PART I

BASIC CONCEPTS IN MEASUREMENT

SCALING

If something exists, it must exist in some amount (Thorndike, 1918). Psychologists generally believe that people have psychological attributes, such as thoughts, feelings, emotions, personality characteristics, intelligence, learning styles, and so on. If we believe this, then we must assume that each psychological attribute exists in some quantity. With this in mind, psychological measurement can be seen as a process through which numbers are assigned to represent the quantities of psychological attributes. The measurement process succeeds if the numbers assigned to an attribute reflect the actual amounts of that attribute.

The standard definition of measurement (borrowed from Stevens, 1946) found in most introductory test and measurement texts goes something like this: "Measurement is the assignment of numerals to objects or events according to rules." In the case of psychology, education, and other behavioral sciences, the "events" of interest are generally samples of individuals' behaviors. The "rules" mentioned in this definition usually refer to the scales of measurement proposed by Stevens (1946).

This chapter is about **scaling**, which concerns the way numerical values are assigned to psychological attributes. Scaling is a fundamental issue in measurement, and it involves a variety of considerations. This chapter discusses the meaning of numerals, the way in which numerals can be used to represent psychological attributes, and the problems associated with trying to connect psychological attributes with numerals. As discussed in the previous chapter, psychological tests are intended to measure unobservable psychological characteristics such as attitudes, personality traits, and intelligence. Such characteristics present special problems for measurement, and this chapter discusses several possible solutions for these problems.

These issues might not elicit cheers of excitement and enthusiasm among some readers or perhaps among most readers (or perhaps in any reader?); however, these issues are fundamental to psychological measurement, to measurement in general, and to the pursuit and application of science. More specifically, they are important because

they help define scales of measurement. That is, they help differentiate the ways in which psychologists apply numerical values in psychological measurement. In turn, these differences have important implications for the use and interpretation of scores from psychological tests. The way scientists and practitioners use and make sense out of tests depends heavily on the scales of measurement being used. Your attention to the material in this chapter should be rewarded with new insights into the foundations of psychological measurement and even into the nature of numbers.

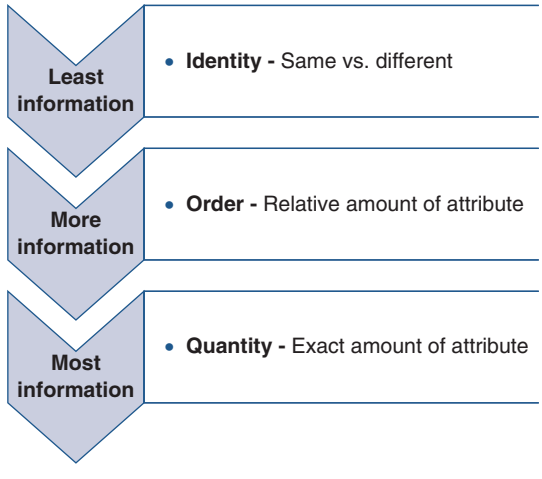
FUNDAMENTAL ISSUES WITH NUMBERS

In psychological measurement, numerals are used to represent an individual's level of a psychological attribute. For example, your numerical score on an IQ test is used to represent your level of intelligence, your numerical score on the Rosenberg Self-Esteem Inventory is used to represent your level of self-esteem, and a numerical value can even be used to represent your biological sex (e.g., males might be referred to as "Group 0" and females as "Group 1"). Thus, psychological measurement is heavily oriented toward numbers and quantification.

Importantly, numerals can represent psychological attributes in different ways, depending on the nature of the numeral that is used to represent an attribute. This section describes important properties of numerals, and it shows how these properties influence the ways in which numerals represent psychological attributes.

As shown in Figure 2.1. this section outlines three important numerical properties, and it discusses the meaning of zero. In essence, the numerical properties of

FIGURE 2.1 ⬢ Properties of Numbers



identity, order, and quantity reflect the ways in which numerals represent potential differences in psychological attributes. Furthermore, zero is an interestingly complex number, and this complexity has implications for the meaning of different kinds of test scores. A "score" of zero can have extremely different meanings in different measurement contexts.

The Property of Identity

The most fundamental form of measurement is the ability to reflect "sameness versus differentness." Indeed, the simplest psychological measurements are those that differentiate between categories or groups of people.

For example, you might ask first-grade teachers to identify those children in their classrooms who have behavior problems. The children who are classified as having behavior problems should be *similar to* each other with respect to their behavior. In addition, those children should be *different from* the children who are classified as not having behavioral problems. That is, the individuals within a category should be the same as each other in terms of sharing a psychological feature, but they should be different from the individuals in another category. In psychology, this requires that we sort people into at least two categories. The idea is that objects, events, or people can be sorted into categories that are based on similarity of features. In many cases, these features are behavioral characteristics reflecting psychological attributes, such as happy or sad, introverted or extroverted, and so on.

Certain rules must be followed when sorting people into categories. The first and most straightforward rule is that, to establish a category, the people within a category must satisfy the property of **identity**. That is, all people within a particular category must be "identical" with respect to the feature reflected by the category. For example, everyone in the "behavioral problem" group must, in fact, have behavioral problems, and everyone in the "no behavioral problem" group must not have behavioral problems. Second, the categories must be *mutually exclusive.* If a person is classified as having a behavioral problem, then they cannot simultaneously be classified as not having a behavioral problem. Third, the categories must be *exhaustive.* If you think that all first-graders can be classified as either having behavioral problems or not having behavioral problems, then these categories would be exhaustive. If, on the other hand, you can imagine someone who cannot be so easily classified, then you would need another category to capture that person's behavior. To summarize the second and third rules, each person should fall into one and only one category.

When numerals have only the property of identity, they represent sameness vs. differentness, and they serve simply as labels of categories. The categories could be labeled with letters, names, or numerals. You could label the category of children with behavior problems as "Behavior Problem Children," you could refer to the category as "Category B," or you could assign a numeral to the category. For example, you could label the group as "0," "1," or "100."

When having only the property of identity, numerals are generally not thought of as having true mathematical value. For example, if "1" is used to reflect the category of children with behavioral problems and "2" is used to represent the category of children without behavioral problems, then we would not interpret the apparent 1-point difference between the numerical labels as having any form of quantitative significance.

The latter point deserves some additional comment. When making categorical differentiations between people, the distinctions between categories represent differences in kind or quality rather than differences in amount. Again returning to the teachers' classifications of children, the difference between the two groups is a difference between *types* of children—those children who have behavioral problems and those who do not. In this example, the classification is not intended to represent the amount of problems (e.g., a lot vs. a little) but rather the presence or absence of problems. In this way, the classification is intended to represent two qualitatively distinct groups of children.

