



CHAPTER ONE

INTRODUCTION TO MEASUREMENT

LEARNING OBJECTIVES

At the end of this chapter, the student should be able to

1. Give two reasons for taking a measurement course.
2. Describe major historical events in the development of measurement theory.
3. List three measurement skills that public health students can use in their careers.
4. Define the term *measurement*.
5. List four advantages of measurement.
6. Compare and contrast the four levels of measurement.
7. Define the terms *reliability* and *validity*.

Role of Measurement in Health Education and Health Behavior Research

Public health students are often surprised to learn that measurement of health behaviors requires special study. Students often associate measurement theory with research and sometimes question its applicability to the everyday work of health professionals. Yet a close examination of the work of health professionals reveals the pervasiveness of measurement issues. Consider, for example, Nelini, who is employed as a health

educator by a large public hospital. Nelini develops and conducts self-management classes for adults with epilepsy. She knows that the people who attend the classes enjoy them, because they tell her so. But the hospital is faced with budget cuts, and the hospital administrator has asked Nelini and others to justify the existence of their programs. Nelini knows that a successful program would make the people who attend her classes more comfortable and confident in dealing with issues that surround living with epilepsy and would encourage them to take their medications on a regular basis. They might have fewer seizures or fewer side effects of seizures as a result of taking their medications properly and living a healthy life. To demonstrate that her program is effective, Nelini must evaluate it. This evaluation is likely to include a survey measuring self-management skills and important outcomes associated with self-management.

Other graduates of health education and behavioral sciences programs have similar experiences. Take Sara, a new graduate with a master's degree in public health (MPH) hired recently by a large research firm. In the few short months since she started working for the company, her assignments have varied, and some involve an understanding of measurement. For instance, she recently consulted with a group of researchers in California on how to use focus groups to generate items for a research questionnaire on smoking behaviors of middle school adolescents. She assisted in the development of the focus group interview guide and the writing of items for the questionnaire; both tasks required an understanding of measurement principles.

Matt works at a university school of public health as the project director of a federally funded study. The purpose of the study is to evaluate an intervention designed to increase physical activity among older men and women. Matt's responsibilities include locating self-report measures of exercise and physical activity appropriate for people over sixty-five years of age who live in both urban and rural areas, and assessing the feasibility of those measures for the current study. Matt must also locate biobehavioral measures of strength and endurance that will be used to augment the self-report measures. During the course of the study, Matt will be involved in the preliminary studies to assess the psychometric properties of the instruments to be used in it. While the main study is conducted, he will monitor the collection of data using the measures and ensure that the data are collected according to protocol.

Rao, a graduate of an MPH program with an emphasis in global health, is also applying his knowledge of measurement, as a member of the HIV task force in India. Street youth attend the HIV prevention program he has developed; classes are held in an open field next to a temple. During the past year, he has expanded the program to include other services such as vocational and literacy training. Currently, Rao is attempting to locate additional funding for these new programs, and potential funding agencies have requested information about the effectiveness of the programs. Rao's task is to decide what outcomes of the program are important to measure and how to measure them. In conducting an assessment of outcomes, he faces special challenges

because the children neither read nor write and cultural differences preclude the use of questionnaires or interviews developed for other populations.

These examples highlight the variety of measurement skills used by public health professionals. They also demonstrate the importance of exposure to measurement issues in training programs, which creates an advantage for students when they take on work assignments. In summary, then, a course in measurement can introduce students in public health to the basic principles of measurement, including concept analysis, item writing and analysis, data collection procedures, and assessment of measures for use in research or evaluation studies.

Brief History of Psychosocial Measurement

The history of measurement dates back to the first man or woman who devised a system to count objects. Unfortunately, we do not know how or why this conceptual leap occurred, but it defined a monumental transition point in human history. Over time, people invented procedures to measure length and weight, and fields of study (for example, astronomy and physics) evolved that required various forms of measurement. Theories of measurement and quantity were also proposed by philosophers and mathematicians, including Aristotle and Euclid (Michell, 1990). The study and assessment of human capabilities have a long history as well. No doubt our early ancestors devised methods to evaluate and reward physical characteristics that were important in everyday and military life. In ancient Greece, for example, the Olympic Games celebrated athletic achievements in human endurance and strength through public competition (Crane, 2004).

Recorded history of the measurement of human capabilities, however, can be traced back only 3,000 years, to ancient China. In the second century B.C., China used a system of examinations to assess eligibility for government positions and for retention and promotion in those positions. (See DuBois, 1970, for a detailed discussion of the history of psychological testing. Major events addressed in that book are described here.) Early examinations included assessments of archery, music, and writing; later ones assessed civil law, military affairs, and geography. This system of examinations, with modifications, continued in China until the early twentieth century. Great Britain adopted a similar system of examinations for government employees in the 1830s, and by the end of the nineteenth century, the U.S. Congress had established competitive testing for entry into government service in the United States.

