MEASUREMENT IN THE SOCIAL SCIENCES

Whatever exists at all exists in some amount.

–Edward L. Thorndike

CHAPTER OUTLINE

Suppose that you and your organization, or you and your project team, or just you are interested in the idea of "perception of rejuvenation" and are interested in exploring the relationship between that idea and a variety of health outcomes. Or, perhaps your organization wishes to offer a credential, a license, or certification, for a particular set of skills designed to foster rejuvenation in the workplace. For both scenarios, however, there may be no way to measure perception of rejuvenation because a *test* for it does not exist. And so, you are looking to develop one yourself.

One purpose of this book is to lead you through the process of creating a high-quality test to measure ideas like this one. Furthermore, it is not sufficient to merely develop the test. It must be done in such a way as to convince individuals and organizations who may wish to use it that it is actually achieving its intended purpose. Collecting evidence to support the interpretation and uses of the test is a process called validation and is the second primary purpose of this book.

1.1 WHAT DO WE CALL THEM?

We have a nomenclature issue to deal with before we get started. What do we call an instrument that measures mental variables? Mental variables, which we also call constructs or latent variables, are variables that we cannot directly observe. Most textbooks and sponsoring organizations call the instruments measuring these variables as mental measurement instruments. But what do you think of when you hear the terms test or examination? Many people think of a series of questions on which the respondent tries to do his or her best. There is often an accompanying feeling of stress or pressure. There's even a large body of research on *test anxiety*. But not all mental measurement instruments are like this. Other instruments are called inventories, indices, scales, profiles, or simply measures. These instruments usually attempt to describe typical behaviors, attitudes, or emotions. Here are the names of some widely used instruments:

Graduate Record *Exam*

National *Assessment* of Educational Progress

Minnesota Multiphasic Personality *Inventory* (MMPI)

GED *Tests*

The School Counselor Self-Efficacy *Scale*

The Parenting Stress *Index*

The School as a Caring Community *Profile*

Adolescent Diagnostic *Interview*

Myers–Briggs Type *Indicator*

Measures of Musical Abilities

Rating Scale of Communication in Cognitive Decline

All of these have several things in common. They require the respondent to answer questions, complete tasks, or make a performance. The responses or behaviors are scored to provide an interpretation. Every year, I teach a course called Instrument Development and Validation, and to be sure, I would prefer here to use the term instrument. However, to eliminate confusion over naming and to be consistent with most of the professional literature, I'm going to refer to all of these instruments as *tests*. For example, the *Standards for Educational and Psychological Testing* (APA, AERA, & NCME, 2014) is a primary source of guidance, sponsored by three prominent professional organizations, for test development and validation. The *Standards* refer to tests but apply more broadly to instruments measuring constructs. Additionally, people who take tests are commonly called examinees. But since not everyone is actually taking an exam, I will refer to them generically as respondents, that is, people who respond to the questions on a test.

You may have noticed that the above list of instruments does not include *surveys* and *questionnaires*. Questionnaires are data collection instruments designed to collect a wide variety of variables, some of which could be constructs. They are used in survey programs to collect data from a target population. But questionnaires also are likely to include many other variables, such as demographic, physical measurements, personal history, and test scores. They may also include tests. For example, the National Education Longitudinal Survey (NELS), sponsored by the National Center of Education Statistics, contained questionnaires that asked high school students questions about constructs such as self-concept and locus of control. These sets of questions could be considered to be tests embedded within the questionnaires. To be clear, this book is *not* intended to guide the development and validation of

questionnaires and surveys. This book considers a set of questions that together are designed to provide a score that measures a construct or mental variable.

