trust him or her and may forget a

research study is going on. Some field researchers say they obtain their best data at this point but at the same time are most ethically vulnerable.

2. **Reciprocation**. Another issue about which the researcher should be concerned is the issue of reciprocity. The people in the research setting have given of themselves to help the researcher, and he or she is indebted. Qualitative researchers need to give participants something in return for their time, effort, cooperation, and just tolerating their extended presence. They might offer to provide a ***written report***, ***present the findings*** at a school or neighborhood meeting, ***give advice or assistance*** on other research projects at the school, help with ***grant writing***, and so forth.
3. *Getting permission to conduct research.* Like the quantitative researcher, the qualitative researcher must get approval for the project from his or her institution's Human Subjects Research Committee, especially if minors are included in the research.
4. *Kind of information obtained*.

Chapter16: Types of Qualitative Research

Basic Qualitative Studies

Basic qualitative studies, also called basic *interpretative studies* by some, provide rich **descriptive accounts** targeted to understanding a phenomenon. The central purpose of these studies is to understand the world or the experience of another. The underlying question the researcher is asking is "*How are events, processes, and activities perceived by participants*?" It has its own roots in the ***social sciences***.

Basic interpretive studies are more simplistic compared to other qualitative approaches. They are not restricted to a particular phenomenon as in case studies. They do not seek to explain sociocultural aspects as in ethnography. They do not seek to enter the subject's conceptual world to explain the "essence" as in phenomenology. They do not seek to define theory as in grounded theory research. They do not convey life stories through narrative analysis, delve into history, or focus on analyzing content. These studies are, as the name implies, basic. They describe and attempt to ***interpret experience***.

Case Studies

Emerging from approaches in business, law, and medicine, a **case study** focuses on a single unit to produce an in-depth description that is rich and holistic. The underlying question is "*What are the characteristics of this particular entity, phenomenon, person, or setting?"* Case studies typically include multiple sources of data collected over time. As indicated, case studies provide an in-depth description of a single **unit.** The "unit" can be an individual, a group, a site, a class, a policy, a program, a process, an institution, or a community.

A specific unit may be selected because it is unique or typical or for a variety of other reasons. The unit is defined within specific boundaries, referred to as a "**bounded system".** To be bounded, the phenomenon must be identifiable within a *specific context*. If it cannot be described in such a way, case study may not be the best approach to study it.

In comparing a ***case study*** with ***single-subject experiments*** (see Chapter 11), both may study a single individual. However, single-subject experiments focus on a single behavior or a very limited number of behaviors, whereas case studies attempt to describe the subject's entire range of behaviors and the relationship of these behaviors to the subject's history and environment. In a case study, the investigator attempts to examine an individual or unit in-depth.

Note: However, case studies need not be limited to the study of individuals. Case studies are made of communities, institutions, and groups of individuals. A more recent ***community case study*** by Matthew Corrigan (2007) examines race, religion, and economic change in the Republican South by focusing on one southern city.

Types of Case Studies

Three types of case studies have been described. The **intrinsic case study** is conducted to understand a particular case that may be ***unusual, unique, or different*** in some way. It does not necessarily represent other cases or a broader trait or problem for investigation. The case in and of itself is of interest to the researcher.

In an **instrumental case study,** the researcher selects the case because it represents some other issue under investigation and the researcher believes this particular case can help provide insights or help to understand that issue. The case is ***illustrative*** of something under investigation.

The **multiple** or **collective case study** uses several cases selected to further understand and investigate a phenomenon, population, or general condition. The researcher believes that the phenomenon is not idiosyncratic to a single unit and studying multiple units can provide better illumination.

Case studies may employ multiple methods of data collection and do not rely on a single technique. Testing, interviewing, observation, review of documents and artifacts, and other methods may be used. The distinction is that whatever techniques are used, all are focused on a single phenomenon or entity (the case).

The case study researcher starts with a particular concern or topic, and from that general area emerge **foreshadowed problems.** A purposeful choice is made of the bounded system to be studied, and then data are collected from multiple sources and analyzed. Two kinds of analysis appropriate for case studies have been described: **holistic analysis** of the entire case and **embedded analysis** that focuses on specific aspects of the case. Multiple case studies require analysis across site.