Of course, you might object that this is a rather crude and imprecise way of measuring or representing behavioral problems. You might suggest that such an attribute is more accurately reflected in some degree, level, or amount than in a simple presence/absence categorization. This leads to additional properties of numerals.

The Property of Order

Although identity is the most fundamental property of a numeral, the property of order conveys more information. As discussed above, when numerals have only the property of identity, they convey information about whether two individuals are similar or different but nothing more. In contrast, when numerals have the property of **order**, they convey information about the relative amount of an attribute that people possess.

When numerals have the property of order, they indicate the rank order of people relative to each other along some dimension. In this case, the numeral 1 might be assigned to a person because they possess more of an attribute than anyone else in the group. The numeral 2 might be assigned to the person with the next greatest amount of the attribute, and so on.

For example, teachers might be asked to rank children in their classrooms according to the children's interest in learning. Teachers might be instructed to assign the numeral 1 to the child who shows the most interest in learning and 2 to the child whose interest in learning is greater than all the other children except the first child, continuing in this way until all the children have been ranked according to their interest in learning.

When numerals are used to indicate order, they again serve essentially as labels. For example, the numeral 1 indicated a person who had more of an attribute than anyone

else in the group. The child with the greatest interest in learning was assigned the numeral 1 as a label indicating the child's rank. In fact, we could just as easily assign letters as numerals to indicate the children's ranks. The child with the most (least) interest in learning might have been assigned the letter A to indicate his or her rank. Each person in a group of people receives a numeral (or letter) indicating that person's relative standing within the group with respect to some attribute. For communication purposes, it is essential that the meaning of the symbol used to indicate rank be clearly defined. We simply need to know what 1, or A, means in each context.

Although the property of order conveys more information than the property of identity, it is still quite limited. While it tells us the relative amount of differences between people, it does not tell us about the actual degree of differences in that attribute. For example, based on ordinal information, we might know that the child ranked 1 has more interest in learning than the child ranked 2, but we do not know *how much* more interest they have. The two children could differ only slightly in their amount of interest in learning, or they could differ dramatically. In this way, when numerals have the property of order, they are still a rather imprecise way of representing psychological differences.

The Property of Quantity

Although the property of order conveys more information than the property of identity, the property of quantity conveys even greater information. As noted above, numerals that have the property of order convey information about which of two individuals has a higher level of a psychological attribute, but they convey no information about the exact amounts of that attribute. In contrast, when numerals have the property of **quantity**, they provide information about the magnitude of differences between people.

At this level, numerals reflect *real numbers* or, for our purposes, numbers. The number 1 is used to define the size of the basic *unit* on any particular scale. All other values on the scale are multiples of 1 or fractions of 1. Each numeral (e.g., the numeral 4) represents a count of basic units.

Think about a thermometer that you might use to measure temperature. To describe how warm the weather is, your thermometer reflects temperature in terms of "number of degrees" (above or below 0). The degree is the unit of measurement, and temperature is represented in terms of this unit.

Units of measurement are standardized quantities; the size of a unit will be determined by some convention. For example, 1 degree Celsius (1°C) is defined (originally) in terms of 1/100th of the difference between the temperature at which ice melts and the temperature at which water boils. We will revisit this important point shortly.

Real numbers are also said to be continuous. In principle, any real number can be divided into infinitely small parts. In the context of measurement, real numbers

are often referred to as *scalar*, *metric*, or *cardinal*, or sometimes simply as *quantitative* values.

The power of real numbers derives from the fact that they can be used to measure the amount or quantity of an attribute of a thing, person, or event. When applied to an attribute in an appropriate way, a real number indicates the amount of something. For example, a day that has a temperature of 50°C is not simply warmer than a day that has a temperature of 40°C; it is precisely 10 units (i.e., degrees) warmer.

When psychologists use psychological tests to measure psychological attributes, they often assume that the test scores have the property of quantity. As we will see later, this often might not be a reasonable assumption.

The Number 0

The number 0 is a strange number (see Seife, 2000), with at least two potential meanings. To properly interpret a score of 0 in any particular situation, you must understand which meaning is relevant in that situation.

In one possible meaning, zero reflects a state in which an attribute of an object or event has no existence. If you said that an object was 0.0 cm long, you would be claiming that the object has no length, at least in any ordinary sense of the term *length*. Zero in this context is referred to as **absolute zero**. In psychology, the best example of a behavioral measure with an absolute 0 point might be reaction time.

The second possible meaning of zero is to view it as an arbitrary quantity of an attribute. A zero of this type is called a relative or **arbitrary zero**. In the physical world, attributes such as time (e.g., calendar, clock) and temperature measured by standard thermometers are examples. In these examples, 0 is simply an arbitrary point on a scale used to measure that feature. For example, a temperature of 0 on the Celsius scale represents the melting point of ice, but it does not represent the "absence" of anything (i.e., it does not represent the absence of temperature or of warmth).

The psychological world is filled, at least potentially, with attributes having a relative 0 point. For example, it is difficult to think that conscious people could truly have no (zero) intelligence, self-esteem, introversion, social skills, attitudes, and so on. Although we might informally say that someone "has no social skill," psychologists would not suggest this formally—indeed, we actually believe that everyone has some level of social skill (and self-esteem, etc.), although some people might have much lower levels than other people.

Despite the fact that most psychological attributes do not have an absolute 0 point, psychological tests of such attributes could produce a score of 0. In such cases, the zero would be considered arbitrary, not truly reflecting an absence of the attribute. Furthermore, you will see that many if not most psychological test scores can be expressed as a type of score called a z score, which will be discussed in Chapter 3.

A *z* score of 0 indicates an average score within the set of score. In this case, zero represents an arbitrary or relative zero.

In psychology, there can be a problem in determining whether a test score of zero should be thought of as relative or absolute. The problem concerns the distinction between the test being used to measure a psychological attribute and that psychological attribute itself.

Consider an example that Thorndike (2005) used to illustrate this problem. Thorndike describes a scenario in which a sixth-grade child takes a spelling test and fails to spell any of the words correctly. The child thus receives a score of 0 on the test. In this case, the spelling test is the instrument used to measure an attribute of the child—the child's spelling ability. The test itself has an absolute 0 point, indicating that the child failed to spell any words correctly. That is, the test score of 0 indicates an absence of correctly spelled words. It is difficult, however, to imagine that a sixth-grade child is incapable of spelling; the child's *spelling ability* is probably not zero. The question then becomes how we are going to treat the child's test score. Should we consider it an absolute zero or a relative zero?