1.2 WHAT IS MEASUREMENT?

In thinking about measuring constructs, it's critical to discuss first what we mean by measurement. According to *Merriam-Webster's Collegiate Dictionary* (11th Edition, 2014), measurement is "1: the act or process of measuring, or 2: a figure, extent, or amount obtained by measuring" (p. 769). Okay, so what is measuring? One of several definitions of this term is "to allot or apportion in measured amounts (measuring out three cups)" (p. 769). The definitions of measurement in dictionaries tend to have this circular form. However, the operative word here is *amount*. That is, measuring a variable, say someone's height, is tantamount to determining how many height units someone has. The units can be inches, feet, centimeters, meters, or miles. Measuring height is an exercise in counting height units. This is the original, classical conception of measurement which dates back to Euclid's *Elements*, Book V (Heath, 1908). Consider the Thorndike quote at the beginning of this chapter. It is well known among measurement professionals, but the sentence that follows is less well known:

> *Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity.*
>
> (Thorndike, 1918, pp. 16–17)

Unfortunately, this view of measurement does not work well with tests measuring constructs. Implicit in the classical concept of measurement is the idea that counting units results in ratio interpretations. For example, if Person A has 20 units of height and Person B has 10 units of height, not only is there a difference of 10 units but also Person A is *twice* as tall as Person B. This interpretation is clearly difficult, if not impossible, for tests measuring constructs. First, there is no clear unit of measurement. Scores on tests or parts of tests are not equal interval in nature. Consider, for example, one mathematics item "1 + 1 =" and another item "Find the square root of 23." They may both be scored one point for a correct answer, but the difference in mathematics skills between scores of 0 and 1 is not equal to the difference between scores of 1 and 2. Second, a ratio interpretation requires a true zero point. A score of 20 for Person A does not mean twice the amount of the construct as a score of 10 for Person B because a score of 0 does not mean zero amount of the construct. If we

adhered to the classical concept of measurement, then it would be nearly impossible to conduct any quantitative-based research in education or the social sciences.

To get around this problem, the psychologist S. S. Stevens expanded the classical concept of measurement to include other interpretations of measuring constructs in addition to ratio scales. His definition of measurement was:

> *Measurement is the assignment of numerals to objects or events according to rule.* (Stevens, 1946)

For example, in a competition that measures diving skill, a diver performs a certain kind of dive and receives a score of 9.1. There were rules for obtaining that score. For example, if there is a lot of splash entering the pool, 0.2 of a point is deducted. This measurement of diving skill obviously does not fit the classical concept of measurement. There is no absolute 0 point, and it is not even clear that scores are on an equal interval scale. Nevertheless, because there were rules for scoring the dive, rules that were applied equally to all divers, Stevens claims that measurement has taken place. He developed four levels of measurement that are widely taught today in introductory quantitative research methods courses:

- *Nominal/categorical*: Numbers of this type are merely labels for categories, e.g., 1 = female, 2 = male. Many variables, such as marital status, political party affiliation, and geographic region are nominal/categorical in nature. Test scores can also be nominal in nature, e.g., pass or fail.
- *Ordinal*: The numbers indicate relative position or ranking but do not claim to be equal interval, e.g., 1 = first, 2 = second, 3 = third. An example of an ordinal level test score is the percentile rank, or the percentage of respondents with that score or lower.
- *Interval*: The numbers represent equal intervals, e.g., the difference between 5 and 10 is the same as the difference between 10 and 15, but there is no true zero. An example of an interval level variable is the centigrade temperature scale.
- *Ratio*: This level of measurement embodies the original classical concept, where the numbers indicate equal intervals, and there is a true zero. Examples of ratio level variables are length, height, velocity, and reaction time.

So, what is the highest measurement level for scores from tests measuring constructs? This is a point of some debate among psychometricians, but there is a general consensus that test scores fall somewhere between ordinal and interval data. Much depends on how the test is scored. If item scores are simply summed, the resulting scores, called raw or observed scores, are most likely ordinal. As discussed in Chapters 3 and 9, some transformations of raw scores claim to result in an interval scale. However, test scores are clearly not ratio level data, so you can never make an interpretation such as, Person A's self-concept is twice as positive as Person B's even if Person A's score is twice that of Person B's score.

The process of moving psychology from the classical view of measurement to Stevens' concept of levels of measurement being widely adopted actually took place over several decades in the early twentieth century. For the reader interested in this history, I recommend Joel Michell's (1999) account in *Measurement in Psychology: A Critical History of a Methodological Concept*.

