CHAPTER 3
DEFINING AND MEASURING VARIABLES

CHAPTER OUTLINE

3.1 CONSTRUCTS AND OPERATIONAL DEFINITIONS
Theories and Constructs
Operational Definitions
Limitations of Operational Definitions
Using Operational Definitions

3.2 VALIDITY AND RELIABILITY OF MEASUREMENT
Consistency of a Relationship
Validity of Measurement
  Face validity
  Concurrent validity
  Predictive validity
  Construct validity
  Convergent and divergent validity
Reliability of Measurement
  Types and measures of reliability
The Relationship between Reliability and Validity

3.3 SCALES OF MEASUREMENT
The Nominal Scale
The Ordinal Scale
Interval and Ratio Scales
  Dealing with equivocal measurements
Selecting a Scale of Measurement

3.4 MODALITIES OF MEASUREMENT
Self-Report Measures
Physiological Measures
Behavioral Measures

3.5 OTHER ASPECTS OF MEASUREMENT
Multiple Measures
Sensitivity and Range Effects
Artifacts: Experimenter Bias and Participant Reactivity
  Experimenter bias
  Demand characteristics and participant reactivity

Selecting a Measurement Procedure

KEY WORDS

theory
constructs or hypothetical constructs
operational definition
validity
face validity
concurrent validity
predictive validity
construct validity
convergent validity
divergent validity
reliability
test-retest reliability
parallel-forms reliability
inter-rater reliability
split-half reliability
ceiling effect
floor effect
experimenter bias
single-blind research
double-blind research
demand characteristics
reactivity
good subject role
negativistic subject role
apprehensive subject role
faithful subject role
laboratory
field

LEARNING OBJECTIVES AND CHAPTER SUMMARY

1. Define a *construct* and explain the role that constructs play in theories.

> Constructs are hypothetical attributes or mechanisms that help explain and predict behavior in a theory.

2. Define an *operational definition* and explain the purpose and the limitations of operational definitions.

> An operational definition is a procedure for indirectly measuring and defining a variable that cannot be observed or measured directly. An operational definition specifies a measurement procedure (a set of operations) for measuring an external, observable behavior, and uses the resulting measurements as a definition and a measurement of the hypothetical construct. Although operational definitions are necessary to convert an abstract variable into a concrete entity that can be observed and studied, you should keep in mind that an operational definition is not the same as the construct itself.

3. Define a *positive relationship* and a *negative relationship* and explain how the consistency of positive and negative relationships can be used to establish validity and reliability.

> A positive relationship is a relationship in which the two variables or measurements tend to change together in the same direction. A negative relationship is a relationship in which the two variables or measurements tend to change together in opposite directions. Note that the reliability or validity of a measurement procedure is usually established with a consistent positive or a consistent negative relationship, depending on how the variables are defined and measured.

4. Define the *validity of measurement* and explain why and how it is measured.

Validity of measurement is the degree to which the measurement process measures the variable it claims to measure. Researchers have developed several methods for assessing the validity of measurement. Six of the more commonly used definitions of validity are: face validity, concurrent validity, predictive validity, construct validity, convergent and divergent validity.

5. Define the *reliability of measurement* and explain why and how it is measured.

The reliability of a measurement procedure is the stability or consistency of the measurement. If the same individuals are measured under the same conditions, a reliable measurement procedure produces identical (or nearly identical) measurements.

6. Compare and contrast the four scales of measurement (nominal, ordinal, interval, and ratio) and identify examples of each.

The categories that make up a nominal scale simply represent qualitative (not quantitative) differences in the variable measured. For example, if you were measuring academic majors for a group of college students, the categories would be art, chemistry, English, history, psychology, and so on. The categories that make up an ordinal scale have different names and are organized sequentially. Often, an ordinal scale consists of a series of ranks (first, second, third, and so on) like the order of finish in a horse race. The categories on interval and ratio scales are organized sequentially, and all categories are the same size. Thus, the scale of measurement consists of a series of equal intervals like the inches on a ruler. Other common examples of interval or ratio scales are the measures of time in seconds, weight in pounds, and temperature in degrees Fahrenheit.

7. Identify the three modalities of measurement and explain the strengths and weaknesses of each.

The many different external expressions of a construct are traditionally classified into three categories that also define three different types, or modalities, of measurement. The three categories are self-report, physiological, and behavioral. Consider, for example, the hypothetical construct “fear,” and suppose that a researcher would like to evaluate the effectiveness of a therapy program designed to reduce the fear of flying. This researcher must somehow obtain measurements of fear before the therapy begins, then compare them with measurements of fear obtained after therapy. Although fear is an internal construct that cannot be observed directly, it is possible to observe and measure external expressions of fear. For example, an individual may claim to be afraid (self-report), may have an increased heart rate (physiological), or may refuse to travel on an airplane (behavioral). One major decision in developing a measurement procedure (an operational definition) is to determine which type of external expression should be used to define and measure fear.