This is important because the type of zero associated with a test affects how we interpret and use the test scores. For example, we might plan to conduct statistical analyses on test scores for a research study. Importantly, the types of analyses that we can legitimately conduct are determined, in part, by the type of zero that is reflected in the test scores. On one hand, if we can assume that a test has an absolute zero, then we can feel comfortable performing the arithmetic operations of multiplication and division on the test scores. On the other hand, if a test has a relative 0 point, then we should restrict arithmetic operations on the scores to addition and subtraction. As a matter of evaluation, it is important to know what zero means—does it mean that a person who scored 0 on a test had none of the attribute that was being measured, or does it mean that the person might not have had a measurable amount of the attribute, at least not measurable with respect to the particular test you used to measure the attribute?

In sum, the three properties of numerals and the meaning of zero are fundamental issues that shape our understanding of psychological test scores. If two people share a psychological feature, then we have established the property of identity. If two people share a common attribute but one person has more of that attribute than the other, then we can establish order. If order can be established and if we can determine *how much* more of the attribute one person has compared with others, then we have established the property of quantity. Put another way, identity is the most fundamental level of measurement. To measure anything, the identity of the thing must be established. Once the identity of an attribute is known, it might be possible to establish order. Furthermore, order is a fundamental characteristic of quantity. As we will see, numbers play a different role in representing psychological attributes depending on their level of measurement.

Most psychological tests are treated as if they provide numerical scores that possess the property of quantity. The next two sections discuss key issues regarding the meaning and use of such quantitative test scores. Specifically, they discuss the meaning of a "unit of measurement," the issues involved with counting those units, and the implications of those counts.

UNITS OF MEASUREMENT

The property of quantity requires that units of measurement be clearly defined. As discussed in the next section, quantitative measurement depends on our ability to count these units. Before discussing the process and implications of counting the units of measurement, we must clarify what is meant by a unit of measurement.

In many everyday cases of physical measurement, the units of measurement are familiar. When measuring the length of a piece of lumber, the width of a couch, or the height of their children, people typically use a tape or ruler marked off in units of inches or centimeters. Length, width, and height are measured by counting the number of these units from one end of the lumber, couch, or child to the other end.

In contrast, in many cases of psychological measurement, units of measurement are often less obvious. When measuring a psychological characteristic such as shyness, working memory, attention, or intelligence, what are the units of measurement? Presumably, they are responses of some kind, perhaps to a series of questions or items. But how do we know whether, or to what extent, those responses are related to the psychological attributes themselves? This book returns to these questions at a later time, as they represent the most vexing problems in psychometrics. At this point, let's focus on the notion of a unit of measurement. This can be illustrated in the context of the measurement of the length of physical objects (Michell, 1990).

Imagine that you are building a bookshelf and you need to measure the length of pieces of wood. Unfortunately, you cannot find a tape measure, a yardstick, or a ruler of any kind—how can you precisely quantify the lengths of your various pieces of wood?

One solution is to create your own unique measurement system. First, imagine that you happen to find a long wooden curtain rod left over from a previous project. You cut a small piece of the curtain rod; let us call this piece an "xrod" (see Figure 2.2). Because your pieces of bookshelf wood are longer than the xrod, you will need several xrods. Therefore, you can use this original xrod as a template to produce a collection of identical xrods. That is, you can cut additional xrods from the curtain rod, making sure that each xrod is the same exact length as your original xrod. You can now use your xrods to measure the length of all your pieces of wood. For example, to measure the length of one of your shelves, place one of the xrods at one end of the piece of wood that you will use as a shelf. Next, as shown in

FIGURE 2.2 ⬢ Measuring a Shelf by Using "Xrods"

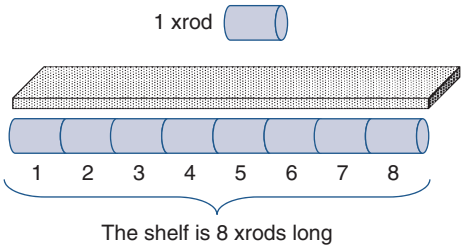


Figure 2.2, place xrods end to end in a straight line until you reach the opposite end of the piece of wood. Now count the number of xrods, and you might find that the shelf is "8 xrods long."

You have just measured length in "units of xrods." You can now use your set of xrods to measure the length of each and every piece of wood that you need. In fact, you could use xrods to measure the length of many things, not just pieces of wood. In many ways, your measure is as good as any measure of length (except that you are the only one who knows what an xrod represents!).

Arbitrariness is an important concept in understanding units of measurement, and it distinguishes between different kinds of measurement units. There are three ways in which a measurement unit might be arbitrary. First, the unit size can be arbitrary. Consider your xrod—the size of your original xrod could have been any length. When you cut that first xrod, your decision about its length could be completely arbitrary—there was no "true" xrod length that you were trying to obtain. You simply chose a length to cut, and that length became the "official" length of an xrod. In this sense, the size of your unit of measurement, the xrod, was completely arbitrary. Similarly, the amount of weight that is represented by a "pound" is an arbitrary amount. Although there is now clear consensus regarding the exact amount of weight represented by a pound, we can ask why a pound should reflect that *specific* amount. The original choice was likely quite arbitrary.

A second form of arbitrariness is that some units of measurement are not tied to any one type of object. That is, there might be no inherent restriction on the objects to which a unit of measurement might be applied. Our xrods can be used to measure the spatial extent of anything that has spatial extent. For example, they could be used to measure the length of a piece of wood, the length of a table, the distance between two objects, or the depth of water in a swimming pool. Similarly, a pound can be used to measure the weight of many different kinds of objects.

A third form of arbitrariness is that, when they take a physical form, some units of measurement can be used to measure different features of objects. For example, the xrods that we used to measure the length of a piece of lumber could also be used as

units of weight. Imagine that you needed to measure the weight of a bag of fruit. If you had a balance scale, you could put the bag in one of the balance's baskets, and you could gradually stack xrods in the other basket. When the two sides of the scale "balance," you would know that the bag of fruit weighs, say, 4 xrods.

Some units of measurement, those called *standard measures*, are based on arbitrary units of measurement in all three ways when they take a physical form. In physical measurement, standard units include units such as pounds, liters, and milliseconds. The fact that they are expressed in arbitrary units gives them flexibility and generality. For example, you can use milliseconds to measure anything from a person's reaction time to the presentation of a stimulus to the amount of time it takes a car to travel down the street.