1.3 SOME KEY TERMS

I have no wish to bore you with an extensive list of definitions. However, a few terms have different meanings for measurement specialists than they do for the general public or scholars in other disciplines. As a result, it's important to make this distinction so that you know what is meant when I use these terms in this book.

Psychometrics/Psychometricians. Psychometrics is the academic discipline associated with mental measurement. It includes scaling and scoring tests, validation, and test theories. It may also be referred to as quantitative psychology. Psychometricians are professionals who work and conduct research in this discipline. In other words, these are the testing experts. On the other hand, in the field of school psychology, psychometricians are professionals who administer tests. That is different from the use in this book.

Test. A test (or exam or scale or one of the other labels mentioned above) is a device for obtaining a measurement. More specifically, it is a "standard procedure for obtaining a sample of behavior from a specified domain" (Crocker & Algina, 1986, p. 4). The

operative words in this definition are *standard* and *sample*. Although we are dealing here with test development, this definition also applies to everyday use. For instance, a thermometer is an instrument for measuring one's body temperature. The measurement is done the same way every time, or by following a set of procedures. Furthermore, each measurement is a sample, e.g., temperature taken in the morning, from a large number of possible occurrences.

Assessment. According to the National Council on Measurement's (NCME) glossary (http://www.ncme.org/resources/glossary), assessment means:

> *Any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects, or programs; a process designed to systematically measure or evaluate the characteristics or performance of individuals, programs, or other entities, for purposes of drawing inferences; sometimes used synonymously with test.*

This rather wordy definition conflates an assessment with a test. In fact, many psychometricians think of assessment as testing. State educational achievement testing programs are often called state assessments, while at the national level, the federal government sponsors the National Assessment of Educational Progress (NAEP). On the other hand, outside of a measurement context (Merriam-Webster, 2014), to assess has a broader meaning: "to determine the importance, value, or size of…." (p.74). This definition of assessment does not need to have anything to do with mental measurement.

Evaluation. Related to assessment (Merriam-Webster, 2014), to evaluate is defined as: "to determine the value of" (p.432). For example, an assessment such as NAEP may discover that 30 percent of third graders are proficient in reading, but reporting that this percentage is below national goals is an evaluation of that assessment.

For other terms related to testing, test development, and validation that are not explicitly defined in this book, I refer the reader to the glossary on the NCME website.

1.4 MEASURING CONSTRUCTS

At the beginning of this chapter, I mentioned that this book is intended to guide the development of tests to measure mental variables most commonly called constructs or latent variables. Much of psychological or educational test development is geared toward measuring how much of some hypothetical variable respondents have. We construct these variables to explain human behavior on the basis of a theory of that construct, but we are never sure if these constructs really exist. These variables are sometimes called traits, but there is a difference in meaning. Traits are thought of as relatively stable characteristics, for example, the Big Five personality traits. So, traits are a type of construct, but the term construct is broader in meaning. For the remainder of this book, for simplicity, I will use the term construct.

1.5 HISTORICAL CONTEXT

Current practice in test development and validation has evolved over the course of time, over 4,000 years as it turns out. It's useful to examine briefly this history in order to see the basic underlying principles on which development procedures rest and why they are important. In particular, there are five major lines of development that converge in terms of methodology even though the purposes of these types of tests vary widely: personnel selection, credentialing, measuring intelligence, educational achievement, and personality.

1.5.1 Personnel Selection Testing in China, the United Kingdom, and United States

The first tests, as we know them today, were used to select individuals for employment. And the first of these dates back over 4,000 years to China. Nation states at that time were led by ruling families, or dynasties. China was no different in that regard. However, unlike other nations, China prioritized hiring the most qualified individuals for government positions. Beginning in 2200 BCE, China administered "examinations" every three years to government officials for determining their fitness for remaining on the job. In 1115 BCE, the Chan dynasty began formal testing to select individuals for high government positions. These positions were highly prized, and millions of Chinese sat for these exams from which a chosen few were hired. The

exams were performance-based and covered music, writing, arithmetic, horsemanship, and prominent ceremonial rites.