Though China is credited with initiating testing for government employees, European universities and schools were the first to implement testing as a means of evaluating student learning. The earliest testing seems to have occurred in the Middle Ages, with the earliest formal examinations in 1219 at the University of Bologna

(DuBois, 1970). Teachers conducted oral examinations until the introduction of paper, after which written examinations predominated. Schools and universities in the American colonies adopted student testing as a means of assessing student learning and eligibility for degrees, and this testing continues today in most U.S. educational institutions.

Whereas today's civil service and academic examinations provide a foundation for the measurement of knowledge and professional skills, the work of European and American psychologists in the late nineteenth century furthered the study of measurement of individual differences, attitudes, and behaviors. Two British citizens, Sir Francis Galton and Karl Pearson, were among the first to work in the area of psychometric testing. Galton introduced the study of human differences in psychological functioning, the basic understanding of correlation between variables, and the development of the questionnaire for psychological research (DuBois, 1970). Pearson expanded Galton's work on correlation and developed the product moment correlation along with the chi-square goodness-of-fit test to examine the relationships between variables.

The initial work in the area of psychological measurement focused on individual differences in tests of motor and sensory functioning. These tests included reaction times for identifying sounds and naming colors, identification of least noticeable differences in weight, and tactile perception of touch and pain. The field of psychology experienced a dramatic shift in the late nineteenth century with the work of Alfred Binet (Binet & Simon, 1973). Binet recognized that tests involving complex mental processes might be more useful than the usual motor and sensory tests in distinguishing the important characteristics of individuals. His work focused on the development of tests to measure psychological processes such as memory, problem solving, judgment, and comprehension. Through careful testing, he established that success on the tests varied with the age of a child, the child's school year, and even his or her class standing as determined by the teacher. In 1905, Binet and Simon introduced the first intelligence test, which was a composite of individual tests. Used to develop norms for different age groups of children and for adults, and used widely in clinical evaluations, these early forays into more complicated testing inspired the development of other forms of mental assessment.

By the time the United States entered World War I, in 1917, a number of psychological tests were available, and psychologists expressed an interest in assisting in the evaluation of recruits. Robert Yerkes served as chairperson of the Psychology Committee of the National Research Council, whose primary objective was the psychological assessment of U.S. Army recruits (DuBois, 1970). Because of the overwhelming task of conducting individual assessments for all recruits, the committee decided to develop a group test. Their efforts resulted in the Army Alpha, the first psychological test administered on a large-scale basis using a group format. The Army used the assessment throughout the war, and by the end of the fighting, the testing program

had opened a new avenue for the administration of tests and had expanded psychological testing into other areas. Interestingly, the Army Alpha used multiple-choice items, the success of which led to the adoption of multiple-choice items among educational measurement specialists.

In addition to measures of intelligence, psychologists developed measures of other psychological qualities. In the first quarter of the twentieth century, E. L. Thorndike, a noted educational measurement specialist, studied the interests and perceived abilities of students in academic subjects. (See, for example, Thorndike, 1913, 1920.) His work is particularly interesting because these measures incorporated ratings of degree of interest or perceived abilities. In 1922, Freyd published a paper describing the study of vocational interests. To assess interests in a variety of occupations, Freyd used a five-point response scale (*L*!, *L*, *?*, *D*, *D*!) in which *L* represented *like* and *D* represented *dislike*. In 1923, Freyd described a type of graphic rating scale that came to be known as a visual analog scale. Other tests measured skills, aptitudes, and personality traits such as neuroticism, creativity, introversion/extroversion, and anxiety. Approaches to the development of scales were devised by several individuals including Thurstone (1925), Likert (1932), and Guttman (1944). These measures and others like them formed the basis for the later development and assessment of instruments designed to measure health behaviors.

The study of measures themselves was also a major research focus during the twentieth century. In the early 1900s, Spearman (1904) introduced the concept of *reliability* to refer to the consistency of a measure. Later he developed a procedure to compute the internal consistency of a measure and coined the term *reliability coefficient* (Spearman, 1910). Through the use of correlation, investigators were able to compare new measures to selected criteria or other, related concepts. In 1937, Kuder and Richardson published a method for calculating the internal-consistency reliability of dichotomous measures, and in 1951, Cronbach, building on their work and that of Spearman (1910) and Brown (1910), devised a general method for calculating internal consistency.