In contrast to many physical measures, most psychological units of measurement (e.g., scores on aptitude tests or intelligence tests) are generally arbitrary only in the first sense of the term *arbitrary* mentioned above. That is, most psychological units of measurement are arbitrary in size, but they are typically tied to specific objects or dimensions. For example, a "unit" of measurement on an IQ test is linked in a nonarbitrary way to intelligence, and it is not applicable to any other dimension. Because of this feature of IQ test scores, we refer to IQ score units as "IQ points"; the points have no referent beyond the test used to measure intelligence. An exception to this observation is when standard measures are used to measure psychological attributes. For example, reaction times are often used to measure various cognitive processes.

ADDITIVITY AND COUNTING

The need for counting is central to all measurement. Whether we are measuring a feature of the physical world or of the psychological world, all measurement involves counting.

For example, when you used xrods to measure the length of a piece of wood, you placed the xrods end to end, starting from one end of the piece of wood and continuing until you reached the other end. You then counted the xrods to determine the length of the object. The resulting count was a measure of length.

Similarly, when you use a behavioral sampling procedure (i.e., a test) to measure a person's self-esteem, you count responses of some kind. For example, you might count the number of test statements that a test respondent marks as "true," and you might interpret the number of "true" marks as indicating the level of the respondent's self-esteem. That is, you count units to obtain a score for your measurement.

Additivity

Importantly, the process of counting as a **facet of measurement** involves a key assumption that might not be valid in many applications of psychological

measurement. The assumption is that the unit size does not change—that all units being counted are identical. In other words, additivity requires unit size to remain constant; a *unit* increase at one point in the measurement process must be the same as a unit increase at any other point.

Recall again the xrod example, where you used the original xrod as a guide to cut additional xrods—you needed to be "sure that each xrod is the same exact length as your original xrod." By doing so, you ensured that anytime you laid xrods side by side and counted them, you could trust that your count accurately reflected a length. Say that you had cut 10 xrods; if they are all identical, then it does not matter which xrods you used when measuring the length of any piece of wood. That is, a piece of wood that you measured as 5 xrods would be measured as 5 xrods no matter which particular 5 xrods you used to measure the piece of wood.

Now imagine that instead of having a collection of equal-length xrods, your xrods had various lengths. In that case, if you measured the same piece of wood on two occasions, you might get two different counts, indicating different lengths! That is, if some xrods were longer than the others, then your piece of wood might be 5 xrods when you use the shorter xrods, but it would be only 3 xrods if you happened to use the longer xrods. Because your xrods differ in size, there is no single amount of length that is represented by an xrod. Thus, your unit sizes are not constant, and your entire measurement system is flawed. This would prevent you from accurately measuring the length of the lumber.

In addition, the size of a measurement unit should not change as the conditions of measurement change. For example, the size of an xrod should remain constant regardless of the time of day that the xrod is used to measure a piece of wood. In effect, you want your measure to be affected by only one attribute of the thing you are measuring, regardless of the conditions that exist at the time or place of measurement. This condition is referred to as conjoint measurement (Luce & Tukey, 1964) and is a complex issue beyond the scope of this book (but see Green & Rao, 1971, for a clear, nontechnical discussion).

Although these issues might be initially clearest in terms of physical measurements (e.g., xrods), our concern is psychological measurement. So now imagine that you are a history teacher who wants to measure a psychological attribute such as "knowledge of American history." This is often done by asking students questions that you believed were diagnostic of their knowledge and recording their responses to the questions. Let's temporarily differentiate between measurement units and psychological units. That is, each test item represents a measurement unit, and again you count the correctly answered items to obtain a score that you interpret as a student's knowledge of American history. In contrast, we will use the crude and informal idea of psychological units to mean "true" levels of knowledge. Ideally, the measurement units will correspond closely with psychological units. That is, we use test scores to represent levels of psychological attributes. With this in mind, you combine each student's test responses in some way (e.g., by counting the

number of questions that each student answered correctly) to create a total score that is interpreted as a measure of true knowledge of American history.

Suppose that one of the questions on your test was "Who was the first president of the United States?" and another was "Who was the first European to sail into Puget Sound?" It should be clear that the amount of knowledge of American history you need to answer the first question correctly is considerably less than the amount you need to answer the second question correctly. In terms of psychological units, let's say that you needed only 1 psychological unit of American history knowledge to answer the first question correctly but you needed three times as much knowledge (i.e., 3 psychological units of knowledge) to answer the second question correctly.

Consider a student who answered both questions correctly. In terms of amount of true knowledge, that student would have 4 psychological units of history knowledge. However, in terms of measurement, that student would have a score of only 2. That is, if you simply summed the number of correct responses to the questions to get a total score, the student would get a score of 2. This would suggest that the person had 2 units of American history knowledge when in fact they had 4 units of knowledge.

This discrepancy occurs because the measurement units are not constant in terms of the underlying attribute that they reflect. That is, the answers to the questions do not reflect equal-sized units of knowledge—it takes less knowledge to answer the first question than it does to answer the second. Thus, the additive count of correct answers is not a good measure of the amount of actual knowledge.

From a psychological perspective, the assumption is often made that a psychological attribute such as knowledge of American history actually exists in some amount. However, unlike a piece of wood, whose "length" can be directly observed, we cannot directly observe "knowledge of American history." As a result, we cannot simply see if a count of American history questions corresponds to the actual amount of American history knowledge possessed by a particular individual.

There is a paradox in this: We want to translate the amount of a psychological attribute onto a set of numbers to measure the attribute. But it appears that this cannot be done because we do not know how much of the attribute actually exists. This tension is, in part, at the heart of the test validation process (see Chapters 8 and 9). That process largely hinges on a back-and-forth between the theory of the attribute and the empirical data collected about that attribute.

Counts: When Do They Qualify as Measurement?

Although all measurement relies on counting, not all forms of counting are forms of measurement. Indeed, a controversy about the relationship between counting and measurement arises when we count *things* rather than *attributes* (Lord & Novick, 1968; Michell, 1986; Wright, 1997). For example, if you count the

number of forks on a table, are you "measuring" something? Similarly, if you count the number of children in a classroom, are you measuring something?

Some experts argue that simply counting the number of some kind of object does not qualify as a "measurement." They would argue that counting qualifies as measurement only when one is counting to reflect the amount of some feature or attribute of an object. For example, if a physical scientist uses a Geiger counter to count radioactive emissions from an object, then she is measuring the radioactivity of the object, where "radioactivity" is a feature of the object. Similarly, if a professor counts the number of correct answers given by a student on a multiple-choice mathematics test, then she might be measuring the amount of mathematical knowledge of the student, where "amount of mathematical knowledge" is the psychological attribute of the student.