8. Define a *ceiling effect* and a *floor effect* and explain how they can interfere with measurement.

Ceiling and floor effects are both range effects. When the range is restricted at the high end, the problem is called a ceiling effect (the measurements bump into a ceiling and can go no higher). Similarly, clustering at the low end of the scale can produce a floor effect. In general, range effects suggest a basic incompatibility between the measurement procedure and the individuals measured. Often, the measurement is based on a task that is too easy (thereby producing high scores) or too difficult (thereby producing low scores) for the participants being tested.

9. Define an *artifact* and explain how examples of artifacts (experimenter bias, demand characteristics, and reactivity) can threaten both the validity and reliability of measurement and how they can influence the results of a research study.

An artifact is a non-natural feature accidentally introduced into something being observed. In the context of a research study, an artifact is an external factor that may influence or distort the measurements. For example, a doctor who startles you with an ice-cold stethoscope is probably not going to get accurate observations of your heartbeat. An artifact can threaten the validity of the measurements because you are not really measuring what you intended, and it can be a threat to reliability.

Experimenter bias occurs when the measurements obtained in a study are influenced by the experimenter’s expectations or personal beliefs regarding the outcome of the study. The term *demand characteristic* refers to any of the potential cues or features of a study that (1) suggest to the participants what the purpose and hypothesis is, and (2) influence the participants to respond or behave in a certain way. Reactivity occurs when participants modify their natural behavior in response to the fact that they are participating in a research study or the knowledge that they are being measured.

OTHER LECTURE SUGGESTIONS

1. Although students are familiar with the concept of measurement, they typically do not have a generalized definition of the process. Measurement involves a systematic procedure for assigning individuals to categories. The procedure can be explained to others who can then repeat the measurements, and the categories make up a scale of measurement.

2. The first general theme for this chapter is that there are multiple ways to measure any particular variable. Often there are many different ways to operationally define a variable. Then you must select a scale of measurement, a modality of measurement, and units of measurement, all of which may have an impact on the outcome of the study.

3. The concepts of (a) hypothetical construct, (b) operational definition, and (c) reliability and validity are all intertwined. Because a hypothetical construct cannot be observed or measured directly, you must devise some indirect method for measurement; that is, you need an operational definition. However, because the measurement is indirect, you have an obligation to show that the measurement procedure is accurate and consistent (valid and reliable).

4. If a variable is relatively stable over time, a measurement procedure that is not reliable is also not valid. For example, intelligence does not change dramatically from day to day. Therefore, if your measurements of intelligence do change (unreliable) then you are not really measuring intelligence (not valid).

5. In addition to the activities presented at the end of the chapter, the following can be used as an in-class activity for this chapter.

   The classic example of an unreliable measurement is a rubber ruler. You can reproduce this classic by cutting a rubber band to form a string about 6 inches long. Then use a marker to draw 12 “one-inch” units on band (the actual marks will be about ½ inch apart). Next, draw two lines about 6 feet apart on the floor or a blackboard. Have students stretch the band until it looks like one foot and then use it to measure the distance between the lines. Repeated measurements should produce different numbers, indicating an unreliable measurement.

NOTES ON END-OF-CHAPTER EXERCISES

1. In addition to the key words, you should also be able to define each of the following terms:

Positive relationship: A positive relationship is a relationship in which the two variables or measurements tend to change together in the same direction.

Negative relationship: A negative relationship is a relationship in which the two variables or measurements tend to change together in opposite directions.

Accuracy: The degree to which a measure conforms to the established standard.

Scale of measurement: The set of categories used for classification of individuals. The four types of measurement scales are nominal, ordinal, interval, and ratio.

Nominal scale: A scale of measurement in which the categories represent qualitative differences in the variable being measured. The categories have different names but are not related to each other in any systematic way.

Ordinal scale: A scale of measurement on which the categories have different names and are organized sequentially (e.g., first, second, third).

Interval scale: A scale of measurement in which the categories are organized sequentially and all categories are the same size. The zero point of an interval scale is arbitrary and does not indicate a total absence of the variable being measured.