Advances in computer technology enabled significant advances to be made in psychometrics in the latter half of the twentieth century. Although researchers developed the basic ideas and calculations for complex techniques such as factor analysis and multidimensional scaling in the early part of the century, personnel and time demands hindered widespread use of these techniques. With the advent of personal computers and the software for them, researchers began to use these and other techniques in the assessment of instruments, further advancing the understanding of instrument development and testing.

In addition to developments in the availability of computers, advances were made in the theoretical approaches to instrument development and testing. The most important contributions were confirmatory factor analysis using structural equation modeling (SEM) (Jöreskog, 1969; James, Mulaik, & Brett, 1982), item response theory (IRT)

(Lord, 1953), and generalizability theory (G theory) (Cronbach, Rajaratnam, & Gleser, 1963). Before 1950, classical test theory (CTT) constituted the basis of instrument development and testing. Though useful, this approach proved inadequate to solve some complex measurement problems. Therefore, alternative approaches to the conceptualization of measurement, along with new evaluation techniques, evolved during the latter part of the century. These newer ideas and techniques are referred to as modern test theory.

Conceptualization of Measurement

As noted earlier, the study of measurement and quantity has a long history, which dates back to Aristotle (Michell, 1990). Until the 1930s, ideas about measurement that had been developed over time were applied to psychological measurement. However, in 1932, the British Association for the Advancement of Science formed a committee to examine psychophysical measurement. And according to Michell (1990), that inquiry precipitated the development of new approaches to understanding psychological measurement, though some of them were not very useful for psychology. S. S. Stevens (1946) proposed one conceptualization of measurement. And although Michell sharply criticizes Stevens's conceptualization, Stevens's ideas have been used extensively in understanding psychological measurement for the past fifty years. Because Stevens's view is so pervasive, the following discussion on measurement is based on his work. The interested student is encouraged to read Michell (1990) for alternative views on the conceptualization of measurement.

A good place to begin the study of measurement is to define the term *measurement*. Stevens (1946) proposed that measurement is the “assignment of numerals to objects or events according to rule.” He later modified his definition to this one: “[M]easurement is the assignment of numerals to aspects of objects or events according to rule[s]” (Stevens, 1959, p. 24). This qualification of measurement was important, because it states that measurement applies to the attributes (properties) of objects rather than to the objects themselves. Limiting the application of measurement to objects alone allows only the counting of objects—for example, determining the number of people in a room. But applying measurement to attributes of objects allows the measurement of numerous properties—such as length, height, and weight. In public health the objects most often studied are people, communities, and systems. So aspects of these objects might include, for people, illnesses, health behaviors, and attitudes; for communities, size, location (urban or rural), air quality, and cultural norms; and for systems, system access (for example, access to care), policies, cost, and quality.

Using rules to assign numbers to attributes of objects implies that the process of assignment must be standardized and clear. The rules for using common measuring devices such as tape measures, bathroom scales, and thermometers are generally clear.

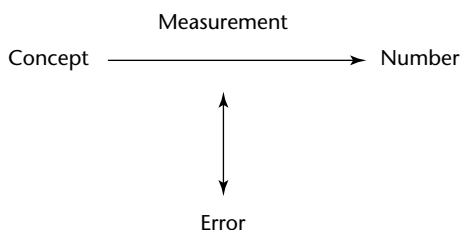
However, instruments measuring constructs such as access to care, quality of life, and health status may be accompanied by more complex administration and scoring procedures. In these cases an explicit statement of the rules is necessary to ensure accurate measurement of the attributes.

Lastly, inherent in the definition of measurement is the notion of quantification. Numbers are assigned to an attribute in such a way as to represent quantities of that attribute (Nunnally & Bernstein, 1994). Quantification of an attribute is necessary. By converting abstract concepts to the numerical system through the process of measurement, we create the capacity to perform mathematical and statistical operations. For example, although we can see the heights of two young children, we cannot determine whether one of them is taller than the other unless there is a markedly visible difference in their heights or they are standing side by side. If the children live in different countries and are similar in height, how can we determine which of them is taller? Converting the concept of height into a number allows us to quantify the height of the child in country A and compare his or her height to that of the child in country B.

By converting concepts to numbers, we can perform other mathematical functions as well. We can measure the height, calculate the body mass index, and determine the developmental stage for each child. We can then compare these aspects to the norms for the child's age group. We can also compute averages for groups and norms for age groups. Without the process of measurement, we would be forced to use a conceptual system to discuss the attributes of the children, thus limiting our statements to broad conceptual comparisons. We could say that the children are large or small, are short or tall, or develop slowly or quickly. We could say that one child appears to have more body fat than another, but we would be unable to give more precise comparisons such as the exact difference in body mass and its impact on health.

