PSYCHOMETRICS AND THE IMPORTANCE OF PSYCHOLOGICAL MEASUREMENT

Your life has probably been shaped, in part, by psychological measurement. Whether you are a student, a teacher, a parent, a psychologist, a physician, a nurse, a patient, a lawyer, a police officer, or a businessperson, you have taken psychological tests, your family members have taken psychological tests, or you have been affected by people who have taken psychological tests. These tests can affect our education, our careers, our family life, our safety, our health, our wealth, and, potentially, our happiness. Indeed, almost every member of an industrialized society is affected by psychological measurement at some point in his or her life—both directly and indirectly.

It is even fair to say that, in extreme situations, psychological measurement can have life or death consequences. Although this might seem overly sensational, far-fetched, and perhaps even simply wrong, it is true. The fact is that in some states and nations, prisoners who have severe cognitive disabilities cannot receive a death penalty. For example, in the state of North Carolina, the General Assembly states that "no defendant with an intellectual disability shall be sentenced to death" (N.C. Gen. Stat. § 15A-2005, 2019); it defines intellectual disability, in part, as general intellectual functioning that is "significantly subaverage." But what is "significantly subaverage" intellectual functioning, and how could we know whether a person's intelligence is indeed significantly subaverage?

These difficult questions are answered in terms of psychological tests. Specifically, the General Assembly states that significantly subaverage intellectual functioning is indicated by a score of 70 or below "on an individually administered, scientifically recognized standardized intelligence quotient test administered by a licensed

psychiatrist or psychologist." Put simply, if a person has an intelligence quotient (IQ) score below 70, then they might not be sentenced to death by the state of North Carolina; however, if a person has an IQ score above 70, then they can legally be put to death. Thus, although it might seem hard to believe, intelligence testing can affect whether men and women might live or die, quite literally. Of course, few consequences of psychological measurement are so dramatic, but they can indeed be real, long- lasting, and important.

Given the important role of psychological tests in our lives and in society more generally, those tests must have extremely high quality. If testing has such robust implications, then it should be done with the strongest possible tools and procedures.

This book is about understanding whether such tools and procedures are indeed strong—how to determine whether a test produces scores that are psychologically meaningful and trustworthy. In addition, the principles and concepts discussed in this book are important for creating tests that are psychologically meaningful and trustworthy. These principles and concepts are known as psychometrics.

WHY PSYCHOLOGICAL TESTING MATTERS TO YOU

Considering the potential real-life impact of psychological testing, you need to understand the basic principles of psychological measurement. Whether you wish to be a practitioner of behavioral science, a behavioral researcher, or a sophisticated member of modern society, your life is likely to be affected by psychological measurement.

You might be considering a career involving psychological measurement. Some of you might be considering careers in the practice or application of a behavioral science. Whether you are a clinical psychologist, a school psychologist, a human resources director, a university admissions officer, or a teacher, your work might require you to make decisions on the basis of scores obtained from some kind of psychological test. When a patient responds to a psychopathology assessment, when a student completes a test of cognitive ability or academic aptitude, or when a job applicant fills out a personality inventory, there is an attempt to measure some type of psychological characteristic.

In such cases, test users have a responsibility to examine and interpret important information about the meaning and quality of the tests they use. Without a solid understanding of the basic principles of psychological measurement, test users risk misinterpreting or misusing the information derived from psychological tests. Such misinterpretation or misuse might harm patients, students, clients, employees, and applicants, and it can lead to lawsuits for the test user. Proper test interpretation and use can be extremely valuable for test users and beneficial for test takers.

Some of you might be considering careers in behavioral research. Whether your area is psychology, education, or any other behavioral science, measurement is at the heart of your research process. Whether you conduct experimental research, survey research, or any other kind of quantitative research, measurement is at the heart of your research process. Whether you are interested in differences between individuals, changes in people across time, differences between genders, differences between classrooms, differences between treatment conditions, differences between teachers, or differences between cultures, measurement is at the heart of your research process. If something is not measured or is not measured well, then it cannot be studied with any scientific validity. If your goal is meaningful and accurate interpretation of your research findings, then you must evaluate critically the measurements that you have collected in your research.

As mentioned earlier, even if you do not pursue a career involving psychological measurement, you will almost surely face the consequences of psychological measurement, either directly or indirectly. Applicants to graduate school and various professional schools might be accepted (or not) partially on the basis of tests of knowledge and achievement. Job applicants might be hired (or not) partially on the basis of scores on personality tests. Employees might be promoted (or passed over for promotion) partially on the basis of supervisor ratings of psychological characteristics such as attitude, competence, or collegiality. Parents must cope with the consequences of their children's educational testing. People seeking psychological services might be diagnosed and treated partially on the basis of their responses to various psychological measures.

Even more broadly, our society receives information and recommendations based on research findings. Whether you are (or will be) an applicant, an employee, a parent, a psychological client, or an informed member of society, the more knowledge you have about psychological measurement, the more discriminating a consumer you will be. You will have a better sense of when to accept or believe test scores, when to question the use and interpretation of test scores, and what you need to know to make such important judgments.

Given the widespread use and importance of psychological measurement, it is crucial to understand the properties affecting the quality of such measurements. This book is about the important *attributes of the instruments* that psychologists use to measure psychological attributes and processes.