FOUR SCALES OF MEASUREMENT

As discussed earlier, measurement involves the assignment of numbers to observations in such a way that the numbers reflect the real differences that exist between the levels of a psychological attribute. Scaling is the particular way in which numbers are linked to behavioral observations to create a measure (Allen & Yen, 1979; Crocker & Algina, 1986; Guilford, 1954; Magnusson, 1967).

In actuality, the definition of scaling is itself controversial. On one hand, some experts might find this book's definition of scaling unacceptably liberal, and they might restrict scaling to the assignment of numbers that, at a minimum, have the property of order (Magnusson, 1967; McDonald, 1999). On the other hand, some experts might prefer an even more restrictive definition that requires the use of scalars (Wright, 1997). This is another controversy in the measurement literature that will not be resolved here. It should be pointed out that for some authors, the terms *scaling* and *measurement* are synonymous (Bartholomew, 1996).

In a well-known framework, Stevens (1946) identified four levels of measurement. In the standard definition of measurement, the assignment of numbers to observations of behaviors is said to be "rule governed." In most cases, these "rules" refer to the scales of measurement proposed by Stevens (1946, 1951). Stevens's measurement scales are "rules" in that they suggest how certain properties of numerals might be linked to particular types of behavioral observations associated with psychological attributes. Table 2.1 integrates these levels of measurement with the fundamental numerical properties outlined earlier.

Nominal Scales

The most fundamental level of measurement is the nominal scale of measurement. In a **nominal scale**, numerals that have the property of identity are used to label observations in which behaviors have been sorted into categories according

to some psychological attribute. For example, we can "measure" biological sex by sorting people into two categories—males and females, represented as Group 0 and Group 1, respectively. Similarly and as described earlier, children in a classroom might be sorted into groups based on the presence or absence of behavioral problems, with the numeral 1 identifying the children with behavioral problems and the numeral 2 identifying the children without behavioral problems. As long as we can be sure that the groups are mutually exclusive and exhaustive, our only concern is our ability to correctly sort the children into the groups.

It is important to distinguish nominal scale labels, as used in the above example, from labels used to identify or name individuals. Nominal scale labels are used to identify *groups* of people who share a common attribute that is not shared by people in other groups. In contrast, numerals that are used to identify individuals, such as Social Security numbers, are generally not intended to establish group membership. The distinction can be clouded, however, when numerals are assigned to individuals in some systematic fashion. For example, it is possible to sort people into groups according to their year of birth, and numerals on the jerseys of individual football players might be used to differentiate people who play different positions on a team (see Lord, 1953, for a humorous discussion of this problem). The important point is that when using numerals to identify people, you need to be clear about your intent. That is, are you using the numerals to identify group membership (as is the case for the nominal level of measurement) or as labels that essentially serve as names for individuals?

Ordinal Scales

As its name implies, an **ordinal scale** defines measurement in terms of numerals that have the property of order. That is, ordinal scales produce ranks in which people are ordered according to the amounts of some attribute that they possess. For example, the members of an athletic team might be ranked according to their athleticism. The team's coaches might create the rankings based on their own

TABLE 2.1 ⬢ Associations Between Numerical Principles and Levels of Measurement

Principle	Level of Measurement			
	Nominal	Ordinal	Interval	Ratio
Identity	X	X	X	X
Order		X	X	X
Quantity			X	X
Absolute zero				X
Example	Biological sex	Class rank	Temperature	Distance

judgments of the athleticism of each team member. The player judged to have the most athleticism might be assigned the numeral 1, the next most athletic player the numeral 2, and so on.

As described earlier, numerals used in this sense are simply labels indicating the relative position of people with regard to the relative levels of the attribute being measured (e.g., athleticism). However, there is no attempt to determine how *much* of that attribute is actually possessed by each person. The numerals simply indicate that one person has more or less of the attribute than another person.

Although this level of measurement conveys more information than a nominal level of measurement, it is limited. Imagine two different athletic teams, one team composed of professional athletes and one team of high school athletes. Players on each team are ranked according to athleticism by their respective coaches; each professional player is ranked in comparison with the other professionals, and each high school player is ranked in comparison with the other students. The most athletic professional player and the most athletic high school student are each given a ranking of "1" by their coaches. Their "scores" tell us that these two people are the most athletic members of their teams, but the fact that both of them scored a "1" on the athleticism rankings clearly does not imply that they are equally athletic. Obviously, we should not infer that the most athletic high school player is as athletic as the most athletic professional player. Such quantitative comparisons would require a measurement that has the property of quantity.

Interval Scales

The property of quantity characterizes two remaining scales of measurement. That is, both interval scales and ratio scales are based on numbers that represent quantitative differences between people in the amount of the attribute being measured. However, the difference between the two scales rests primarily on the meaning of zero.

Interval scales have an arbitrary zero. As noted earlier, temperature expressed in Celsius (or Fahrenheit) units is a classic example of an attribute (temperature) measured on an interval scale. A temperature of 0°C (or 0°F, Fahrenheit) is arbitrary because it does not represent the absence of any attribute. It does not represent the complete absence of heat.

In interval scales, the size of the unit of measurement is constant and additive, but the scale does not allow multiplicative interpretations. You can add 2° to 40° and get 42°, or you can add 2° to 80° and get 82°. In each case, a 2° change on the thermometer represents the same change in the underlying amount of heat. That is, the amount of heat required to change the temperature from 40 to 42°C is the same as the amount of heat required to change it from 80 to 82°C. This means that the size of the units on the Celsius scale are additive and constant. However, it is not appropriate to interpret a temperature of 80°F as having "twice as much

heat" or being "twice as warm" as 40°F. In that sense, multiplication (and division) on the units of the Celsius scale does not produce meaningful results, in terms of reflecting ratios of "amount of heat."

As discussed later, many psychological tests are used and interpreted as if they are based on an interval scale of measurement. For example, the vast majority of intelligence tests, personality tests, achievement tests, developmental tests, and many other types of psychological assessments are treated as if they are interval scales. By assuming that a test's scores have the property of quantity and that the units of measurement have a constant magnitude, test users can make many research-based and practice-based applications of test scores.

Unfortunately, according to many measurement experts, few psychological tests can be truly said to yield interval-level scores (Ghiselli et al., 1981). Scores obtained from some well-known academic assessment tests, such as the SAT and the American College Testing (ACT) program, are probably on an interval scale. However, it has been argued that scores from the vast majority of psychological tests are, in fact, not on an interval scale. We will return to the implications of this issue soon.