Figure 1.1 is a visual representation of the conversion of concepts to numbers. The model shows that the conversion occurs through a process of measurement and that error is associated with this process. As you read this book, you will learn about different types of measurement processes, errors of measurement associated with them, and strategies used to identify and reduce error.

FIGURE 1.1. CONCEPTUALIZATION OF THE MEASUREMENT PROCESS.



Reasons for Measuring Concepts

Assigning numbers that represent quantity to concepts allows us to perform operations related to the concepts that would be impossible without the conversions. For example, if we moved into a new apartment and wanted to purchase a table for the dining room, how would we know the size of the table to buy? We could go to the store and look for tables that seemed about the right size. But a better alternative would be to measure the size of the room, to allow space to walk and to move the chairs around the table, and from those figures to estimate the maximum size of an appropriate table for the room. Or we could go to the store, measure a table we liked, and then go home and measure the size of the area to see whether the table would fit. The point is that because we are able to measure the attributes of length and width, we can determine, without actually bringing the table home, whether it will fit the assigned space. Measurement, then, creates *flexibility*. It also allows us to compare similar concepts (for example, the length of the table and the length of the room) and to make judgments about things without actually seeing them (for example, the fact that a table 10 feet long is larger than one that is 3 feet long).

The benefits of measurement also inform the daily work of health practice. A health educator who wants to determine the success of a seminar for people with hypertension can develop a test to evaluate changes in the participants' knowledge of hypertension. Rather than assume that participants have learned certain facts about hypertension, the health educator can determine how many participants have understood the information by administering a test as an objective measure of this knowledge. Using the information gained from the test results, the health educator can determine how many items participants answered correctly, thus quantifying the results. Using statistical tests, the posttest results can be compared to the pretest results to determine the amount of change in knowledge that can be attributed to the seminar. The health educator can evaluate the total scores on the test as well as on individual items to determine whether participants have learned some facts but not others. The information, in turn, is useful for the improvement of the seminar or the *communication* of the results of the program to others. For example, the health educator can share the results with the participants or with the funding agency. Once developed the test is available for every seminar, thus saving the time and expense that would be involved in creating new methods of assessment for each seminar. The health educator can compare the results over time to determine whether the program is meeting its objectives. The test itself can be shared with other health educators, thus increasing the ability to compare the effectiveness of different types of hypertension programs in improving knowledge.

Measurement also provides the means by which we can test propositions from theory. In the present example, the health educator might have based the program on a

theory of behavioral change such as social cognitive theory, and thus might propose that participants who report higher levels of self-efficacy would also report higher rates of behavioral change. The health educator also might propose that participants in the seminar would be more likely to adopt strategies to control hypertension than participants in a control group. Using data collected from the pretests and posttests, the health educator could answer the research questions or test the hypotheses. The results of these tests could determine whether or not the hypotheses were supported, which, in turn, could provide support (or not) for the theory itself.

Scales of Measurement

Another introductory concept that is important to the conceptualization of measurement proposed by Stevens is the categorization of measurement scales, also called levels of measurement. Stevens (1946) proposed that one can classify measurement scales according to the kind of transformations they can undergo without losing their properties. The four scale types he originally proposed are *nominal*, *ordinal*, *interval*, and *ratio*; the four properties of the scales, respectively, are *kind*, *order*, *equal intervals*, and *ratios* of individual values. Each scale can be transformed, but the transformations must maintain the properties of the scale. Stevens refers to this characteristic as *invariance*: that is, the scale must retain its properties following transformation. Nominal scales are the least restrictive and can undergo a variety of transformations while maintaining their properties, whereas ratio scales are the most restrictive and can undergo few transformations without changes in their properties. Because scale properties and their transformations are often confusing for students, we present a detailed description of each scale type, along with examples of measures used in health behavior research.

Nominal Scale

Nominal scales are used to classify variables that we can place in categories based on equivalence. Of the four scale properties listed above, the nominal-level variable is characterized by *kind*. More specifically, the categories of nominal-level variables differ in kind only. In regard to the other three properties of scales, nominal scales have no meaningful order such that assignment to one category indicates a greater or smaller amount of the variable. Likewise, the intervals between numbers on the scale are not equivalent, nor are ratios of individual values meaningful. An example of a nominal variable is the variable *gender* (also referred to as *sex*). It is composed of two categories: *male* and *female*. People are classified as either male or female, and all members of one group (for example, female) are the same on the variable (gender) but different from the members of the other group (male). Marital status is another common variable measured on the

nominal scale. Although there are several possible categories of marital status, for this example we use the following: *married*, *never married*, *separated*, *widowed*, and *divorced*. People placed together in one of these categories have the same marital status, but differ on the marital status variable from people placed in the other categories.