This book addresses several fundamental questions related to the logic, development, evaluation, and use of psychological measures.

- What does it mean to attribute scores to characteristics such as intelligence, memory, self-esteem, shyness, happiness, or executive functioning?

- How do you know if a particular psychological measure is trustworthy and interpretable?
- How confident should you be when interpreting an individual's score on a particular psychological test?
- What kinds of questions should you ask to evaluate the quality of a psychological test?
- What are some of the different kinds of psychological measures?
- What are some of the challenges to psychological measurement?
- How is the measurement of psychological characteristics similar to and different from the measurement of physical characteristics of objects?
- How should you interpret some of the technical information regarding psychological measurement?

The goal of this book is to address these kinds of questions in a way that provides a deep and intuitive understanding of psychometrics. This book is intended to help you develop the knowledge and skills needed to evaluate psychological tests intelligently. Psychological testing plays an important role in psychological science and in psychological practice, and it plays an increasingly important role in our society.

Hopefully, this book helps you become a more informed consumer and, possibly, producer of psychological information.

OBSERVABLE BEHAVIOR AND UNOBSERVABLE PSYCHOLOGICAL ATTRIBUTES

People use many kinds of instruments to measure observable properties of the physical world. For example, if you want to measure the length of a piece of lumber, then you might use a tape measure. People also use various instruments to measure the properties of the physical world that are not directly observable. For example, clocks are used to measure time, and voltmeters are used to measure the change in voltage between two points in an electric circuit.

Similarly, psychologists, educators, and others use psychological tests as instruments to measure observable events in the physical world. In the behavioral sciences, these observable events are typically some kind of behavior, and behavioral measurement is usually conducted for two purposes. Sometimes, psychologists measure a behavior because they are interested in that specific behavior in its own right. For example, some psychologists have studied the way facial expressions affect the perception of emotions. The Facial Action Coding System (FACS; Ekman & Friesen,

1978) was developed to allow researchers to pinpoint movements of very specific facial muscles. Researchers using the FACS can measure precise "facial behavior" to examine which of a person's facial movements affect other people's perceptions of emotions. In such cases, researchers are interested in the specific facial behaviors themselves; they do not interpret them as signals of some underlying psychological process or characteristics.

Much more commonly, however, behavioral scientists observe human behavior as a way of assessing unobservable psychological attributes such as intelligence, depression, knowledge, aptitude, extroversion, or ability. In such cases, they identify some type of observable behavior that they think represents the particular unobservable psychological attribute, state, or process. They then measure the behavior and try to interpret those measurements in terms of the unobservable psychological characteristics that they think are reflected in the behavior. In most but not all cases, psychologists develop psychological tests as a way to sample the behavior that they think reflects the underlying psychological attribute.

For example, suppose that we wish to identify which of two students, Sam and William, had greater working memory. To do this, we must measure both students' working memories. Unfortunately, there is no known way to observe directly working memory—we cannot directly "see" memory inside a person's head. Therefore, we must look for something that we can see (e.g., some type of behavior) and that could indicate how much working memory someone has. For example, we might ask the students to repeat a series of numbers presented to them rapidly. If the two students differ in their performance on this task, then we might assume that they differ in their working memory. That is, if we observe a difference in their behavior, then we interpret it as revealing a difference in their working memory. If Sam repeats more of the numbers than William, then we might conclude that Sam's working memory is greater than William's. This conclusion requires that we make an inference—that an observable behavior, the number of recalled numbers, is systematically related to an unobservable mental attribute, working memory.

There are several things to notice about this attempt to measure working memory. First, we make an inference from an observable behavior to an unobservable psychological attribute. That is, we assume that the particular behavior that we observe reflects or reveals working memory. If this inference is reasonable, then we would say that our interpretation of the behavior has a degree of *validity*. Although validity is a matter of degree, if the scores from a measure seem to be actually measuring the mental state or mental process that we think they are measuring, then we say that our interpretation of scores on the measure is valid.

Second, for our interpretation of "number recall" scores to be considered valid, the recall task must be theoretically linked to working memory. It would not have made theoretical sense, for example, to measure working memory by timing William's

and Sam's running speed in a footrace. In the behavioral sciences, we often make an inference from an observable behavior to an unobservable psychological attribute. Therefore, measurement in psychology often, but not always, involves some type of theory linking a psychological characteristic, process, or state to an observable behavior that is thought to reflect differences in that psychological attribute.

There is a third important feature of our attempt to measure working memory. Working memory is itself a theoretical concept. When measuring working memory, we assume that working memory is more than a figment of our imagination. Psychologists, educators, and other social scientists often use theoretical concepts such as working memory to explain differences in people's behavior. Psychologists refer to these theoretical concepts as *hypothetical* **constructs** or **latent variables**. They are theoretical psychological characteristics, attributes, processes, or states that cannot be directly observed, and they include things such as knowledge, intelligence, self-esteem, attitudes, hunger, memory, personality traits, depression, and attention. The operations or procedures that we use to measure these hypothetical constructs, or for that matter to measure anything, are called **operational definitions**. In our example, the number of recalled numbers was used as an operational definition of some aspect of working memory, which itself is an unobservable hypothetical construct.