In developing these tests, the Chinese government developed methodology that is still used today. First, standardized procedures were implemented so that everyone was measured the same way. For example, scorers did not know who the candidates were. Two examiners scored each response, with a third brought in to mediate differences between the scorers. Additionally, candidates were tested under environmental conditions that were as similar as possible. Finally, the exams themselves were small samples of behavior under controlled conditions that could fairly well predict how individuals would perform in broader conditions.

The Han dynasty (202 BCE–220 AD) undertook major revisions of the civil service examinations. The program became a set of separate exams for law, revenue, military, agriculture, geography, and moral standards. After the invention of paper in 105 AD, these became the first written exams. From this point forward, the Chinese civil service examinations evolved in content and purpose. One key purpose was the promotion of meritocracy, or the hiring of the best qualified candidates as a path to upward mobility. This idea later became a primary driver for the widespread use of college admissions tests. For the reader interested in the Chinese civil service examination system and its intended and unintended consequences, I recommend Benjamin A. Elman's (2013) *Civil Examinations and Meritocracy in Late Imperial China*.

The Chinese examinations became a model for other nations. In the mid-1800s, British visitors to China were so impressed with the examination system that they modeled a testing system after it for selecting candidates for the civil service in India and later in Great Britain. In the late 1800s, legislators in the United States modeled the American civil service examination program after the British system.

1.5.2 Credentialing

Many professions today have a credentialing process in place, for certification or licensure. This process usually involves candidates needing to pass an examination or a series of exams. In particular, licensure procedures for the medical and legal professions have been in existence for several thousand years. In fact, the early Chinese civil service examinations included credentialing for those two professions. In the

early part of this millennium, legal regulations of law and medicine became established in Europe. By the 14th century, credentialing took two primary forms. First, there were educational or training requirements. Second, candidates had to pass an examination.

The methodology for the examinations was taken largely from the Chinese civil service procedures. They tended to be performance-based, for example, requiring surgeon candidates to operate on an animal. A major methodological problem was inconsistency in scoring, and so in the 20th century, multiple-choice items replaced or supplemented essay and performance-based questions in many credentialing programs. In recent years, there has been a trend back to performance-based items as multiple-choice exams have been criticized for being unrealistic and unable to measure accurately some important professional skills. In addition, professional organizations, such as the National Board of Medical Examiners for medicine, arose to oversee the credentialing process.

1.5.3 Measuring Intelligence

In the late 1800s, psychology developed into a new discipline, distinct from philosophy. Early on, psychologists tried to apply the idea of measuring physical variables, such as height and velocity, to the measurement of constructs. The first attempts focused on intelligence. I will outline this history in some detail because it illustrates some of the controversies and abuses of testing. For an excellent history of intelligence testing, I refer the reader to Stephen J. Gould's *The Mismeasure of Man* (1996).

One of the first theories about intelligence was based on craniometry: the idea that smarter people had larger brains. An early proponent of this theory was Samuel G. Morton, a physician from Philadelphia. To be honest, Morton set out to show that racial groups differed in intelligence and that these differences were inherited. He hoped to support this view by exhuming bodies from different parts of the world and measuring their cranial capacities. Unfortunately, as Gould (1996) shows, Morton manipulated his data in order to come to the conclusion he desired.

Alfred Binet was a French psychologist who directed the Laboratory of Physiological Psychology at the Sorbonne at the turn of the 20th century. His research with his student Theodore Simon focused on child development. In 1904, the French

government appointed a commission to study broadly the education of at-risk children and specifically to develop a test to identify these children. As members of this commission, Binet and Simon worked to develop the first intelligence test. At the time, craniometry was still the most prominent theory of intelligence. However, Binet knew that this approach would not work for their test (craniometry would recommend simply lining up the children and picking the smallest). Binet had earlier conducted some research on cranial capacity. He found that there were extremely small differences in head size between children identified as the brightest and slowest by teachers. The brightest children's heads were slightly larger, but this difference could be attributed to better nutrition and health. As a result, he gave up on physical measurement and decided to measure reasoning more directly. In 1905, he and Simon created a short series of tasks, related to everyday life but not explicitly taught in school (i.e., no mathematics or reading). These tasks included things such as repeating sentences and sequences of numbers and stating the difference between pairs of things. The score on the test was a child's mental age, or the average chronological age of children who performed the same tasks correctly as the respondent. Binet thought of his test as measuring different types of intelligence with mental age as an average of these intelligences. The test succeeded in its mission, identifying children with special educational needs.