Ratio scale: A scale of measurement in which the categories are sequentially organized, all categories are the same size, and the zero point is absolute or nonarbitrary, and indicates a complete absence of the variable being measured.

Self-report measure: A measurement obtained by asking a participant to describe his or her own attitude, opinion, or behavior.

Physiological measure: Measurement obtained by recording a physiological activity such as heart rate.

Behavioral measure: A measurement obtained by the direct observation of an individual's behavior.

Range effect: The clustering of scores at one end of a measurement scale. Ceiling effects and floor effects are types of range effects.

Artifact: In the context of a research study, an external factor that could influence or distort measures. Artifacts threaten the validity of the measurement, as well as both internal and external validity.

Subject roles or subject role behaviors: The different ways that participants respond to experimental cues based on whatever they judge to be appropriate in the situation. Also known as subject role behavior.

2. (LO1 and 2) Hypothetical concepts, such as *honesty*, are variables that cannot be observed or measured directly and, therefore, require operational definitions.
(a) Describe one procedure that might be used to measure honesty.
(b) Use the procedure you described in (a) to explain why there may not be a one-to-one relationship between the actual variable and the procedure by the operational definition of the variable.

Student answers will vary. The idea is for them to learn to operationalize the variable.

3. (LO2) Briefly explain what an operational definition is and why operational definitions are sometimes necessary.

An operational definition is a procedure for indirectly measuring and defining a variable that cannot be observed or measured directly. An operational definition specifies a measurement procedure (a set of operations) for measuring an external, observable behavior and uses the resulting measurements as a definition and a measurement of the hypothetical construct.

4. (LO4) A researcher evaluates a new cholesterol medication by measuring cholesterol levels for a group of patients before they begin taking the medication and after they have been taking the medication for eight weeks. A second researcher measures quality of life for a group of 40-year-old men who have been married for at least 5 years and a group of 40-year-old men who are still single. Explain why the first researcher is probably not concerned about the validity of measurement, whereas the second researcher probably is. (Hint: What variable is each researcher measuring and how will it be measured?)

The first researcher is using a physiological variable, and thus is not concerned about the validity because cholesterol level is not an abstract concept. The second researcher is measuring a very abstract concept (“quality of life”) and thus needs to operationalize it in a valid way.

5. (LO3 and 4) A clinical researcher has developed a new test measuring impulsiveness and would like to determine the validity of the test. The new test and an established measure of impulsiveness are both administered to a sample of participants. Describe the pattern of results that would establish concurrent validity for the new test.

This researcher would hope to find a positive relationship between the two measures. That is, as the scores on the new social anxiety test increase, so too should the scores on the existing test increase.

6. (LO5) Suppose that a social scientist has developed a questionnaire intended to measure the quality of romantic relationships. Describe how you could evaluate the reliability of the questionnaire.

The questionnaire would be evaluated using test-retest reliability analysis and split-half reliability analysis. Test-retest reliability is established by comparing the scores obtained from two successive measurements of the same individuals and calculating a correlation between the two sets of scores. Split-half reliability is obtained by splitting the items on a questionnaire or test in half, computing a separate score for each half, and then measuring the degree of consistency between the two scores for a group of participants.

7. (LO5) Explain how inter-rater reliability is established.

When measurements are obtained by direct observation of behaviors, it is common to use two or more separate observers who simultaneously record measurements. For example, two psychologists may watch a group of preschool children and observe social behaviors. Each individual records (measures) what she observes, and the degree of agreement between the two observers is called inter-rater reliability. Inter-rater reliability can be measured by computing the correlation between the scores from the two observers, or by computing a percentage of agreement between the two observers.

8. (LO4 and 5) A researcher claims that intelligence can be measured by measuring the length of a person’s right-hand ring finger. Explain why this procedure is very reliable but probably not valid.

A person’s right-hand ring finger is unlikely to change much over time, so this measure would be considered reliable. However, it is not valid. Intelligence is an abstract concept that cannot be operationalized by something as arbitrary as the length of one’s finger.

9. (LO4 and 5) For each of the following operational definitions, decide whether you consider it to be a valid measure. Explain why or why not. Decide whether you consider it to be a reliable measure. Explain why or why not.

(a) A researcher defines social anxiety in terms of the number of minutes before a child begins to interact with adults other than his or her parents.

This operational definition is not valid because other factors may affect the child’s decision not to interact with other adults; for example, the child might be sleepy or daydreaming, might not consider the adults to be very interesting, etc. This operational definition may not be reliable either since children may feel more or less inclined to interact with other adults on various occasions (depending on the factor(s) that influence their decision).