You should not be dismayed by the fact that psychologists, educators, and other social scientists rely on unobservable hypothetical constructs to explain human behavior. This reliance is true of many branches of science. Measurement in the physical sciences, as well as the behavioral sciences, often involves making inferences about unobservable events, things, and processes based on observable events. As an example, physicists write about four types of "forces" that exist in the universe: (1) the strong force, (2) the electromagnetic force, (3) the weak force, and (4) gravity. Each of these forces is invisible, but their effects on the behavior of visible events can be seen. For example, objects do not float into space off the surface of our planet. Theoretically, the force of gravity is preventing this from happening. Physicists have built equipment to create opportunities to observe the effects of some of these forces on observable phenomena. In effect, the equipment is used to create scenarios in which to measure observable phenomena that are believed to be caused by the unseen forces.

To be sure, the sciences differ in the number and nature of unobservable characteristics, events, or processes that are of concern to them. Some sciences might rely on relatively few, while others might rely on many. Some sciences might have strong empirical bases for their unobservable constructs (e.g., gravity), while others might have weak empirical bases (e.g., penis envy). Nevertheless, all sciences rely on unobservable constructs to some degree, and they all measure those constructs by measuring some observable events or behaviors.

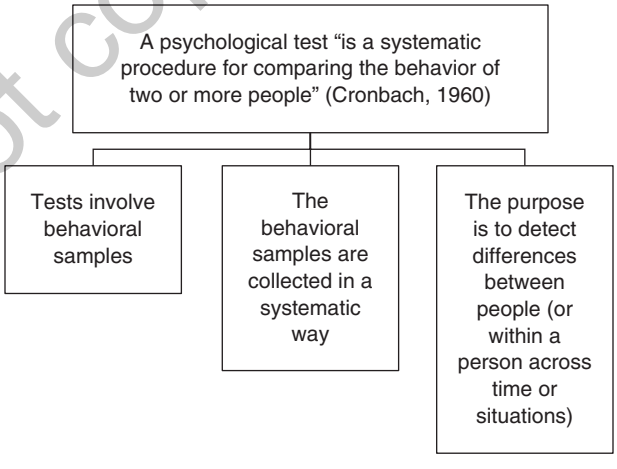
PSYCHOLOGICAL TESTS: DEFINITION AND TYPES

What Is a Psychological Test?

According to Cronbach (1960), a psychological test "is a systematic procedure for comparing the behavior of two or more people" (p. 21). As shown in Figure 1.1, this definition includes three important components: (1) tests involve behavioral samples of some kind, (2) the behavioral samples must be collected in some systematic (i.e., clear and standardized) way, and (3) the purpose of the tests is to detect differences between people. The third component could be modified to include a comparison of performance by the same individuals at different points in time or in different situations, but otherwise the definition is appealing. This appeal is based on several important features.

One appealing feature of the definition is its generality. The idea of a test is sometimes limited to paper-and-pencil tests, but psychological tests can come in many forms. For example, the Beck Depression Inventory–II (BDI-II; Beck et al., 1996) is a fairly traditional 21-item paper-and-pencil test designed to measure depression. People who take the test read each question and then choose an answer from one of several supplied answers. A person's degree of depression is evaluated by counting the number of answers of a certain type that they gave to the questions. The BDI is clearly a test, but other methods of systematically sampling behavior are also tests. For example, in laboratory situations, researchers ask participants to respond in various ways to well-defined stimulus events; participants might be asked to watch for a particular

FIGURE 1.1 ⬢ Cronbach's Definition of a Psychological Test, With Three of Its Key Components Emphasized



visual event and respond by pressing, as quickly as possible, a response key. In other laboratory situations, participants might be asked to make judgments regarding the intensity of stimuli such as sounds. By Cronbach's definition, these are also tests.

The generality of Cronbach's definition also extends to the type of information produced by tests. Some tests produce numbers that represent the amount of some psychological attribute possessed by a person. For example, the U.S. National Assessment of Educational Progress (NAEP; http://nces.ed.gov/nationsreportcard/reading/whatmeasure.aspx) uses statistical procedures to select test items that, at least in theory, produce data that can be interpreted as reflecting the amount of knowledge or skill possessed by children in various academic areas, such as reading. Other tests produce categorical data—people who take the test can be sorted into groups based on their responses to test items. The House-Tree-Person Test (Burns, 1987) is an example of such a test. Children who take this test are asked to draw a house, a tree, and a person. The drawings are evaluated for certain characteristics, and on the basis of these evaluations, children can be sorted into groups (however, this procedure might not be "systematic" in Cronbach's terms). Chapter 2 discusses more about the types data produced by psychological tests.

Another extremely important feature of Cronbach's definition concerns the general purpose of psychological tests. Specifically, tests must be capable of comparing the behavior of different people (*interindividual differences*) or the behavior of the same individuals at different points in time or under different circumstances (*intraindividual differences*). The purpose of measurement in psychology is to identify and, if possible, quantify such interindividual or intraindividual differences. This purpose is a fundamental theme throughout this book, and we will return to it in every chapter. Inter- and intraindividual differences on test performance contribute to test score variability, a necessary component of any attempt to measure any psychological attribute.