The success of the Binet–Simon test attracted the attention of Henry H. Goddard, the Director of the Vineland (NJ) Training School for Feeble-minded Girls and Boys. In 1913, he translated the Binet–Simon test into English and used it at the Vineland School to identify and rank students. However, unlike Binet, Goddard subscribed to the idea that intelligence is inherited and fixed. It is interesting to note that he coined the technical terms moron, imbecile, and idiot to classify increasingly lower levels of intelligence. Furthermore, he thought the United States would be a better nation if it could reduce the number of low-intelligence people. Goddard promoted two ways to accomplish this. First, he proposed to sterilize low-intelligence people and was one of the founders of the eugenics movement. Second, the test could be used to prevent low-intelligence people from immigrating. To this end, Goddard established an intelligence testing program at Ellis Island.

At the same time, Lewis Terman, a psychologist at Stanford University, made several major revisions to the last (1911) version of the Binet–Simon test. He added

adult-level items and developed an alternate scoring method, the intelligence quotient, or IQ. This revised version became the Stanford–Binet Intelligence Test, which is still used today by psychologists. He and his colleagues used the test in his famous longitudinal study of gifted individuals and their offspring. The Stanford–Binet became widely used and ushered in the field of testing as a business. On the other hand, Terman, like Goddard, believed that intelligence was a fixed, inherited trait and promoted the test as a measure of personal worth. He became an active participant in the eugenics movement.

In 1917, as the United States entered World War I, the army faced a serious problem. Who among the new recruits could serve as officers? In other words, the army wanted to find the most intelligent recruits. The problem was that, because the Stanford–Binet was individually administered and took over an hour to complete, it was unfeasible to test nearly two million recruits quickly. The US Army hired a team of psychologists, led by Robert Yerkes and including Goddard and Terman, to develop large-scale group-administered intelligence tests. They developed two tests, the Alpha, administered to literate recruits, and the Beta, administered to recruits who could not read or speak English. The exams used methods that are common in practice today, including objectively scored item formats such as multiple-choice and analogies. The exams were administered to nearly 1.75 million recruits in a short period of time. The US Army was satisfied with the results, in terms of classifying the recruits for various positions.

In addition to supporting the war effort, the Army Alpha and Beta had enormous consequences for the testing industry, government policy, and the discipline of psychometrics. First of all, the development team was still steeped in the concept of intelligence as a fixed, inherited trait. As a result, the team conducted numerous statistical analyses and used these to advance the policy agenda of Goddard, Terman, and others. For example, Table 1.1 shows a set of results reported by Yerkes (1921) in which Alpha scores were analyzed in terms of years of residence in the United States. Many, if not most, US citizens were immigrants from other nations. The table appears to show a relationship between how long a recruit has lived in the United States and intelligence, with a longer time associated with greater intelligence. Further analysis showed that longer-term recruits tended to come from northern and western Europe while more recent recruits came from southern and eastern Europe. Yerkes

TABLE 1.1 ⬢ Yerkes's (1921) Table for Average Mental Age by Years of Residence in United States

Years of Residence	Mean Mental Age
0–5	11.29
6–10	11.70
11–15	12.53
16–20	13.50
20–	13.74

concluded that northern and western Europeans were more intelligent than southern and eastern Europeans. Analyses such as these led to the Immigration Restriction Act of 1924, which set immigration quotas based on nation of origin. And fourteen years later, before World War II, when Jews wanted to escape to the United States from Germany and other parts of Europe, the United States used the 1924 Act to justify turning them away. Not our finest hour.

Of course, today we come to a far different interpretation of the Yerkes results. Many of the Alpha items required some familiarity with American culture and geography. Plus, the test was in English. Recruits who had been here longer simply had more time to learn English literacy skills and more time to be familiar with the American environment.