The categories of a nominal scale must be mutually exclusive, meaning that a person, case, or event can be assigned to only one of the categories. The categories must also be exhaustive, so that all instances of an event or all of its characteristics can be categorized. If a researcher were interested in knowing how many of his or her participants were living with partners but not married, the marital status classification just discussed would fail to work. The researcher might elect to add another category, *living with a partner*. Adding this category would lead to overlapping categories, violating the principle that the categories must be mutually exclusive. A participant who has never been married but is living with a partner would be able to check two boxes, as would one who is divorced and living with a partner. In such a case, the researcher must determine exactly what information is important and develop categories that are both mutually exclusive and exhaustive.

Because the categories on a nominal scale differ only in kind and not in degree or amount, we can assign any number to each category as long as we use a different number for each. We could code male as 0 and female as 1 or vice versa. Because the particular numbers, and therefore any mathematical or statistical manipulation of the numbers, have no meaning, we could code male as 1003 and female as 62. By convention, though, the numbers 0 and 1 or 1 and 2 are used for coding the categories of gender. Nominal scales are the least restrictive in regard to the invariant criterion because no matter what numbers we choose, the property of the scale that categories differ in kind remains unchanged. The other properties of scales do not apply to nominal-level variables. That is, nominal-level variables have no meaningful order that would lead us to consider that a male has more gender than a female or vice versa.

The literature contains some controversy about whether classification as employed in the nominal scale is a true form of measurement (Lord & Novick, 1968; Stevens, 1959). Recall that we said that numbers are assigned to represent quantities of attributes. With nominal level scales, assigned numbers do not represent quantity, a seeming violation of the definition of measurement. Stevens (1959) contends that classification is a form of measurement because the assignment of numbers is done according to the rule of not assigning the same number to different categories or different numbers to the same category. Nunnally and Bernstein (1994, p. 1) provide a more refined definition of measurement that addresses this issue directly: "Measurement consists of rules for assigning symbols to objects so as to (1) represent quantities of attributes numerically (scaling) or (2) define whether the objects fall in the same or different categories with respect to a given attribute (classification)." Thus, nominal level variables meet the conditions for the second definition of measurement.

Ordinal Scale

Variables measured on an ordinal scale are those whose categories have a meaningful order that is *hierarchical* in nature. For example, the variable *college standing*, with the categories of *freshman*, *sophomore*, *junior*, and *senior*, is an ordinal variable. The categories represent an increasing number of credit hours completed; *freshman* represents the smallest number of credit hours and *senior* represents the largest. An important point to note, however, is that there is likely to be variability in the attributes within the categories. Thus, the category of *freshman* might include students who are just beginning their college careers and have not completed any coursework as well as those who have completed the maximum of credit hours for freshmen. Likewise, sophomores include those whose numbers of credit hours range from the least to the most for that classification. The grade of heart murmurs is another ordinal variable. Heart murmurs are graded from 1 to 6 depending on their severity. The variable is ordinal because there are higher levels of cardiac dysfunction as the grade of the murmur increases from 1 to 6. The numbers 1 through 6 indicate increasing severity in the heart murmur, but do not reflect the actual amount of increased severity.

Rating scales that use responses such as *strongly agree*, *agree*, *neither agree nor disagree*, *disagree*, and *strongly disagree* are considered ordinal scales. The categories of such a scale are distinct and differ in degree. The categories are ordered so that a report of *agree* suggests a more favorable attitude than that of *disagree*. However, there is likely a range of feelings among people responding in each category. For those who respond *agree*, some may lean more toward *strongly agree* and others may be closer to *neither agree nor disagree*. The numbers assigned to the categories can only reflect the order of the categories. The numbers cannot represent the precise degrees of difference in attitude between the categories.

Variables measured on an ordinal scale will remain unchanged under transformations in which the order of the categories is always maintained. Transformations include adding a constant to all the numbers or multiplying the numbers by a constant. Thus we could code college class status as 1, 2, 3, and 4 to represent the four classes from *freshman* to *senior*. The numbers selected must always reflect the order. Thus, if we are using these numbers, freshmen must always be coded as 1 and seniors must always be coded as 4. But as shown in Exhibit 1.1, if we added 18 to each of the numbers or multiplied each

EXHIBIT 1.1. TRANSFORMATION OF AN ORDINAL VARIABLE BY ADDITION AND BY MULTIPLICATION.

$$(1 + 18), (2 + 18), (3 + 18), (4 + 18) = 19, 20, 21, 22$$

$$(1 \times 44), (2 \times 44), (3 \times 44), (4 \times 44) = 44, 88, 132, 176$$

number by 44, the actual numbers would change, but the rank order of the numbers would remain the same. Thus, any transformation that maintains this original order is acceptable.