Researchers conducting case studies provide a detailed report that may build on narratives, vignettes, tables, charts, figures, visual displays, text, and more. Typically, the report is written to provide both an **emic,** or insider, **perspective** (the perspective of the individuals who are part of the case) as well as an **etic,** or outsider, **perspective** (the interpretations of the researcher).

Content or Document Analysis

Content or **document analysis** is a research method applied to written or visual materials for the purpose of identifying specified characteristics of the material. The materials analyzed can be textbooks, newspapers, web pages, speeches, television programs, advertisements, musical compositions, or any of a host of other types of documents. It is rooted in ***communication*** studies.

An advantage of content analysis is its **unobtrusiveness.** The presence of the observer does not influence what is being observed. Another advantage of content analyses is that they are easily replicated. However, content analysis can be slow and time-consuming.

Ethnographic Studies

Ethnography is the in-depth study of naturally occurring behavior within a culture or entire social group. Ethnographers typically describe, analyze, and interpret culture over time using observations and fieldwork as the primary data collecting strategies. The final product is a **cultural portrait** that incorporates the views of participants (emic perspective) as well as views of researcher (etic perspective). Ethnographic studies consider where people are situated and how they go about daily activities as well as cultural beliefs.

Anthropologists immerse themselves in the lives of the people they study, using primarily **extended observation** (participant and nonparticipant) and occasionally in-depth interviewing to gain clarification and more detailed information. The ethnographer explores and tests hypotheses, but the hypotheses evolve out of the fieldwork. Ethnographers refer to the people from whom they gather information as "**informants"** rather than participants, and they study "sites" rather than individuals.

Creswell (2007) describes two approaches to ethnography. **Realist ethnography** is the more *traditional* approach. In realist ethnography, the researcher tries to provide ***an objective account of the situation***, typically from a third-person point of view. Standard categories are used, and factual information and closely edited quotes are presented as data. The researcher's ***interpretation occurs at the end***. In **critical ethnography,** the researcher takes an **advocacy perspective** and has a value-laden orientation. The researcher is ***advocating for a marginalized group***, challenging the status quo, or attempting to empower the group by giving it voice.

Grounded Theory Studies

Grounded theory has its roots in sociology. Its goal is to inductively build a theory about a practice or phenomenon using **interviews** and **observation** as the primary data collection tools. This emphasis on theory distinguishes it from other qualitative approaches.

The *personal open-ended interview* is the primary method of data collection in grounded theory studies. The interviewer asks questions about what happened to individuals, why it happened, and what it means to them. Choose a sample where each individual has had the experience and can contribute to theory development. The study may include as many as 20 to 25 subjects who are interviewed on the topic until no new information is forthcoming (**data saturation**).

The concept of **saturation** was first defined in the context of grounded theory as theoretical saturation. In qualitative research the word saturation is extensively used almost interchangeably with data ***saturation, thematic saturation***, ***theoretical saturation*** and ***conceptual saturation***.

Saturation point determines the sample size in qualitative research as it indicates that adequate data has been collected for a detailed analysis. However, there are no fixed sizes or standard tests that determine the required data for reaching saturation.

Documentary materials (letters, speeches, etc.) and literature can also be potential data sources. In reviewing text materials, it is important to identify whether the text is **extant** (those the researcher did not shape, such as letters or diaries) or **elicited** (those in which the researcher involved participants in writing, such as through an internet survey). Text used in the study must be situated in the context.

After forming categories having similar units of meaning, the researcher searches for ***underlying themes and relationships*** among the categories. This analysis of the data results in insights, conditional propositions, and questions that are pursued through further data collection. The researcher constructs ***tentative theoretical statements*** about the relationships among constructs, explores these theoretical propositions through further data collection, and so on. This cyclical process of testing the explanatory adequacy of the theoretical constructs by comparing with additional empirical data continues until the comparative analysis no longer contributes anything new (**theoretical saturation**).