Similarly, much intelligence research was originally intended to discover differences in intelligence among race/ethnic groups, the results of which were used to promote discrimination. An examination of peer-reviewed journal articles from the 1930s into the 1960s shows that many researchers considered IQ differences between subgroups a legitimate question for inquiry. Things changed in 1969 with the publication of Arthur Jensen's (Jensen, 1969) article in which he promoted research that he believed suggested racial differences in intelligence. He was widely criticized (and threatened) for his article. Since then, researchers have largely viewed subgroup differences as resulting from environmental factors. However, there is still a small group of scholars, most notably Richard J. Hernstein and Charles Murray (1994) in *The Bell Curve*, who, in the face of severe public pressure and sanctions, persist in promoting genetic differences in intelligence.

Secondly, the Army tests also spurred the development of other large-scale tests, many of which are still in use. Among these are:

- Otis Group Intelligence Scale (1918): Arthur Otis, a student of Terman, created a group intelligence test modeled on the Alpha. This was the first commercial group-administered psychological test.
- ACE's Psychological Examination (1924): Sponsored by the American Council on Education, this was a test developed by L. L. Thurstone (see Chapter 11 on dimensionality), for college admissions. Patterned on the Alpha and Beta, respectively, it had two scores, linguistic and quantitative. As the SAT became more widely used, the Psychological Examination gradually faded.
- Scholastic Aptitude Test (SAT) (1926): The SAT began as Harvard College's test for awarding scholarships. Its two parts were patterned after the Alpha and Beta, but it was designed for entering college students. It became Harvard's admissions test and before long was adopted by all the Ivy League colleges, and then more widely as a nationwide college admissions test. This necessitated the formation of the College Entrance Examination Board, now called simply The College Board, to oversee its development and administration.
- Weschler Intelligence Scales (1939): David Weschler worked on the Army tests. The Weschler scales report two IQs, a Verbal IQ based on the Alpha and a Performance IQ based on the Beta. The Weschler scales are a competitor to the Stanford–Binet and are widely used in psychological assessments.

Finally, the Army tests led to the formation of large-scale testing as an academic and professional discipline as well as a major industry, as distinguished by the following:

- Professional organizations: Psychometric Society (1935), National Council on Measurement in Education (NCME) (1938), and the American Psychological Association (APA) Division 5 (1945).
- Peer-reviewed journals: *Psychometrika* (1936) and *Educational and Psychological Measurement* (1941).

- Textbooks by Kelley (1927) and Guilford (1936).
- Test publishers: Psychological Corporation, California Test Bureau, and the Iowa Testing Programs.
- Guidance in the development and use of tests:
  - *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).
  - *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 1988).
  - *Educational Measurement* (4th Ed.) (NCME, ACE, 2006).
  - *Handbook of Statistics Volume 26: Psychometrics* (Rao & Sinharay, 2007).
  - The International Test Commission has published numerous guidelines on its website: for test use, computer-based testing, test security and disposal, translating and adapting tests, assessment of diverse populations, a test-taker's guide, and using tests for research (https://www.intestcom.org/page/28).

1.5.4 Educational Achievement

Before 1845, educational achievement as an indicator of school quality was measured by exhibitions—public, well-rehearsed demonstrations of skills, such as recitations of facts and figures and speeches about topics like bravery and loyalty. These were done primarily to secure funding and recruit students for schools and to earn awards and recognition for students. Actual examinations were sometimes administered, but not every student was tested, often only the best students. In addition, what examinations that were given were often of dubious quality. Here is one sample item (reported by Reese, 2013, p. 89):

> *If 11 young men can become fools by drinking 6 bottles of wine, at $3 a bottle, what would it cost a dinner party of 25, to become fools in the like manner?*

As a result, there was no way to compare schools directly. In 1845, legislators from Boston and New York began a lively debate on which city had the better schools. To

be able to compare schools and school systems, a common achievement test was needed. And so, a movement was begun in Boston, led by Horace Mann, the Secretary of the new state Board of Education, to create a central examination administered to all students. Several years before, Mann had visited the United Kingdom and Europe and was impressed by their standardized civil service exams. Mann's ulterior motive for achievement testing was that he thought US schools were falling behind those from other nations. He viewed a central examination program as an agent of wholesale reform. To that end, he led a committee to develop the first standardized achievement tests in the United States. These were administered first to students in Boston and consisted of sections on reading, mathematics, writing, and geography.