Ratio Scales

In contrast to interval scales with an arbitrary 0 point, **ratio scales** have an absolute 0 point. For example, measures of physical distance are ratio scales. We might intend to measure the distance between two objects, and we find that the distance is 0. In such a case, the zero indicates a true "absence of distance." That is, the zero indicates an absence of the feature being measured.

Ratio scales are considered a "higher" level of measurement than interval, ordinal, and nominal scales, because they provide more information and allow for more sophisticated inferences. Specifically, ratio scales allow additivity as well as multiplicative interpretations in terms of ratios. For example, it is appropriate to interpret a distance of 80 miles as "twice as far" as a distance of only 40 miles.

This important issue has implications for our interpretations of the differences between objects. In applied settings, a ratio scale would allow a test user to make statements such as "Psychiatric Patient A is twice as mentally disturbed as Psychiatric Patient B." In research settings, a ratio scale would allow researchers to interpret the results of certain statistical procedures in terms of the underlying attributes being measured.

According to most testing experts, there probably are no psychological tests that yield ratio-level data. This might be surprising to those of you who are familiar with attempts to measure psychological attributes using standard measures. For example, reaction times (e.g., measured in milliseconds) are a common form of measurement in cognitive psychology, and they are becoming more popular in

other areas of psychology. Standard measures such as "time in milliseconds" are ratio-level measures. So why do some experts claim that there are no ratio-level psychological tests?

Remember that ratio scales have an absolute 0 point. A moment's reflection, however, will show that it is impossible for a person to respond to anything in 0 seconds (or milliseconds). The measuring device—for example, a stop clock—has an absolute zero, but a person's reaction time can never be zero. This is not to suggest that reaction time measures are poor measures of psychological processes. In fact, reaction times might be the most natural way to measure mental activity (Jensen, 2005). The point is that test users must distinguish between the zero associated with a measuring device and the zero associated with the psychological attribute being measured. Although a measuring instrument might have an absolute zero, this does not mean that the psychological attribute being measured has an absolute zero (Blanton & Jaccard, 2006).

SCALES OF MEASUREMENT: PRACTICAL IMPLICATIONS

As noted earlier, a test's scale of measurement can have important implications. Among behavioral researchers, it is commonly suggested that this issue can have implications for the meaningfulness of specific forms of statistical analysis. That is, it has been argued that some of the most common, fundamental, and familiar statistical procedures should be used only with measurements that are interval or ratio, not with nominal or ordinal measurements.

A variable's scale of measurement has important implications for the meaningfulness of certain descriptive statistics, such as a mean or correlation (see Chapter 3). For a fairly concrete example, consider a nominal variable such as hair color. Let's say we have a sample of 20 participants and we categorize each person in terms of hair color: Group 1 = blond (13 participants), Group 2 = black (4 participants), and Group 3 = brown (3 participants). If we used these numbers to compute the "average hair color," we would get a value of 1.5:

$$\frac{1+1+1+1+1+1+1+1+1+1+1+1+1+2+2+2+2+3+3+3}{20}=\frac{30}{20}=1.5$$

Although it is obviously *mathematically* possible to compute the mean of scores on a nominal variable, the question is whether it is *meaningful* to do so. What exactly does an "average hair color" mean, and more to the point, what would a value of 1.5 mean in this context? Does it mean that the average hair color is blondish-black, since 1.5 is between 1 and 2 (with those values representing blond and black in this context)? To make things worse, we could have chosen to categorize the groups

differently: Group 1 = black (4 participants), Group 2 = brown (3 participants), and Group 3 = blond (13 participants). This would be totally fine, since the numbers here (for a nominal scale) are simply labels. However, in this case, the average score would be 2.45, and again what would that even mean in this context—the average person now has blondish-brown hair? Our answer to the question of the "average hair color" changes depending on arbitrary choices about the numbers we use to represent hair color. This is a problem.

Hopefully, this example conveys the point that the meaningfulness of certain descriptive statistical concepts can depend heavily on the scale of measurement. Certain descriptive statistical concepts simply are not very meaningful when applied to scores based on certain scales of measurement.

In contrast to the relative clarity of descriptive statistics, there is significant disagreement about whether scales of measurement have clear implications for the use of parametric statistics (such as *t* tests or analysis of variance, which some of you might be familiar with). On one hand, many statistics textbooks and researchers state or believe that parametric statistical procedures are appropriate only for interval or ratio scales. For example, Cohen (2001) states that "parametric statistics are truly valid only when you are dealing with interval/ratio data" (p. 7). On the other hand, many experts reject this view. For example, Howell (1997) asserts that "the underlying measurement scale is not crucial in our choice of statistical techniques" (p. 8). Rather more pointedly, Gaito (1980) suggests that those who believe that scaling has direct implications for the appropriateness of parametric (vs. nonparametric) procedures "apparently do not read the statistical journal literature, inasmuch as a number of articles on this topic showed clearly that measurement scales are not related to statistical techniques" (p. 564). Reflecting on such disagreements, Maxwell and Delaney (2004) admit that "level of measurement continues to be controversial as a factor that might or might not influence the choice between parametric and nonparametric approaches" (p. 143).

Regardless of ambiguities and disagreements, behavioral researchers generally treat most tests and measures as having an interval level of measurement. Particularly for aggregated scores obtained from multi-item scales, researchers assume that scores are "reasonably" interval level. For very brief or single-item scales, this assumption is more tenuous. In such cases, researchers should either consider alternative analytic strategies or acknowledge the potential problem.

ADDITIONAL ISSUES REGARDING SCALES OF MEASUREMENT

Stevens's rules for assigning symbols, including numbers, to behavioral observations used as tests should be taken as heuristic devices rather than as definitive

and exhaustive. In fact, other authors have proposed additional levels of measurement, with corresponding rules for creating scales. For example, Coombs (1950) argues for a level of measurement between nominal and ordinal levels and another between ordinal and interval levels. Similarly, Mosteller and Tukey (1977) present a set of six types of scales, and Michell (1986) suggests that, by some definitions, only interval-level and ratio-level scales count as "measurements." Moreover, counting can be considered a level of measurement in its own right, and when used to quantify a psychological attribute, it can be thought of as a measure with an absolute zero and a fixed nonarbitrary unit of measurement (the number 1). Our discussion focused on Stevens's framework because it is the most common such framework and because it provides a reasonable foundation for understanding the key issues as related to psychometrics.