Thus, through **induction** and **verification** techniques, the researcher progressively elaborates a general theoretical statement well-grounded in the data. The **constant comparative method of analysis** is typically used in grounded theory. In this method, the researcher compares units of data with each other to generate **tentative categories**, eventually reducing these to **conceptual categories** that evolve into an overall framework or theory. Generating the theory is not easy; it requires insight and understanding and, as indicated, many reviews of the data.

Description of Coding Types Used in Grounded Theory Studies

Open coding: It deals with labeling and categorizing phenomenon in the data. It uses the comparative method. Data are broken down by asking what, where, how, when, how much, etc. Similar incidents are grouped together and given the same conceptual label. Concepts are grouped together into categories. The purpose is to develop core concepts, categories, and properties.

Axial coding: It is designed to put data back together that were broken apart in open coding. It develops connections between a category and its subcategories (not between discrete categories). Its purpose is to develop main categories and subcategories.

Selective coding: It shows the connections between the discrete categories. Categories that have been developed to build the theoretical framework are integrated. Its purpose is to bring the categories together into an overall theory.

Historical Studies

Historical studies are oriented to the past rather than to the present and thus use different data collection methods from those used in other qualitative approaches. **Historical research** is included in qualitative research because of its emphasis on interpretation and its use of nonnumeric data.

Historical research is an attempt to establish facts and **arrive at conclusions** concerning the past. The historian systematically locates, evaluates, and interprets **evidence** from which people can learn about the past. Based on the evidence gathered, conclusions are drawn regarding the past so as to increase knowledge of how and why past events occurred and the process by which the past became the present.

The historian operates under different handicaps from those of researchers in other fields. Control over treatment, measurement, and sampling is limited, and there is **no opportunity for replication**. Another limitation impinging on historical researchers is that **no assumption** about the past can be made.

The historian classifies materials as primary and secondary sources. **Primary sources** are original documents (correspondence, diaries, reports, etc.), relics, remains, or artifacts. These are the direct outcomes of events or the records of participants. With **secondary sources,** the mind of a nonobserver comes between the event and the user of the record.

Two ideas that have proved useful in evaluating historical sources are the concepts of **external** (or lower) criticism and **internal** (or higher) criticism. Basically, **external criticism** asks if the evidence under consideration is authentic and, depending on the nature of the study, may involve such techniques as authentication of signatures, chemical analysis of paint, or carbon dating of artifacts. After the authenticity of a piece of evidence has been established, the historical investigator proceeds to **internal criticism,** which requires evaluating the worth of the evidence, for instance, whether a document provides a true report of an event.

Narrative Research

Narrative research has its roots in different humanities disciplines and focuses on stories (spoken or written) told by individuals about their lives. The researcher emphasizes **sequence** and **chronology** and a collaborative **re-storying** process. The researcher seeks to understand the **lived experience** of an individual or small group.

Narrative research is not designed to be an historical record but, rather, it is designed to understand the perspective of the storyteller in the context of his or her life.

Narrative research is not simply content based; it does not lend itself to the thematic approach, it does not focus on the analysis of elements of language, and **there are not clear rules** on analysis

as there are in grounded theory or phenomenology. Narrative analysis attempts to capture **individual representations of phenomena** that are event and experience based.

Phenomenological Research

A **phenomenological study** is designed to describe and interpret an experience by determining the meaning of the experience as perceived by the people who have participated in it. What is the experience of an activity or concept from the perspective of particular participants? That is the key question in phenomenology. **Phenomenology** addresses questions about common human experience.

The concept of **bracketing** is used in phenomenological research. Bracketing involves the researcher intentionally setting aside his or her own experiences, suspending his or her own beliefs in order to take a fresh perspective based on data collected from persons who have experienced the phenomenon. The bracketing or suspension of belief is also referred to as **epoche.**

From an analysis of the interview data, the researcher writes descriptions of the participants' experiences and how those experiences were perceived. From the analysis, the researcher derives an overall description of the general meaning of the experience. This is done through a process called **reduction.** Think of reduction as a way to reflect. It is a phenomenological device that aims to bring aspects of meaning into nearness or focus.