Interval Scale

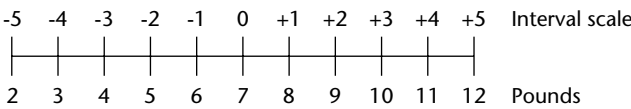
Variables measured on the interval scale have a *meaningful order*: higher levels have more of the attribute. Unlike those of the nominal and ordinal scales, the numbers used to represent the attribute measured on an interval scale are themselves meaningful and provide information about the amount of the attribute. In addition, interval scales have equal intervals, so that the distances between the successive units of measure are equivalent. The interval scale does not have a true zero point, a point at which there is nothing left of the attribute. Without the true zero point, the interpretation of ratios of individual values on the scale is not meaningful. However, the interpretation of ratios of differences is still meaningful. This concept is usually confusing for students, so we will give two examples to explain it.

The most common example of an interval scale is temperature measured in degrees Fahrenheit. A temperature of 100 degrees Fahrenheit is greater than one of 60 degrees Fahrenheit. Moreover, all the degrees of heat (molecular activity) are equivalent. Thus the difference in the amount of heat between 100 degrees and 60 degrees (40 degrees) is equivalent to the difference between 30 degrees and 70 degrees (40 degrees). We can also say that the difference in the amount of heat between 100 degrees and 60 degrees (40 degrees) is twice that of the difference between 50 degrees and 30 degrees (20 degrees) (ratios of differences). We cannot say, however, that 60 degrees is twice as hot as 30 degrees (ratio of individual values).

For a second example, suppose a researcher knows that the average birth weight for a particular group of infants is 7 pounds. The researcher wants to determine how much each infant deviates from the average weight. If the researcher recalibrated the scale with the mean weight set at 0, then 7 pounds would be coded as 0 (Figure 1.2). A baby weighing 8 pounds would receive a score of +1; a baby weighing 9 pounds, a score of +2; and so on. A baby weighing 6 pounds would receive a score of -1; one weighing 5 pounds, a score of -2. Using the new interval scale, as shown in Figure 1.2, we could say that the difference in weight between -2 and 0 (2 pounds) is equivalent to the difference in weight between +2 and +4 (2 pounds). However, we could not say that a baby with a score of +4 (11 pounds) weighs twice as much as one with a score of +2 (9 pounds).

When a scale has an arbitrary zero point, we can transform the scale by adding or multiplying by a constant. Thus, subtracting 32 from a Fahrenheit temperature and then multiplying the result by $\frac{5}{9}$ will change the scale from Fahrenheit degrees to Celsius degrees. Exhibit 1.2 shows the conversion of the Fahrenheit temperatures mentioned previously to Celsius temperatures. Note that when the Fahrenheit temperatures

FIGURE 1.2. INTERVAL SCALE FOR WEIGHTS OF NEWBORNS USING THE MEAN WEIGHT AS THE ORIGIN.



are converted to Celsius, the differences between the high and low temperatures in the two examples are equivalent (for example, 22 degrees). That is, the difference between 37.4°C and 15.4°C is equal to the difference between 20.9°C and -1.1°C. However, on the Fahrenheit scale, 60 degrees is twice 30 degrees, whereas the same temperature measured in Celsius—15.4°C—is not twice as much as -1.1°C. Thus, the interpretation of the ratios of individual values is not invariant when temperatures are converted from Fahrenheit to Celsius degrees.

An interval scale is considered to have an underlying continuous measure, whereas categories of variables measured on the ordinal scale are separated into distinct groups (for example, *freshman*, *sophomore*, *junior*, and *senior*). Although it is possible to convert interval-level variables into categories, the interpretation of the categories must take into

EXHIBIT 1.2. CONVERSION OF TEMPERATURE FROM FAHRENHEIT TO CELSIUS DEGREES.

Proof showing ratios of differences are equivalent:

$$\frac{5}{9} (^{\circ}\text{F} - 32) = ^{\circ}\text{C}$$

$\frac{5}{9}(100^{\circ}\text{F} - 32) = 37.7^{\circ}\text{C}$	$\frac{5}{9}(70^{\circ}\text{F} - 32) = 21.1^{\circ}\text{C}$
$\frac{5}{9}(60^{\circ}\text{F} - 32) = 15.5^{\circ}\text{C}$	$\frac{5}{9}(30^{\circ}\text{F} - 32) = -1.1^{\circ}\text{C}$
Differences: 40°F 22.2°C	Differences: 40°F 22.2°C

100°F – 60°F is equivalent to 70°F – 30°F (both are 40°F).
37.7°C – 15.5°C is equivalent to 21.1°C – (-1.1)°C (both are 22.2°C).