(b) A professor classifies students as either introverted or extroverted based on the number of questions each individual asks during one week of class.

> This operational definition is not valid because other factors could influence a student's decision to ask a question in class (e.g., an extroverted person may understand everything and not feel the need to ask a question). It is also likely not reliable either since students may ask a lot of questions the first week in one class and no questions in another, or they may ask more or fewer questions as the weeks go on based on their understanding of the material.

(c) A sports psychologist measures physical fitness by measuring how high each person can jump.

> This operational definition lacks validity because jumping ability is only one measure of physical fitness. However, it is reliable since a person's jumping ability is not likely to change much from one day to the next (although it will likely change over the course of a person's lifetime).

(d) Reasoning that bigger brains require bigger heads, a researcher measures intelligence by measuring the circumference of each person's head (just above the ears).

> This operational definition is not valid because the size of a person's skull is not directly linked to intelligence. This definition is reliable though, because a person's head size is not likely to change from one day to the next.

10. (LO6) In this chapter, we identified four scales of measurement: nominal, ordinal, interval, and ratio.

(a) What additional information is obtained from measurements on an ordinal scale compared to measurements from a nominal scale?

> Whereas a nominal scale tells us if a difference exists, an ordinal scale will tell us the direction of that difference.

(b) What additional information is obtained from measurements on an interval scale compared to measurements from an ordinal scale?

> Whereas an ordinal scale will tell us if a difference exists and the direction of that difference, an interval scale will tell us information about the magnitude of difference.

(c) What additional information is obtained from measurements on a ratio scale compared to measurements from an interval scale?

> Whereas an interval scale will tell us if a difference exists, the direction of that different, and the magnitude of that difference, a ratio scale will allow us to measure the absolute amount of a variable and compare measurements in terms of ratios.

11. Select one construct from the following list:

happiness
exhaustion
creativity
hunger
motivation
fear

Briefly describe how it might be measured using:
(a) (LO2 and 7) an operational definition based on self-report (e.g., a questionnaire)

> Participants could be asked if they are happy/exhausted/motivated/etc., or could rate their level of happiness/exhaustion/motivation/etc. on a scale of 1 to 10.

(b) (LO2 and 7) an operational definition based on behavior (e.g., what kinds of behavior would you expect to see from an exhausted individual?)

> Participants could be observed acting in ways that indicate happiness (e.g., smiling), exhaustion (e.g., frequent yawning), etc.

12. (LO7) Describe the relative strengths and weaknesses of self-report measures compared to behavioral measures.

> The advantage of self-report measures is that they are probably the most direct way to assess a construct. The disadvantage is that participants could lie or be influenced by the researcher or the phrasing of the questions, which undermines the validity of the data. Behavioral measures give researchers a wide range of options regarding what to observe, however, one problem is that a particular participant behavior may be temporary.

13. (LO8) What is a ceiling effect, and how can it be a problem?

> A ceiling effect is the clustering of scores at the high end of a measurement scale, allowing little or no possibility of increases in value. It prevents participants from scoring any higher on a scale, which skews the measurement.

14. (LO9) Explain how an artifact can limit the validity and reliability of a measurement.

> An artifact is a non-natural feature accidentally introduced into the phenomenon being observed. It can influence or distort measurements, which can affect validity (because you may no longer be measuring what you want to measure) and reliability (because your measurements may no longer be consistent).

15. (LO9) What are demand characteristics, and how do they limit the validity of the measurements obtained in a research study?

The term *demand characteristic* refers to any of the potential cues or features of a study that (1) suggest to the participants what the purpose and hypothesis is, and (2) influence the participants to respond or behave in a certain way. They can change participants' normal behavior and thereby influence the measurements they produce.

16. (LO9) Describe how the concept of participant reactivity might explain why a person's behavior during a job interview is very different from a person's behavior with friends.

Reactivity occurs when participants modify their natural behavior in response to the fact that they are participating in a research study or the knowledge that they are being measured. A person in a job interview may act the way he or she thinks the potential employer would expect, but that may not be an accurate reflection of how the person behaves when relaxing with friends.

WEB RESOURCES

Scales of Measurement in Statistics from Stat Trek:
http://stattrek.com/statistics/measurement-scales.aspx

Standardized tests: Test reliability from UC Davis:
http://psychology.ucdavis.edu/faculty_sites/sommerb/sommerdemo/stantests/test_rel.htm

What Is Validity? From Simply Psychology:
http://www.simplypsychology.org/validity.html

Operationalization from Explorable Psychology Experiments:
https://explorable.com/operationalization