Another point is that, although they are often used to reflect nominal scales, dichotomous variables that have been assigned binary codes (such as 0 and 1) can sometimes be thought of as producing interval-level data. If you have reason to believe that discrete dichotomous categories were created based on some underlying quantitative psychological attribute, then the binary codes possess all of the properties associated with quantity. For example, imagine that you have a measure of depression. You give the test to a large group of people and sort people into two categories based on their scores—those who are depressed and those who are not depressed. If you assigned numerical codes to these categories, then the numbers can be seen as reflecting the differences in the amount of depression in the people in the two categories. In this case, the values could be conceptualized on an interval scale. On the other hand, if a sort into categories is not based on a quantitative attribute, then it would not make sense to treat the codes as having quantitative properties. An example might be a case in which people are sorted into categories based on whether they have used an illegal drug.

TECHNICAL APPENDIX: R SYNTAX

As noted in the preface, many chapters in the fourth edition of this book now include a technical appendix that shows readers how to carry out psychometric procedures described in those chapters. The goal is to help readers turn psychometric concepts into psychometric action. It is one thing to describe the conceptual basis of psychometrics and even to provide or illustrate the statistical equations that underlie those concepts. It is another thing to show how to apply those concepts and derive real psychometric information based on real data using real statistical software.

These syntax appendices will highlight the use of R programming language for psychometric analyses, for the reasons noted in the preface (e.g., no-cost availability of R, wide range of R's capabilities, vast popularity of R in data science more broadly, etc).

These appendices will assume some familiarity with R, and they are not intended to provide a general introduction to R. They will focus on getting test data into R, loading relevant R packages, using functions relevant to psychometric analysis, and explaining output from those analyses. For readers who need a more general introduction to R, there are many online resources available, including tutorials, online classes, and videos.

Each appendix will include one or more "chunks" of syntax that carry out one or more procedures. Data will be available via the SAGE website at **edge.sagepub.com/furr4e**. Readers are encouraged to download the relevant data and try using the syntax to analyze those data. You should be able to obtain the same results reported and interpreted in the appendices. Each chunk of syntax will have pieces of syntax numbered, and each piece will be explained briefly in the text. In addition, the output from each piece of syntax will be presented and explained.

Chapter 2 discussed issues with scaling. Most of these issues are conceptual and do not directly involve analysis, although they do have significant implications for psychometric analyses.

That said, this appendix will present syntax relevant to a few issues discussed in Chapter 2:

1. Accessing the relevant R data set. This includes data from approximately 300 people who completed the Moral Relativism and Moral Tolerance scales (Collier-Spruel et al., 2019), along with some demographic information.
2. Viewing item-level test data.
3. Getting a sense of the types of variables (items) in the data set.
4. Dealing with "factors" that represent nominal variables.

Syntax

```
#################################################################
## Syntax for Chapter 2 of "Psychometrics: An Introduction"     ##
##                                                              ##
## Demonstrates how to access an R data set                     ##
## Data are from the Moral Relativism and Moral Tolerance scales ##
## (Collier-Spruel et al., 2019) along with some demographics   ##
## Shows how to view data                                       ##
## Helps get a sense of the types of variables (items)          ##
## Deals with "factors" that represent nominal variables        ##
#################################################################

#1. Open the data set
load(file="C:…../…./…../…../…./MRMTch2.Rdata"
```

```
#2. Obtain a snapshot of the data (variable names and top few rows of data set)
names(MRMTch2)
head(MRMTch2)

#3. See each variable's type, value labels, etc.
str(MRMTch2)

#4. View value labels for the nominal variable 'Gender' and the item "MRS_1"
attr(MRMTch2$Gender,"value.labels")
attr(MRMTch2$MRS_1,"value.labels")

#5. Obtain a frequency count for the Gender variable
table(MRMTch2$Gender)

#6. Determine the "mean" of 'Sex,' even though it's not meaningful in this context
mean(MRMTch2$Gender)

#7. Change the Sex variable from numeric to "factor"
MRMTch2$Gender <- factor(MRMTch2$Gender,
                         levels = c(1,2),
                         labels = c("Male", "Female"))
str(MRMTch2$Gender)
table(MRMTch2$Gender)
mean(MRMTch2$Gender)

#8. Save the revised data set under a new name
MRMTch3 <- MRMTch2
save(MRMTch3, file = "MRMTch3.Rdata")
```

Output

1. The load() function opens an R-formatted data set. You can download the MRMTch2.Rdata data set from the SAGE website. Be sure to revise the syntax to find the file in the directory in which you have saved it. This line of syntax produces no output.

2. The names() function prints (to the R "console" screen) the names of all the variables in the MRMTch2.Rdata data set. The head() function prints the first few rows of the data set. This provides a quick glance at the data. The following shows the full output of the names() function and the partial output of the head() function.

```
> names(MRMTch2)
 [1] "Partid"   "Gender"     "Age"  "Ethnicity"    "Race"  "Religion"
 [7]  "MRS_1"    "MRS_2"    "MRS_3"      "MRS_4"    "MRS_5"     "MRS_6"
[13]  "MRS_7"    "MRS_8"    "MRS_9"     "MRS_10"    "MTS_1"     "MTS_2"
[19]  "MTS_3"    "MTS_4"    "MTS_5"      "MTS_6"    "MTS_7"     "MTS_8"
[25]  "MTS_9"   "MTS_10"    "SWL_1"      "SWL_2"    "SWL_3"     "SWL_4"
[31]  "SWL_5"      "MRS"      "MTS"        "SWL"
 > head(MRMTitemsch2)
  Partid Gender Age Ethnicity Race Religion MRS_1 MRS_2 MRS_3 MRS_4 [etc]…
1      1      1  18         2    5        1     1     3     1     1…[etc]…
2      2      1  18         2    5        1     3     4     2     3 [etc]…
3      3      2  18         2    3        1     2     2     3     1 [etc]…
4      4      2  18         2    5        8     4     5     5     5 [etc]…
5      5      2  18         2    5        2     5     5     5     5 [etc]…
6      6      2  17         2    5        2     5     4     4     3 [etc]…
```

There are 31 variables in the data set, including a participant identification number, five demographic variables, 10 items from the Moral Relativism Scale (MRS), 10 items from the Moral Tolerance Scale (MTS), five items from the Satisfaction With Life Scale (SWLS), and three "total scores" for the three scales. The head() output shows that all the demographic variables (gender, age, etc) are coded numerically. Responses to the MRS, MTS, and SWLS items are on 5-point scales of agreement.