Other Types of Qualitative Research

Portraiture is a form of qualitative research that seeks to join science and art in an attempt to describe complex human experiences within an organizational culture. The "portrait" is shaped by the dialogue between the researcher (portraitist) and the subject and attempts to reveal the "essence" of the subject and to tell the "central story". Data can be collected using in-depth interviews and observations over a period of time, which typically result in a personal relationship between the researcher and participants.

Critical research seeks to empower change through examining and critiquing assumptions. Questions focus on power relationships and the influence of race, class, and gender. Whereas other forms of qualitative research described in this text have as a key purpose the understanding of a phenomenon and the meanings people attach to events, the purpose in critical research is to critique and challenge the status quo. Critical research may analyze texts or artifacts such as film or other communication forms such as drama or dance to reveal underlying assumptions. **Feminist research** and **participatory research** are sometimes classified as critical research.

Semiotics and **discourse analysis** study linguistic units to examine the relationship between words and their meanings. Texts or signs and their structural relationships are the subject of study for semiotics and there is no neutral text. These approaches stress the system of relations between words as a source of meaning and view language as a social construction.

Chapter 17: Analyzing and Reporting Qualitative Research

Qualitative analysis is messy and **nonlinear**. Data analysis in qualitative research is often done concurrently or simultaneously with data collection through an iterative, recursive, and dynamic process. Data collection, analysis, and report writing do not occur in distinct steps as is typical in quantitative studies.

The task of analyzing qualitative data can appear overwhelming but becomes manageable when broken down into key stages. Creswell (2007) describes the **data analysis spiral.** Once data are collected, they must be organized and managed. The researcher must become engaged with the data through reading and reflecting. Then data must be described, classified, and interpreted. Finally, the researcher represents or visualizes the data for others. Creswell describes how this spiral fits with various approaches to qualitative inquiry (narrative, phenomenology, grounded theory, ethnography, and case study).

There are three stages in analyzing the Qualitative Data: (1) *organizing and familiarizing,* (2) *coding and reducing*, and (3) *interpreting and representing*.

The first stage in analyzing qualitative data involves **familiarization** and **organization** so that the data can be easily retrieved. Initially, the researcher should become familiar with the data through reading and rereading notes and transcripts, viewing and reviewing videotapes, and listening repeatedly to audiotapes. The researcher must be immersed in the data.
As you are thus familiarizing yourself with the data, write notes or memos (also called a **reflective log**) to capture your thoughts as they occur. Notes may be taken in the margins of the transcripts indicating key ideas. Once you have made notes in the margins, review them and make a complete list of the different types of information you see. This is an essential preliminary step to developing a coding scheme.

Coding and Reducing

After familiarizing yourself with the data and organizing them for easy retrieval, you can begin the **coding** and **reducing** process. This is the ***core*** of qualitative analysis and includes the identification of categories and themes and their refinement.

Coding is about developing concepts from the raw data. The first step in coding is referred to as **axial coding, open coding, preliminary coding,** or **provisional coding.** The most common approach is to read and reread all the data and sort them by looking for units of meaning−words, phrases, sentences, subjects' ways of thinking, behavior patterns, and events that seem to appear regularly and that seem important.

The categories developed from the coded data should be internally consistent and distinct from one another. The researcher's interests and style and the research question influence to a great extent the categories chosen. **Organizational categories** typically could have been anticipated

and may have been established prior to data collection. However, these are not usually a good mechanism for making sense of the actual data. Substantive or theoretical categories help provide insights.

Substantive categories are primarily descriptive and not generally related to more abstract theories. ***Emic substantive categories*** are those from participants' perspectives and words. However, substantive categories are more likely to be based on the researcher's interpretation of what is going on (etic categories). **Theoretical categories** are more abstract and can be from prior theory or from inductively developed theory. They are more likely to be etic categories. Often, novice researchers use only organizational categories.

Perhaps the best known qualitative analysis strategy is the **constant comparative method,** which combines inductive category coding with simultaneous comparison of all units of meaning obtained. The researcher examines each new unit of meaning (topics or concepts) to determine its distinctive characteristics. Then he or she compares categories and groups them with similar categories.