Types of Tests

There are tens of thousands of psychological tests in the public domain (Educational Testing Service, 2016). These tests vary from each other along dozens of different dimensions, some of which are reflected in Table 1.1.

TABLE 1.1 ⬢ Some Key Ways in Which Psychological Tests Differ

Differences	Examples
Content	Aptitude, achievement, intelligence, personality, etc.
Response required	Open ended vs. closed ended
Method of administration	Individual vs. group
Use	Criterion refenced vs. norm referenced
Timing	Speeded vs. power
The meaning of "indicators"	Reflective/effect vs. formative/causal

For example, tests can vary in content: There are achievement tests, aptitude tests, intelligence tests, personality tests, attitude surveys, and so on. Tests also vary with regard to the type of response required: There are **open-ended tests**, in which people can answer test questions by saying anything they want in response to the questions on the test, and there are **closed-ended tests**, which require people to answer questions by choosing among alternative answers provided in the test. Tests also vary according to the methods used to administer them. Some are individually administered, in which one person administers the test to one test taker at a time. Other tests can be administered to multiple people all at the same time.

Another major distinction concerns the intended purpose of test scores. Psychological tests are often categorized as either *criterion referenced* (also called domain referenced) or *norm referenced* (Glaser, 1963). **Criterion-referenced tests** are most often seen in settings in which a decision must be made about a person's skill level. In those settings, a cutoff test score is established as a criterion, and it is used to sort people into two groups: (1) those whose performance exceeds the criterion score and (2) those whose performance does not. In contrast, **norm-referenced tests** are usually used to understand how a person compares with other people. This is done by comparing a person's test score with scores from a **reference sample** or **normative sample**. A reference sample is typically a sample of people who complete a test, and the sample is thought to be representative of some broader population of people. Thus, a person's test score can be compared with the scores obtained from the people in the reference sample, telling us, for example, whether the individual has a higher or lower score than the "average person" (and how much higher or lower) in the relevant population. Scores on norm-referenced tests can be valuable when the reference sample is representative of some population, when the relevant population is well defined, and when the person being tested is a member of the relevant population. In principle, none of these issues arise when evaluating a score on a criterion-referenced test.

In practice, the distinction between norm-referenced tests and criterion-referenced tests is often blurred. Criterion-referenced tests are always "normed" in some sense. That is, criterion cutoff scores are not determined at random. The cutoff score will be associated with a decision criterion based on some standard or expected level of performance of people who might take the test. Most of us have taken written driver's license tests. These are criterion-referenced tests because a person taking the test must obtain a score that exceeds some predetermined cutoff. The questions on these tests were selected to ensure that the average person who is qualified to drive has a good chance of answering enough of the questions to pass the test. The distinction between criterion- and norm-referenced tests is further blurred when scores from norm-referenced tests are used as cutoff scores. Institutions of higher education might have minimum SAT or American College Testing (ACT) score requirements for admission or for various types of scholarships. Public schools use cutoff scores from intelligence tests to sort children into groups. In some cases, the use of scores from norm-referenced tests can have life or death consequences,

as noted at the beginning of this chapter. Despite the problems with the distinction between criterion-referenced tests and norm-referenced tests, there are slightly different methods used to assess the quality of criterion-referenced and norm-referenced tests (Kane, 1986; Popham & Husek, 1969).

Yet another common distinction is between *speeded tests* and *power tests.* **Speeded tests** are time-limited tests. In general, people who take a speeded test are not expected to complete the entire test in the allotted time. Speeded tests are scored by counting the number of questions answered in the allotted time period. It is assumed that there is a high probability that each question will be answered correctly; each of the questions on a speeded test should be of comparable difficulty. In contrast, **power tests** are not time limited, and test takers are expected to answer all the test questions. Often, power tests are scored also by counting the number of correct answers made on the test. Test items must range in difficulty if scores on these tests are to be used to discriminate among people with regard to the psychological attribute of interest. As is the case with the distinction between criterion-referenced tests and norm-referenced tests, slightly different methods are used to assess the quality of speeded and power tests (Angoff, 1953; Cronbach & Warrington, 1951).

It is worth noting that most of the procedures outlined in this book are relevant mainly for scores based on what are called **reflective (or effect) indicators** (Bollen & Lennox, 1991). For example, scores on intelligence or personality tests are of this kind. A person's responses on an intelligence test are typically seen as being caused by his or her actual level of intelligence. That is, the hypothetical construct (i.e., intelligence) determines, in part, a person's responses to the items on the intelligence test, and these responses are seen as "indicators" of the construct. Such tests are very common in psychology. There are, however, different types of scores that are based on what are called **formative (or causal) indicators**. Socioeconomic status (SES) is the classic example. You could quantify a person's SES by quantitatively combining "indicators" such as her income, education level, and occupational status. In this case, the indicators are not viewed as being "caused" by the person's SES. Instead, the indicators of SES are, in part, exactly what define SES. A full discussion of the distinction between formative/effect and reflective/causal scores—or of the usefulness of the supposed distinction—is beyond the scope of this section (interested readers are directed to Bollen & Diamantopoulos, 2017a, 2017b; Bollen & Lennox, 1991; Diamantopoulos & Winklhofer, 2001; Edwards, 2011; Edwards & Bagozzi, 2000; Hardin, 2017; Howell et al., 2007; MacKenzie et al., 2005; Markus, 2018; Myszkowski et al., 2019; Rhemtulla et al., 2020). The goal here is simply to note the existence of this important distinction and to acknowledge that this book focuses on test scores derived from reflective/effect indicators—as is typical for most tests and measures used in psychology.