3. The str() function reveals some key information about each variable, including the variable "type" and the meaning of the numerical values (note, "str" refers to "structure"). Here is partial output from that function:

```
> str(MRMTch2)
'data.frame':294 obs. of  34 variables:
 $ Partid   : num  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender   : num  1 1 2 2 2 2 2 1 2 2 ...
  ..- attr(*, "value.labels")= Named chr  "4" "3" "2" "1"
  .. ..- attr(*, "names")= chr  "Prefer not to answer" "Do not identify as
male or female" "Female" "Male"
…..
 $ MRS_1    : num  1 3 2 4 5 5 5 4 2 4 ...
  ..- attr(*, "value.labels")= Named chr  "5" "4" "3" "2" ...
  .. ..- attr(*, "names")= chr  "Strongly Agree" "Somewhat Agree" "Neither
Agree nor Disagree" "Somewhat Disagree" ...
…
```

In R parlance, the MRMT data set is called a "data frame." It includes responses and scores from 294 participants on each of the 34 variables. The gender variable is numeric, with four possible values ("prefer not to answer," "do not identify as male or female," "female," and "male"). As illustrated by the first item on the MRS ("MRS_1"), each item on the MRS, MTS, and SWL is also numeric with five possible values (e.g., "strongly disagree," "somewhat disagree," etc.)

4. This use of attr() tells us the meaning of each numerical value of a variable. This helps us interpret the data. Here is the output:

```
> attr(MRMTch2$Gender,"value.labels")
       Prefer not to answer       Do not identify as male or female
                        "4"                                      "3"
                     Female                                     Male
                        "2"                                      "1"
> attr(MRMTch2$MRS_1,"value.labels")
     Strongly Agree       Somewhat Agree      Neither Agree nor Disagree
                "5"                  "4"                             "3"
  Somewhat Disagree    Strongly Disagree
                "2"                  "1"
```

This tells us, for example, that a gender value of 1 represents male, while a gender value of 2 represents female. Similarly, for all of the MRS, MTS, and SWL scales (as illustrated by MRS_1), 1 indicates strongly disagree, 2 indicates somewhat disagree, and so on. Thus, "higher scores" represent greater levels of agreement to the MRS, MTS, and SWL items.

5. The table() function provides frequency counts for each level of a variable (i.e., the number of people who fall into each category, who make each response, or who get a given score). Here, we request frequency counts for the nominal gender variable:

```
> table(MRMTch2$Gender)

  1   2
126 168
```

These results show that 126 participants are in gender category 1, which we know from the attr() output is "male"; 168 participants are in gender category 2, which we know is "female." Although participants also had the option of responding "prefer not to answer" or "do not identify as male or female" to the gender question, apparently none chose either category.

6. The mean() function computes the average, or arithmetic mean, of a set of values. Here, it is used to demonstrate that R will compute the mean of any variable that is encoded as a "numeric" type of variable. Gender is, in principle, a nominal variable that is coded here with numerical values. Because it is currently encoded as a numeric variable (see the earlier str() output), R will compute its mean. In this case, however, the mean is not very interpretable (actually in this very special case of two actual values of the nominal variable coded in this specific way, the mean reveals that approximately 57.1% of the sample identified themselves as female). This line of code is here simply to demonstrate that R will compute the mean, even if it's not psychologically meaningful.

```
> mean(MRMTch2$Gender)
[1] 1.571429
```

7. Because gender is a nominal variable, we will reclassify it as a "factor" for R. In R, a factor is treated as a special type of variable that does not have quantitative properties. That is, even if its values are numerals (e.g., values of "1" and "2"), those values are treated simply as labels without quantitative properties. In this use of the factor(), we reclassify gender and we recode its values to "male" and "female" rather than "1" and "2." This recoding is not necessary, but it simplifies some subsequent work that we might do, and it ensures that we know the meaning of each category. In addition, we will run the str() function on the reclassified variable to make sure that it has indeed been reclassified as a factor. We will also rerun the table() function to make sure that the frequency counts map onto the categories as they should. Finally, we run the mean() function on the reclassified factor variable.

```
> MRMTch2$Gender <- factor(MRMTch2$Gender,
+                          levels = c(1,2),
```

```
+                              labels = c("Male", "Female"))
> str(MRMTch2$Gender)
 Factor w/ 2 levels "Male","Female": 1 1 2 2 2 2 2 1 2 2 ...
> table(MRMTch2$Gender)

  Male Female
   126    168
> mean(MRMTch2$Gender)
[1] NA
Warning message:
In mean.default(MRMTch2$Gender) :
  argument is not numeric or logical: returning NA
```

The output tells us that the gender variable has successfully been recoded as a factor, with category levels as we expected. In addition, the "NA" result from the mean() function indicates that R will not compute the mean of a factor variable. Similarly, the warning tells us explicitly that the argument (i.e., the line of syntax trying to compute the mean of gender as a factor variable) is dealing with a variable that is not numeric.

8. Since we have changed the data set by changing the way the gender variable is classified, we will save the revised data set under a new name. This may not be necessary in "real" analysis, but here it can keep our data well organized as we move from chapter to chapter. Because we will use the revised data set for our R syntax examples in Chapter 3, we save it here as "MRMTch3.Rdata."

Summary

This chapter has addressed a variety of important theoretical issues in an attempt to outline the foundations of psychological measurement. The core goal of scaling in the context of this book is to link numerical values to people's psychological attributes. As outlined in this chapter, fundamental issues in scaling concern (a) the connection between the observations of a behavior and numerical symbols and (b) the degree to which this connection is made in such a way that the symbols identify the real differences that exist between the behaviors under observation.

The scaling of people's psychological attributes faces challenges that partly arise from the fact that psychological attributes (e.g., traits, abilities, skills, attitudes) are not directly observable. Therefore, in many cases of psychological measurement, psychologists are likely to rely on nonquantitative measures of psychological attributes or simply assume that quantitative measurement models work well enough to approximate quantities of psychological attributes. Nevertheless, all psychological scaling procedures have one feature in common—they are all procedures for representing the differences among people. The next chapter discusses the statistical concepts that are used to quantify these individual differences.

Suggested Readings

This is the classic article on psychological scaling:

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

This is an article that discusses many variations on Stevens's scales of measurement:

Coombs, C. H., Raiffa, H., & Thrall, R. M. (1954). Some views on mathematical models and measurement theory. *Psychological Review*, *61*(2), 132–144. https://doi.org/10.1037/h0063044

For an accessible and interesting history of zero:

Seife, C. (2000). *Zero: The biography of a dangerous idea*. Penguin.

For an in-depth look at different ways of conceptualizing measurement and the use of numbers in science:

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, *100*(3), 398–407. https://doi.org/10.1037/0033-2909.100.3.398

The following is a good discussion of one of the most fundamental problems of measurement in psychology:

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*(1), 27–41. https://doi.org/10.1037/0003-066x.61.1.27

Student Resources

This text includes access to datasets and files in R for the Technical Appendix. To learn more, visit **edge.sagepub.com/furr4e**.