Another approach used in analysis is the **negative case analysis** or **discrepant data analysis.** Look for data that are negative or discrepant from the main body of data collected.

Interpreting and Representing

Interpreting involves reflecting about the words and acts of the study's participants and abstracting important understandings from them. It is an **inductive process** in which you make generalizations based on the connections and common aspects among the categories and patterns. You may develop hypotheses that have evolved during the analysis. **Interpretation** is about bringing out the meaning, telling the story, providing an explanation, and developing plausible explanations.

Rigor in Qualitative Research

Credibility

Validity concerns the accuracy or truthfulness of the findings. The term most frequently used by qualitative researchers to refer to this characteristic is **credibility.** Credibility in qualitative research concerns the truthfulness of the inquiry's findings. Credibility or truth value involves how well the researcher has established **confidence in the findings** based on the research design, participants, and context. The term *credibility* in qualitative research is analogous to *internal validity* in quantitative research.

A number of methods have been identified in the literature for enhancing the credibility (internal validity) of qualitative studies. These methods may be categorized according to five types of evidence: structural corroboration, consensus, referential or interpretive adequacy, theoretical adequacy, and control of bias.

Evidence Based on Structural Corroboration

Eisner (1998) defines **structural corroboration** as a means through which multiple types of data are related to each other to support or contradict the interpretation and evaluation of a state of affairs" (p. 110). The use of multiple sources of data, multiple observers, and/or multiple methods is referred to as **triangulation.** Structural corroboration uses different sources of data (data triangulation) and different methods (methods triangulation).

Evidence Based on Consensus

Validity based on consensus is defined as "agreement among competent others that the description, interpretation, evaluation, and thematics" are correct (Eisner, 1998, p. 112). This type of validity is primarily demonstrated through two methods: *peer review* and *investigator triangulation*. In **peer review,** also called **peer debriefing,** the question is asked, "Given the evidence presented, is there consensus in the interpretation?" Colleagues or peers are provided with the raw data along with the researcher's interpretation or explanation. Discussions then determine whether the reviewer(s) considers the interpretation to be reasonable, given the evidence. **Investigator triangulation** involves having multiple researchers collect data independently and compare the collected data.

Evidence Based on Referential or Interpretive Adequacy

Referential or **interpretive** evidence of validity refers to "accurately portraying the meaning attached by participants to what is being studied by the researcher" and "the degree to which the participants' viewpoints, thoughts, feelings, intentions, and experiences are accurately understood . . . and portrayed" (Johnson & Christensen, 2000, p. 209).

Two primary strategies are used to enhance referential adequacy: ***member checks*** and low-***inference descriptors***.

Evidence Based on Theoretical Adequacy

Theoretical adequacy or **plausibility** concerns the degree to which a theoretical explanation developed from the study fits the data and is defensible. There are three key strategies for promoting theoretical adequacy: extended fieldwork, theory triangulation, and pattern matching.

Evidence Based on Control of Bias

Researcher bias is a source of invalidity in qualitative studies. Bias may result from selective observations, hearing only what one wants to hear, or allowing personal attitudes, preferences, and feelings to affect interpretation of data.

The most common strategy to control for bias in qualitative studies is reflexivity. **Reflexivity** is the use of self-reflection to recognize one's own biases and to actively seek them out. The researcher should refer to his or her journal reflections during the process of data analysis.

Another strategy used to control for bias is **negative case sampling,** in which researchers intentionally seek examples of the opposite of what they expect. To avoid the appearance of bias, researchers should show that they have searched for and explained any discrepant or contradictory data.

Approaches to Enhancing Credibility in Qualitative Studies	
Criterion	**Strategies**
Structural Corroboration	Methods triangulation Data triangulation
Consensus	Peer review/peer debriefing Investigator triangulation
Referential or interpretive adequacy	Member checks/participant feedback Low-inference descriptors/thick, rich description
Theoretical adequacy	Extended fieldwork Theory triangulation Interdisciplinary triangulation Pattern matching
Control of bias	Reflexivity Negative case sampling

Transferability (external validity)

Transferability is the degree to which the findings of a qualitative study can be applied or generalized to other contexts or to other groups. In quantitative research, the term *external validity* is used to refer to the generalizability of the findings.