Proof showing ratios of individual values are not equivalent:

The ratio $\frac{60}{100}$ is not equivalent to $\frac{15.5}{37.7}$. The ratio $\frac{30}{70}$ is not equivalent to $\frac{-1.1}{21.1}$.

account the context of the conversion. Any attempt to categorize temperatures as low, medium, or high (ordinal scale) would depend on the situation. Categories of low, medium, and high for the temperature of the human body would differ from low, medium, and high air temperatures. Another consideration is that although we might report temperature in whole numbers, we can also measure temperature to a fraction of a degree. In some situations, it might make sense to measure temperature to the tenth or one-hundredth of a degree. The continuous scale allows these fine distinctions, whereas an ordinal scale does not.

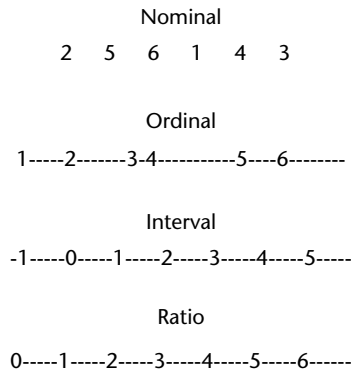
As we have implied above, because an interval scale has an arbitrary zero point, negative values represent actual amounts of the attribute and thus have meaning. We know that Fahrenheit temperature is a measure of heat. Because 0 degrees is an arbitrary number, at 0 degrees Fahrenheit there is still molecular activity (heat) that can be measured. Indeed, air temperatures as low as -50 degrees Fahrenheit have been recorded. Likewise, when the researcher codes infant weights using the average weight as the zero point, a score of -3 is associated with an actual value of weight and has meaning.

Ratio Scale

The ratio scale has all the characteristics of the interval scale except for the arbitrary zero point. A ratio scale has a true zero point, a point at which nothing is left of the attribute. The true zero point permits the *meaningful expression of ratios*. Thus, as with an interval scale, we can say that the difference between 2 and 4 is the same as the difference between 4 and 6 (ratio of differences). However—and this is not the case with an interval scale—we can also say that 8 is twice as much as 4 (ratio of individual values). Both height and weight are measured on a ratio scale. Nothing is left to measure when height or weight is 0. A person with a blood pressure of 0 will not live long without medical intervention. Likewise, molecules do not move and generate heat where the Fahrenheit temperature is -460 (0 Kelvin, or absolute temperature). Because the measure of weight has a true zero point, a baby weighing 10 pounds is twice as heavy as one weighing 5 pounds.

Ratio scales have more restrictions on the types of transformations that can be applied, and they remain invariant only when values are multiplied by a constant. Thus, multiplying weight in pounds by 16 (the number of ounces in a pound) will convert 10 pounds to 160 ounces. When measured in ounces rather than pounds, the difference between the weights of a baby weighing 80 ounces and one weighing 160 ounces is the same as the difference between the weights of a baby weighing 112 ounces and one weighing 192 ounces. Likewise, after the conversion of weight from pounds to ounces, a baby who weighs 160 ounces remains twice as heavy as a baby who weighs 80 ounces.

A useful schema of the differences among the four measurement scales is shown in Figure 1.3.

FIGURE 1.3. SCHEMA FOR LEVELS OF MEASUREMENT.

Levels of Measurement and Statistics

Stevens (1959) extended his classification of variables to include allowable statistical operations for each level of measurement. In general, an inverse relationship exists between allowable transformations and allowable statistics. The variables with the fewest restrictions on transformation (nominal scale) have the most restrictions in terms of statistical operations. Thus, if a data set included only variables measured on the nominal scale, the number of possible statistical tests would be limited. Frequencies and percentages could describe single variables, and we could use the mode as the measure of central tendency. However, significance tests would be limited to the chi-square or a similar statistic. For example, we could calculate the number and percentage of men or women in the study, thereby determining whether we had more men or women in the study (mode). We could make comparisons among variables using cross-tabulations evaluated with the chi-square statistic. Then we could determine whether a higher percentage of women as compared to men were married.