Mann's vision for achievement tests largely came true in the form of school reform: superintendents were hired for centralized school districts, classrooms were graded, and female teachers were hired. And yet, controversies arose, debates about which are still relevant today. For one thing, teacher quality was evaluated using test scores. And in a statement easily found in today's media:

> *Is it not a wonder that so many of our American boys and girls survive the almost continual examinations to which they are subjected? America is gripped by examination mania, turning pupils into walking encyclopedias and threatening to send teachers to an early grave.*
>
> (Charles Parker, high school principal, 1878, in Reese, 2013, p. 1)

However, achievement testing caught on. By the 1870s, most city school districts used written examinations to evaluate student achievement and school and teacher quality. By 1990, achievement tests were administered in nearly all public school districts. In the 1920s, on the business side, achievement testing intersected with the Army Alpha exam. As mentioned earlier, the methodology developed for the Alpha led to the creation of corporations devoted to large-scale testing. In the 1920s, corporations and organizations were developed to administer and score achievement tests, including the Stanford Achievement Test (1923) and those from the Iowa Testing Programs (1925). What these larger entities made possible were achievement test batteries that enabled comparisons across states and with the nation as a whole.

1.5.5 Personality

Unlike tests for personnel selection, credentialing, intelligence, and educational achievement, in which respondents are instructed to do their best, group-administered tests intended to measure personality constructs instead ask respondents to report their *typical* behavior. As a result, one criticism often made against these tests is that it is uncertain how honest respondents really are.

Ironically, the first personality tests were actually selection tests designed to identify respondents with personality or behavior disorders. The stakes were then quite high. Like the Army Alpha, the first formal personality test was developed during World War I where American military leaders were concerned about the emotional stability of soldiers, many of whom experienced long-term traumatic symptoms, such as nausea, night shakes, and heart palpitations. These symptoms, called "shell shock," are now called post-traumatic stress disorder (PTSD). The US Army commissioned Robert S. Woodworth (1919) to develop a test that would identify soldiers who were emotionally unstable and unfit for further combat service. The test was called the scale of Psychoneurotic Tendencies (PT). The Army used the PT to screen soldiers for further psychiatric intervention. After the war, Woodworth adapted the test for industrial research and renamed it the Woodworth Personal Data Sheet (WPDS).

Many personality tests followed the WPDS, through the 1940s. Nearly all of them focused on maladaptive aspects of personality, such as "emotional instability" and "lack of emotional control." The prevailing view was that emotional problems would cause problems at the workplace, and so, these tests were designed to assist employers in identifying individuals with such problems.

The focus on maladaptive behavior began to change in the 1940s. Katherine Briggs and Isabel Briggs Meyers developed the Myers–Briggs Type Indicator (MBTI). Based on the theory of personality types of Carl Jung, the MBTI has become today the most popular personality test. Its primary appeal is that the sixteen personality types of individuals are presented as having strengths and weaknesses which can explain much human behavior. No stigma is placed on any of the types. As a result, respondents can learn their type, be proud of it, and not worry about being labeled as maladaptive. The MBTI ushered in a new type of personality test for respondents seeking greater self-knowledge and self-awareness.

Beginning in the 1940s, personality tests began to include inventories designed to assist in clinical diagnoses in psychiatric and medical settings. These included the Minnesota Multiphasic Personality Inventory (MMPI) and the Sixteen Personality Factor (16PF) Questionnaire. These tests did attempt to identify maladjustment but in a different way. Their purpose was to identify serious psychopathology.