Although the qualitative researcher typically does not have generalizability as a goal, it is his or her responsibility to provide sufficiently rich, **detailed, thick descriptions of the context** so that potential users can make the necessary comparisons and judgments about **similarity** and hence transferability. This is referred to as **descriptive adequacy.** The researcher must strive to provide accurate, detailed, and complete descriptions of the context and participants to assist the reader in determining transferability.

One strategy to enhance transferability is to include **cross-case comparisons.** The researcher may investigate more than one case. If findings are similar, this would increase the possibility of transferability of findings to others settings or contexts. In some cases, even a single case can be compared with other cases in the published literature that might demonstrate transferability.

Be aware that there are threats to transferability, such as **selection effects** (the fact that the constructs being investigated are unique to a single group), **setting effects** (the fact that results may be a function of the specific context under investigation), and **history effects** (the fact that unique historical experiences of the participants may militate (influence) against comparisons).

The researcher should recognize limitations of the study in the description. Detailing of circumstances helps the reader to understand the nature of the data and what might be peculiar to your particular study.

Reactivity (the effect of the research itself) might also limit transferability. Although eliminating the *influence of the researcher may be impossible* in a qualitative study because the researcher is the key data collection instrument, the researcher can help the reader understand the potential influence by describing his or her own biases **through a reflective statement** and providing detailed descriptions of such things as observation strategies and interview questions. Reactivity is a more serious threat in studies using interview techniques.

Approaches to Enhancing Transferability in Qualitative Studies	
Criterion	**Strategies**
Descriptive adequacy	Thick, rich description
Similarity	Cross-case comparisons Literature comparisons Describing limitations
Limiting Reactivity	Reflective statement Detailed description of methods

Dependability

Qualitative researchers speak of **dependability** rather than reliability. Unlike quantitative research, in which tight controls enhance replicability, qualitative studies expect variability because the context of studies changes. Thus, consistency is viewed as the extent to which variation can be tracked or explained. This is referred to as *dependability* or **trustworthiness.**

Some strategies to investigate dependability are using an ***audit trail***, ***replication logic***, ***stepwise replication***, ***code-recoding***, ***interrater comparisons***, and ***triangulation***. To enhance reliability, the researcher wants to demonstrate that the methods used are reproducible and consistent, that the approach and procedures used were appropriate for the context and can be documented.

Documentation

One of the best ways to establish dependability is to use an **audit trail.** Audit trails provide a mechanism by which others can determine how decisions were made and the uniqueness of the situation. It documents how the study was conducted, including ***what*** was done, ***when***, and ***why***. The audit trail contains the ***raw data*** gathered in interviews and observations, records of the inquirer's decisions about whom to interview or what to observe and why, files documenting how working hypotheses were developed from the raw data and subsequently refined and tested, the findings of the study, and so forth. A complete ***presentation of procedures and results***

enables the reader to make a judgment about the replicability of the research within the limits of the natural context.

Consistent Findings

Dependability can be demonstrated by showing consistent findings across multiple settings or multiple investigators. **Replication logic,** which involves conducting the study in multiple locations or with multiple groups, is suggested for determining dependability of a study. According to this logic, the more times a finding is found true with different sets of people or in different settings and time periods, the more confident the researcher can be in the conclusions. **Stepwise replication** is another technique suggested for enhancing dependability. In this strategy, two investigators divide the data, analyze it independently, and then compare results. Consistency of results provides evidence of dependability.

Coding Agreement

Intrarater and **interrater agreement** are strategies for assessing dependability (reliability). An intrarater method is the **code–recode strategy:** A researcher codes the data, leaves the analysis for a period of time, and then comes back and recodes the data and compares the two sets of coded materials.

Corroboration

Triangulation, which we have previously discussed, is also used to establish the dependability of qualitative studies. If multiple data sources or multiple methods result in similar findings, it enhances the reliability of the study.