As with variables measured on a nominal scale, statistical manipulation is limited with ordinal-level variables. In addition to the mode, a median could be calculated for an ordinal scale, along with percentiles and a rank-order correlation. We could use nonparametric statistics as well. With interval-level variables, the researcher can calculate an arithmetic mean, a standard deviation, and a product moment correlation. Parametric tests, such as analysis of variance (ANOVA), are also possible. At the ratio level, all statistical tests are possible, including calculation of the geometric and harmonic mean, percentage of variation, and higher-level statistics such as regression and structured equation modeling. The student should remember that certain statistical tests, such as the t test and ANOVA, require a combination of categorical and continuous variables. For these tests, the independent variable must be categorical and the dependent variable continuous. Independent variables that are continuous

can be converted into categorical variables by grouping. For example, if weight is an independent variable, it can be converted into a categorical variable by grouping participants into weight groups—those who weigh less than 100 pounds in one group and those who weigh 100 pounds or more in another.

When Stevens (1951) proposed the categorization schema and associated statistical tests, he had no idea how significant and controversial this classification would become. Yet, writing in 1959, Stevens was aware that future researchers might discover additional levels of measurement. He even proposed a logarithmic interval scale and suggested the existence of scales between the ordinal and interval levels. Some researchers, however, have applied rigid standards to the use of the schema. These investigators insist that statistics applied in any study must correspond to the appropriate level of measurement of the variables (representational view), whereas other investigators apply more flexible standards (operational or classical view; see Nunnally & Bernstein, 1994). The greatest area of controversy for behavioral researchers is the interpretation of the summated rating scales. Those who espouse the representational viewpoint insist that rating scales be evaluated using the median and nonparametric statistics. However, those in the operational and classical camps note that under certain circumstances, interval-level scales may be created from ordinal scales. One example, provided by Nunnally and Bernstein (1994), is test scores. Individual items (for example, multiple-choice or true/false) on a test are considered ordinal-level variables. However, summing the items produces an interval-level scale to which one can apply an arithmetic mean and the associated statistical tests. These statistical tests include mean comparisons (t and F tests), multiple regression, and other parametric tests. The same type of transformation is possible with summated rating scales. Individual items with response options such as *strongly disagree* to *strongly agree* are considered ordinal scales. However, summing the responses provides an interval-level variable. Nunnally and Bernstein (1994), along with others (Pedhazur & Schmelkin, 1991), note that little damage is done by the application of the higher-level statistics to the analysis of data using summated rating scales. The proof, they say, is in the results. If the literature is a reflection of health behavior researchers' beliefs in this area, most researchers are in the classical camp. A review of studies shows that it is a common practice for researchers to apply parametric statistics to the study of variables that are measured with ordinal scales, including summated rating scales.

Major Concepts of Measurement: Reliability and Validity

Reliability

The two fundamental concepts of measurement are reliability and validity. *Reliability* refers to consistency: that is, does the instrument produce scores that are internally consistent or stable across time? We do not consider a scale reliable if it indicates

that a person weighs 145 pounds at one moment and 162 pounds ten minutes later (given, of course, no obvious change such as the addition of winter clothing for a trek to Kathmandu). Likewise, a thermometer is expected to provide consistent readings, as are many other instruments we use in our daily lives.

In this book we will address reliability as related to instruments designed to measure health behaviors and factors associated with health behaviors. As we will learn, several different ways to assess reliability are available. The selection of reliability procedures depends on a number of factors, including the attribute being measured, the type of instrument, the investigator's skill and available time, the availability of research participants, and data collection time and efforts. We will learn about three major procedures to assess reliability: equivalence, stability, and internal consistency.

Validity

Validity refers to the legitimacy of the scores as a measure of the intended attribute. That is, does the instrument measure what it is intended to measure? We know a thermometer measures heat, but we would not expect degrees on a thermometer to give us an indication of a person's level of self-esteem. However, measures that collect self-reports of behaviors, attitudes, and psychosocial states are not as easy to assess. Researchers encounter more difficulty in determining whether or not these measures are valid. A variety of factors—including conceptual factors, respondent factors, item factors, and methods of administration—might lead to measurement errors that would influence validity. The latest set of standards for education and psychological testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999) lists five sources of validity evidence. These are test content, response processes, relations to other variables, internal structure, and consequences of testing. We will discuss these sources in later chapters.

Summary

Health educators and behavioral scientists holding a variety of health and research positions often need to understand measurement and to use measurement techniques. As we have learned in this chapter, measurement has a long history, dating back to ancient times. However, the measurement theories and techniques we draw on today were developed and refined primarily in the twentieth century. The work of S. S. Stevens has had a powerful impact on the conceptualization of measurement among psychologists and behavioral scientists. In particular, his definitions of measurement and categorization of variables are used extensively in instrument development and testing. In later chapters we will become familiar with the work of other giants in the field and their contributions to reliability and validity assessment.