A brief note concerning terminology: Several different terms are often used as synonyms for the word *test.* The words *measure*, *instrument*, *scale*, *inventory*, *battery*,

schedule, and *assessment* have all been used in different contexts and by different authors as synonyms for the word *test.* This book will sometimes refer to tests as instruments and sometimes as measures. The word *battery* will refer to bundled tests, which are tests that are intended to be administered together but are not *necessarily* designed to measure a single psychological attribute. The word *measure* can be used as a verb, as in "The BDI was designed *to measure* depression." It is also often used as a noun, as in "The BDI is a good *measure* of depression." This book will use both forms of the term and rely on the context to clarify its meaning.

WHAT IS PSYCHOMETRICS?

Psychometrics

Just as psychological tests are designed to measure psychological attributes of people (e.g., anxiety, intelligence), **psychometrics** is the science concerned with evaluating the attributes of psychological tests. Three of these attributes will be of particular interest: (1) the type of information (in most cases, scores) generated by the use of psychological tests, (2) the reliability of data from psychological tests, and (3) issues concerning the validity of data obtained from psychological tests. The remaining chapters in this book describe the procedures that psychometricians use to evaluate these attributes of tests. This book addresses the process of testing to a much lesser extent, and it describes particular tests only when illustrating important principles and concepts.

Note that just as psychological attributes of people (e.g., anxiety) are most often conceptualized as hypothetical constructs (i.e., abstract theoretical attributes of the mind), psychological tests also have attributes that are represented by theoretical concepts such as validity or reliability. Just as psychological tests are about theoretical attributes of people, psychometrics is about theoretical attributes of psychological tests. Just as psychological attributes of people are unobservable and must be measured, psychometric attributes of tests are also unobservable and must be estimated. Psychometrics is about the procedures used to estimate and evaluate the attributes of tests.

A Brief History of Psychometrics

The field of psychometrics has been built on two key foundations. One foundation is the practice of psychological testing and measurement. As most textbooks in psychological testing point out (e.g., Dubois, 1970; Miller & Lovler, 2016), the practice of using formal tests (of some kind) to assess individuals' abilities goes back 2,000 or perhaps even 4,000 years in China, as applicants for governmental positions completed various exams. Psychological measurement increased in the 19th century as psychological science emerged and as researchers began systematically measuring various qualities and responses of individuals in experimental studies. The practice of psychological measurement increased even more dramatically in the

20th century, with the development of early intelligence tests and early personality inventories. Over the course of the past 100+ years, the number, kinds, and applications of psychological tests have exploded. With such development comes the desire to create high-quality tests and to evaluate and improve tests. This desire inspired the development of psychometrics as the body of concepts and tools to do this.

A second and related historical foundation is the development of particular statistical concepts and procedures. Starting in the 19th century, scholars began to develop ways of understanding and working with the types of quantitative information that are produced by psychological tests. Among the early pioneers of this work are scholars such as Charles Spearman, Karl Pearson, and Francis Galton, all making key contributions in the late 1800s and early 1900s. Galton in particular is sometimes considered the founding father of modern psychometrics. He had diverse scholarly interests, including—it should be acknowledged—an advocacy for the now-rejected theory of eugenics. However, it is Galton's, Spearman's, and Pearson's important conceptual and technical innovations that are relevant for our discussion. In fact, you might already be familiar with some of these—the standard deviation and the correlation coefficient (see Chapter 3), factor analysis (see Chapters 4 and 12), the use of the normal distribution (or "bell curve"; see Chapter 3) to represent many human characteristics, and the use of sampling for the purpose of identifying and treating measurement error. These crucial statistical concepts and tools were adopted quickly and sometimes developed explicitly in order to make sense out of the numerical information gathered through the use of psychological tests. We will examine such concepts and tools in detail in this book.

Based on the application of these new statistical tools to the evaluation of psychological tests, the field of psychometrics truly came into its own by the 1930s and 1940s. During this period, the journal *Psychometrika* began publication, the Psychometric Society was formed, the American Psychological Association created its "Division of Evaluation and Measurement," and scholars such as J. P. Guilford and L. L. Thurstone published field-defining texts (Jones & Thissen, 2007). By this time, many tenets of what is now known as classical test theory (CTT) had been articulated (see Chapters 5–7)—providing the foundation for the most widely known perspective on test scores and test attributes. Somewhat later (1970s), CTT was expanded into generalizability theory by Lee Cronbach and his colleagues (see Chapter 13). At approximately the same time (or a bit earlier, in the 1950s and 1960s), an alternative to CTT was emerging, leading to what's now known as item response theory (IRT; see Chapter 14). Also in the 1950s, the crucial concept of test validity was undergoing robust development and articulation, with additional important reconceptualizations in the 1990s—leading to the framework addressed in Chapters 8 and 9 (Angoff, 1988).