Approaches to Enhancing Dependability in Qualitative Studies	
Criterion	**Strategies**
Documentation	Audit trail
Consistent Findings	Replication logic Stepwise replication
Coding Agreement	Code‑recode/intrarater agreement Interrater/interobserver agreement
Corroboration	Data triangulation Methods triangulation

Confirmability

Confirmability in qualitative research is the same as the quantitative researchers concept of objectivity. Both deal with the idea of **neutrality** or the extent to which the research is free of bias in the procedures and the interpretation of results.

Because it may be impossible to achieve the levels of objectivity that quantitative studies strive for, qualitative researchers are concerned with whether the data they collect and the conclusions they draw would be confirmed by others investigating the same situation. Thus, in qualitative

studies, the focus shifts from the neutrality of the researcher to the confirmability of the data and interpretations.

The ***audit trail*** is the main strategy for demonstrating confirmability. By providing a complete audit trail, the researcher enables another researcher to arrive or not arrive at the same conclusions given the same data and context. Other strategies used to enhance confirmability include **triangulation** of methods**, peer review**, and **reflexivity**−all discussed previously.

Approaches to Enhancing Confirmability in Qualitative Research

Criterion	**Strategy**
Documentation	Audit Trail
Corroboration	Triangulation Peer review
Control of Bias	Reactivity

Chapter 18: Action Research

Action research is about taking action based on research and researching the action taken. Action research is based on the premise that local conditions vary widely and that the solutions to many problems cannot be found in generalized truths that take no account of local conditions.

There are three main characteristics of action research:

1. The research is situated in a local context and focused on a local issue.

2. The research is conducted by and for the practitioner.

3. The research results in an action or a change implemented by the practitioner in the context.

Today, action research has gained popularity in the United States and elsewhere and is seen as important in the work of **improving schools**.

Approaches to Action Research

Collaborative action research

It involves multiple researchers. In education, this may include school and university personnel or teachers and school administrators. Its main purpose is to share expertise and foster dialogue among stakeholders.

Critical action research

It involves wide collaboration. In education, this may include university researchers, school administrators, teachers, and community members. Its main objective is to evaluate social issues and use the results for social change.

Classroom action research

It Involves teachers in their classrooms; it can involve groups of teachers examining common issues. Its main purpose is to improve classroom practice or to improve practices in the school.

Participatory action research

It involves collaboration among stakeholders in a social process. Its main purpose is to explore practices within social structures (emancipatory); to challenge power differences and unproductive ways of working (critical); and to change theory and practice (transformational)

Strategies for Identifying the Problem

Reflection

Reflection is one strategy for identifying problems. Think about your own setting and consider what is working well and what might need improvement.

Description

Description is another strategy for determining and focusing on the problem to be investigated. Insights can be gained by describing the who, what, when, where, how, and why of a situation. These descriptions come from observations.

Literature Review

Conducting a limited **literature review** can also help in developing your explanation and clarifying the research question. Reviewing the literature helps in assessing what, if anything, other researchers have found out about the topic and what theoretical perspectives relate to the topic, as well as providing promising practices.

Brainstorming

Johnson (2008) advises that if all else fails, simply **brainstorm** by drawing a line down the center of a blank sheet of paper and listing on the left side any topics of interest that come to mind. Then talk to others about some of these ideas and continue to develop the list. Once you have the topic list, on the right side begin to list specific questions for each topic.

Data Collection for Action Research

In action research, as with other types of research, different research questions require different research approaches. Both quantitative and qualitative approaches may be used in action research, and one approach is not better than the other.

Triangulation is important in action research. Using multiple sources of data and avoiding reliance on a single source enhances corroboration of the findings.

Data Collection Strategies

Three types of data are gathered in action research commonly known as the *three E's*: **experiencing, enquiring,** and **examining.**

Experiencing

First, data may be gathered through the researcher's own experience. This category focuses on observational data that may be recorded in various ways. **Field notes** are the most common data collection strategy used in action research to provide a record of what is going on during an observation. Field notes can include descriptions of places (locations, physical layouts, etc.), people (individuals, types, positions, etc.), objects (buildings, furniture, equipment, materials, etc.), acts (single actions that people take), activities (sets of related acts), events (sets of related activities), purposes (what people are trying to accomplish), time (times, frequency, duration, sequencing, etc.), and feelings (emotional orientations and responses).

Enquiring

Second, data may be collected by asking participants to respond in some manner– that is, enquiring of them. The most common action research strategy for collecting enquiring data is through interviews. During the first phase of study, **grand tour questions** that are global allow participants to describe something in their own terms. ("Tell me about your school?") s

In the second phase, **extension questions** or mini-tour questions ask for more detail. ("Can you tell me more about that?").

In the third phase, **prompt questions** are used so that more details are revealed.

Examining

Third, data may be collected through examining artifacts and other materials that already exist or that are routinely collected in the setting. Student records and teacher records are useful sources of information.

Rigor in Action Research

Action researchers should be concerned about the issues of **rigor** or quality addressed by other researchers: validity, credibility, reliability, dependability, neutrality, confirmability, and transferability. These concepts are covered in other chapters of this text. There are a few comments about rigor in action research, however, worth noting here.

Action research in schools often relies on **authentic student work,** which Sagor (2000) compares to primary source materials and claims enhances credibility.

Credibility is described as the researcher's ability to take into account the complexities that present themselves in a particular setting and to deal with patterns not easily explained.

Being able to generalize is not a primary goal of action research; rather, the primary goal is to understand what is happening in a **specific context** and to determine what might improve things in that context. Action researchers believe that everything is context bound and that the goal is not to develop a generalizable statement but to **provide rich and detailed descriptions of the context** so that others can make comparisons with their contexts and judge for themselves whether the findings might apply (be transferable).

Data Analysis in Action Research

Coding

One key analysis strategy often described in action research is **coding** as typically described in qualitative research. First, the researcher breaks down and categorizes the data into manageable segments **(open coding).** Then, the researcher puts the data back together again, making connections between and across categories **(axial coding).** Sometimes, the researcher has a clear and selective focus and is systematically reviewing the data for that specific category **(selective coding).**

Stages of Analysis

There are two stages of action research analysis, description and **sense making.** During the description stage, you ***review the data*** and ask yourself what did you see and what was happening. During the sense-making stage, you try to consider how the pieces fit together and what stands out.

Data Interpretation in Action Research

Data interpretation focuses on the implications or meanings that emerge from the analysis. Interpretation is used to help make the experiences being studied understandable, using description and conceptual frameworks or theories.

Using Visuals

Concept mapping can be used to plot elements diagrammatically so you can visualize what different components of the situation relate to the problem under investigation. **Problem analysis** using visuals of antecedents and consequences can also be helpful in interpretation.

Reflecting

The interpretation phase of action research is a process of ***ongoing reflection*** and is the most challenging aspect of action research. The researcher continuously reviews the data as the action research process unfolds, remembering that any interpretations reached and conclusions arrived at are not for all time, are not generalizable, and are certainly not conclusive.

Data interpretation in action research is about making educated guesses or reasonable inferences. Once drawn, the interpretation can be connected with personal experience and contextualized. The interpretation provides a rationale for action planning. After interpretation, the researcher must decide what the implications are for practice.

Action Plan

The most important step in action research follows analysis and interpretation. That step is acting on the knowledge you have gained. What do you believe is an effective choice or course of action based on what you now know? What will you do differently? Did you discover a new problem? Does something need modification? But before you act, you must develop an **action plan.**

Given what is known from the research, the researcher must determine what precisely to do, what is the course of action. This step also returns the researcher to the problem formulation step.

Chapter 19: Mixed Methods Research

Classifying Mixed Methods

In the research field, there is continuing discussion regarding clarification of mixed methods research in relationship to *monomethod* research, *multimethod* research, and *mixed model* research.

In a **monomethod research** design, one method, either qualitative or quantitative, with corresponding data collection, analysis, and accompanying procedures, is used to answer the research question. **Multimethod research** employs different types of data collecting methods– for example, both survey and archival data. Multimethod research occurs when the research questions are investigated by using two different data collection procedures (e.g., observations and focus groups) or by combining two research methods (critical theory, grounded theory, or case study) from the same research tradition (qualitative or quantitative).