Over the past few decades, the field of psychometrics has expanded in all of these directions. CTT itself has evolved, as, for example, researchers recognize the limits

of commonly used indices of reliability. IRT has enjoyed increased attention as well, with the development of various models and applications. Moreover, as statistical tools such as structural equation modeling have evolved, researchers have discovered ways of using those tools to conceptualize and examine key psychometric concepts.

In sum, psychometrics, as a scientific discipline, is relatively young but has enjoyed a quick evolution and widespread application. From this point on, this book focuses very little on history, devoting attention instead to contemporary concepts, tools, and practices that have grown out of the pioneering work of Galton, Spearman, Pearson, Thurstone, Cronbach, and many others.

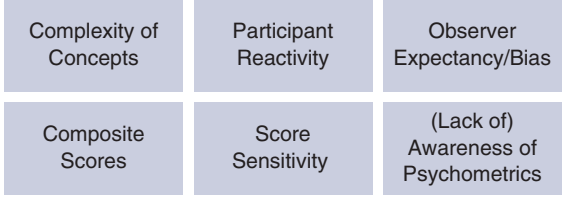
CHALLENGES TO MEASUREMENT IN PSYCHOLOGY

We can never be sure that a measurement is perfect. Is your bathroom scale completely accurate? Is the odometer in your car a flawless measure of distance? Is your new tape measure 100% correct? When you visit your physician, is it possible that the nurse's measure of your blood pressure is off a bit? Even the use of highly precise scientific instruments is potentially affected by various errors, not the least of which is human error in reading the instruments. All measurements, and therefore all sciences, are affected by various challenges that can reduce measurement accuracy.

Despite the many similarities among all sciences, measurement in the behavioral sciences has special challenges that do not exist or are greatly reduced in the physical sciences (see Figure 1.2). These challenges affect our confidence in our understanding and interpretation of behavioral observations.

One of these challenges is related to the complexity of psychological phenomena; notions such as intelligence, self-esteem, anxiety, depression, and so on may have many different aspects to them. Thus, one key challenge is to identify and capture the important aspects of these types of human psychological attributes in a single number or score.

FIGURE 1.2 ⬢ Difficult Challenges in Psychological Measurement



You may hear people object to the very idea of psychological assessment on the grounds that, for example, "you can't reduce people to a number" or "you just can't quantify creativity." Indeed, no reasonable psychologist would try to use a single number to represent an individual's unique totality. Given the richness of human psychology and the extraordinary variety of ways in which people differ from each other, no single number or set of numbers would fully represent any individual in some general or holistic sense. We cannot reduce someone's "total psychology" to a single number any more than we can reduce their "total physicality" to a single number.

However, it might indeed be possible to quantify something like creativity, or at least specific aspects or dimensions of creativity. Again, no one seriously attempts to quantify an individual's "total physicality"; however, we do quantify specific physical dimensions such as height, weight, and blood pressure. In a similar way, psychologists and others attempt to quantify specific psychological dimensions such as verbal intelligence, self-esteem (or specific forms of self-esteem), achievement motivation, attentional control, and so on. A key challenge is to make sure that the way in which we quantify such specific psychological dimensions does indeed reflect the complexity of those dimensions adequately. If psychologists can identify specific, coherent dimensions along which people differ, then they may be able to quantify those differences quite precisely. Chapters 4 and 12 address this crucial issue of dimensionality.

Participant reactivity is a second difficult challenge. Because, in most cases, psychologists are measuring psychological characteristics of people who are conscious and generally know that they are being measured, the act of measurement can itself influence the psychological state or process being measured. For example, suppose we design a questionnaire to assess racism. People's responses to the questionnaire might be influenced by their desire not to be thought of as a racist rather than by their true attitudes toward particular ethnic or racial groups. Therefore, people's knowledge that they are being observed or assessed can cause them to react in ways that obscure the meaning of their behavior. This is usually not a problem when measuring features of inanimate objects that do not know they are being measured; the weight of a bunch of grapes is not influenced by the act of weighing them, and black holes do not mind when astrophysicists attempt to measure their size.

Participant reactivity can take many forms. In research situations, some participants may try to figure out the researcher's purpose for a study, changing their behavior to accommodate the researcher (**demand characteristics**). In contrast, in both research and applied measurement situations, some people might become apprehensive, others might change their behavior to try to impress the person doing the measurement (**social desirability**), and still others might even change their behavior to convey a poor impression to the person doing the measurement (*malingering*). In each case, the validity and meaning of the measure is compromised—the person's "true" psychological characteristic is obscured by a

temporary motivation or state that is a reaction to the very act of being measured. Chapter 10 discusses this important issue in detail.

Yet another challenge to psychological measurement is that, in the behavioral sciences, the people collecting the behavioral data (observing the behavior, scoring a test, interpreting a verbal response, etc.) can bring their own biases and expectations to their task. Measurement quality is compromised when these factors distort the observations that are made. *Expectation* and *bias* effects can be difficult to detect. In most cases, we can trust that people who collect behavioral data are not consciously cheating; however, even subtle, unintended biases can have effects. For example, a researcher might give intelligence tests to young children as part of a study of a program to improve the cognitive development of the children. The researcher might have a vested interest in certain intelligence test score outcomes, and as a result, they might allow a bias, perhaps even an unconscious one, to influence the testing procedures. **Observer (or scorer) bias** of this type can occur in the physical sciences, but it is less likely to occur because physical scientists rely more heavily than do social scientists on mechanical devices as data collection agents.

The measures used in the behavioral sciences tend to differ from those used by physical scientists in a fourth important respect as well. Psychologists tend to rely on **composite scores** when measuring psychological attributes. Many of the tests used by psychologists involve a series of questions, all of which are intended to measure a specific psychological attribute or process. For example, a personality test might have 10 questions designed to measure extroversion. Similarly, class examinations that are used to measure learning or knowledge generally include many questions.

It is common practice to score each question and then to sum or otherwise combine the items' scores to create a total or composite score. The composite score represents the final measure of the relevant construct—for example, an extroversion score or a "knowledge of algebra" score. Although composite scores do have their benefits (as we will discuss in later chapters, including Chapter 6), several issues complicate their use and evaluation. In contrast, the physical sciences are less likely to rely on composite scores in their measurement procedures (although there are exceptions to this). When measuring a physical feature of the world, such as the length of a piece of lumber, the weight of a molecule, or the speed of a moving object, scientists can usually rely on a single value obtained from a single type of measurement.

A fifth challenge to psychological measurement is **score sensitivity**. **Sensitivity** refers to a measure's ability to discriminate between meaningful amounts of the dimension being measured. For a physical example, consider someone trying to measure the width of a hair with a standard yardstick. Yardstick units are simply too large to be of any use in this situation. Similarly, a psychologist may find that a particular procedure for measuring a psychological attribute or process may not

be sensitive enough to discriminate between the real differences that exist in the attribute or process.

For example, imagine a clinical psychologist who wishes to track her clients' emotional changes from one therapeutic session to another. If she chooses a measure that is not sufficiently sensitive to pick up small differences, then she might miss small but important differences in mood. For example, she might ask her clients to complete this very straightforward "measure" after each session:

Check the box below that best describes your general emotional state over the past week:

☐ Good ☐ Bad

The psychologist might become disheartened by her clients' apparent lack of progress because her clients might rarely, if ever, feel sufficiently happy to checkmark the "Good" box. The key measurement point is that her measure might be masking real improvement by her clients. That is, her clients might be making meaningful improvements—originally feeling extremely anxious and depressed and eventually feeling much less anxious and depressed. However, they might not actually feel good enough to checkmark "good," even though they feel much better than they did at the beginning of therapy. Unfortunately, her scale is too crude or insensitive, in that it allows only two responses and does not distinguish among important levels of "badness" or among levels of "goodness." A more precise and sensitive scale might look like this:

Choose the number that best describes your general emotional state over the past week:

1	2	3	4	5	6	7	8	9
Extremely Good			Somewhat Good		Somewhat Bad		Extremely Bad	

A scale of this kind might allow more fine-grained differentiation along the "good versus bad" dimension as compared with the original scale.

For psychologists, the sensitivity problem is exacerbated because we might not anticipate the magnitude of meaningful differences associated with the mental attributes being measured. Although this problem can emerge in the physical sciences, physical scientists are usually aware of it before they do their research. In

contrast, social scientists may be unaware of the scale sensitivity issue even after they have collected their measurements.

A final challenge to mention at this point is an apparent lack of awareness of important psychometric information. In the behavioral sciences, particularly in the application of behavioral science, psychological measurement is often a social or cultural activity. Whether it provides information from a client to a therapist regarding psychiatric symptoms, from a student to a teacher regarding the student's level of knowledge, or from a job applicant to a potential employer regarding the applicant's personality traits and skill, applied psychological measurement often is used to facilitate the flow of information among people. Unfortunately, such measurement often seems to be conducted with little or no regard for the psychometric quality of the tests.

For example, most classroom instructors give class examinations. Only on very rare occasions do instructors have any information about the psychometric properties of their examinations. In fact, instructors might not even be able to clearly define the reason for giving the examination. Is the instructor trying to measure knowledge (a latent variable or hypothetical construct), determine which students can answer the most questions, or motivate students to learn relevant information? Some classroom tests might have questionable quality as indicators of differences among students in their knowledge of a particular subject. Even so, the tests might serve the very useful purpose of motivating students to acquire the relevant knowledge.

Although a poorly constructed test might serve a meaningful purpose in some community of people (e.g., motivating students to learn important information), psychometrically well-formed information is better than information that is not well formed. Furthermore, if a test or measure is intended to reflect the psychological differences among people, then the test must have strong psychometric properties. Knowledge of these properties should inform the development or selection of a test—all else being equal, test users should use psychometrically sound instruments.

In sum, this survey of challenges should indicate that although measurement in the behavioral sciences and measurement in the physical sciences have much in common, there are important differences. These differences should always inform our understanding of data collected from psychological measures. For example, we should be aware that participant reactivity can affect responses to psychological tests.

At the same time, it is important to emphasize that behavioral scientists have significant understanding of these challenges and that they have generated effective methods of minimizing, detecting, and accounting for various problems. Similarly, behavioral scientists have developed methods that reduce the potential impact of experimenter bias in the measurement process. This book covers many of the

extensive methods that psychometricians have developed to handle the challenges associated with the development, evaluation, and process of measurement of psychological attributes and behavioral characteristics.

THE IMPORTANCE OF INDIVIDUAL DIFFERENCES

The ability to identify and characterize psychological differences is at the heart of all psychological measurement, and it is the foundation of all methods used to evaluate tests. Indeed, the purpose of measurement in psychology is to identify and quantify the psychological differences that exist between people, over time, or across conditions. These psychological differences contribute to differences in test scores and are the basis of all psychometric information. Even when a practicing psychologist, educator, or consultant makes a decision about a single person based on that person's test score, the meaning and quality of the person's score can be understood only in the context of the test's ability to detect differences among people.

All measures in psychology require that we obtain behavioral samples of some kind. Behavioral samples might include scores on a paper-and-pencil test, written or oral responses to questions, or records based on behavioral observations. Useful psychometric information can be obtained only if people differ with respect to the behavior that is sampled. If a behavioral sampling procedure produces scores that differ between people (or that differ across time or condition), then the psychometric properties of those scores can be assessed. This book presents the logic and analytic procedures associated with these psychometric properties.

If we think that a particular test is a measure of a particular psychological attribute, then we must be able to argue that differences in the test scores are related to differences in the relevant underlying psychological attribute. For example, a psychologist might be interested in measuring visual attention. Because visual attention is an unobservable hypothetical construct, the psychologist must create a behavioral sampling procedure or test that reflects individual differences in visual attention. However, before firmly concluding that the procedure is indeed interpretable as a measure of visual attention, the psychologist must accumulate evidence that there is an association between individuals' scores on the test and their "true" levels of visual attention. The process by which the psychologist accumulates this evidence is called the validation process; it will be examined in later chapters.

The following chapters show how individual differences are quantified and how their quantification is the first step in solving many of the challenges to measurement. Individual differences represent the currency of psychometric analysis—they provide the data for psychometric analyses of tests.

BUT PSYCHOMETRICS GOES WELL BEYOND "DIFFERENTIAL" PSYCHOLOGY

Although the previous section highlights the fact that measurement is based on the existence and detection of psychological differences among people, it is important to avoid a common misunderstanding. The misunderstanding is that psychometrics, or even a general concern about psychological measurement, is relevant only to those psychologists who study a certain set of phenomena that are sometimes called "individual difference" variables.

It may be true that psychometrics evolved largely in the context of certain areas of research, such as intelligence testing, that would be considered part of "differential" psychology. Indeed, while many early pioneers in psychology pursued general laws or principles of mental phenomena that apply to all people, Galton, Spearman, and others focused on the variability of human characteristics. For example, Galton was primarily interested in the ways in which people differ from each other—some people are taller than others, some are smarter than others, some are more attractive than others, and some are more aggressive than others. He was interested in understanding the magnitude of those types of differences, the causes of such differences, and the consequences of such differences.

Thus, the approach to psychology that was taken by Galton, Spearman, and others became known as **differential psychology**, the study of individual differences. There is no hard-and-fast definition or classification of what constitutes differential psychology, but it is often seen to include intelligence, aptitude, and personality. This is usually seen as contrasting with experimental psychology, which focused mainly on the average person instead of the differences among people.

Perhaps because Galton is closely associated with both psychometrics and differential psychology, people sometimes view psychometrics as an issue that concerns only those who study "individual differences" topics such as intelligence, ability/aptitude, or personality. Some seem to believe that psychometrics is not a concern for those who take a more experimental approach to human behavior. This belief is incorrect.

Psychometric issues are by no means limited to so-called differential psychology. Rather, all psychologists, whatever their specific area of research or practice, must be concerned with measuring behavior and psychological attributes. Therefore, they should all understand the problems associated with measuring behavior and psychological attributes, and these problems are the subject matter of psychometrics.

Regardless of one's specific interest, all behavioral sciences and all applications of the behavioral sciences depend on the ability to identify and quantify variability in human behavior. The book will revisit this issue later in depth, with specific

examples and principles underscoring the wide relevance of psychometric concepts. Psychometrics is the study of the operations and procedures used to measure variability in behavior and to connect those measurements to psychological phenomena.

Suggested Readings

For a history of early developments in psychological testing:

DuBois, P. H. (1970). *A history of psychological testing*. Allyn & Bacon.

For a history more focused on psychometrics specifically:

Jones, L. V., & Thissen, D. (2007). A history and overview of psychometrics. In C. R. Rao & Sinharay (Eds.), *Handbook of statistics, 26: Psychometrics* (pp. 1–27). North Holland.

For a modern historical and philosophical treatment of the history of measurement in psychology:

Michell, J. (2003). Epistemology of measurement: The relevance of its history for quantification in the social sciences. *Social Science Information*, *42*(4), 515–534. https://doi.org/10.1177/0539018403424004

For an overview of contemporary tests and issues in psychological testing:

Miller, L. A., & Lovler, R. L. (2016). *Foundations of psychological testing: A practical approach* (5th ed.). SAGE.