More recently, personality tests have been used in research settings in an attempt to explain human behavior more broadly. Applications of the MBTI have expanded greatly to include work settings, personal life, relationships, and goal setting. Much research has suggested that many personality variables are changeable in nature, and so respondents' scores on such tests can change over time, challenging their use for high-stakes decisions. The exceptions appear to be the "Big Five" personality traits (Openness, Conscientiousness, Extraversion, Neuroticism, and Agreeableness). Much work has been done in recent decades to explore the explanatory power of other personality constructs, such as self-efficacy, self-concept, locus of control, grit, and motivation. Tests to measure these constructs are often developed by individuals or small teams of researchers.

1.5.6 Current State of Test Development

Test development has exploded in the past two decades for personnel selection, educational achievement, credentialing, and the measurement of cognitive and noncognitive constructs. Some of the work has been undertaken by large testing organizations such as Educational Testing Service, the National Board of Medical Examiners, and programs for statewide educational achievement testing. In addition, however, much work is undertaken by universities and nonprofit organizations, sometimes by small teams or individual faculty and students. These projects often lack the resources of large-scale organizations, but it is still possible for them to develop high-quality tests.

In recent years, four major trends have altered the test development process in fundamental ways.

Norm- Versus Criterion-Referenced Score Interpretations. Until the 1980s, most test scores provided *norm-referenced* information, that is, how respondents compared to a target population. In educational achievement, for example, much emphasis was placed on how far above- or below-average students, schools, and school districts

were. More recently, greater emphasis has been placed on how well individuals and groups score compared to performance standards, such as "proficient," resulting in a *criterion-referenced* score interpretation. This distinction is discussed in greater detail in Chapters 3, 4, and 6. Chapter 12 discusses methods for establishing performance standards.

Test Theory. Test theories are conceptions of how to scale and score tests and how to estimate how much measurement error they contain. Until the 1970s, the dominant test theory, now called Classical Test Theory (CTT), was used as the basis for test development. CTT has several drawbacks that were widely known, but until new test theories and corresponding computer applications to implement them were developed, CTT was the best feasible option. One of the modern test theories, Rasch measurement theory, is introduced and implemented in Chapter 9. Many current testing programs use Rasch theory or its relative, Item Response Theory (IRT), to scale and score tests. One of the potential advantages of the modern theories is that test scores are believed to be closer to an interval level measurement than scores developed using CTT.

Technology. One of the critical questions to be addressed in any test development project is how test data are to be collected. For many decades, the best option was a printed test booklet with answers recorded on answer sheets, that is, a paper-pencil format. Initially, the data were transcribed by hand into a storage medium. Later, optically scanned answer sheets greatly speeded up the process. In the past two decades, tests have been increasingly administered by computer. There are several potential benefits to computer-administered tests, including greater data security, instant data entry, and instant score results to the respondent. In the past ten years, advances in computer technology have increased capacity to the point where new item formats are possible, including various types of "technology-enhanced" items that allow the measurement of skills not possible with traditional item formats. Some of these advances are discussed in Chapter 6.

By the same token, advances in artificial intelligence (AI) have supported the automation of many aspects of test development and scoring (NCME, 2020). For example, scoring engines have been *trained* to identify and score salient features of essays, to the point that these engines can operationally score essays with the same degree of accuracy as human raters. Another area enhanced by AI is automated item

generation. This technology is aimed at writing items from a template model and can support the large number of items needed for some forms of computer-based test administration.

Distancing From IQ. For many decades, IQ tests and their cousins were used widely to make decisions such as personnel selection and placement and college admissions. IQ is known to correlate positively with almost any cognitive construct. More recent test development frameworks have focused on the specific knowledges, skills, and processes that underlie more global constructs, such as intelligence and achievement. These are discussed in Chapter 5.

1.6 SOME DESCRIPTIVE STATISTICS (CENTRAL TENDENCY, VARIATION, CORRELATION)

Many of you reading this book have had coursework in introductory statistics and/or quantitative research methods in which you studied some of the basic descriptive and inferential statistics. For you, this section will constitute a review. If you are completely new to statistics, then this section serves as an introduction to the statistics that feature prominently in test development and validation.

Statistics are numbers that summarize or characterize some property of the data. For example, statistics that measure central tendency attempt to find the value of a variable that best represents the dataset, while statistics that measure dispersion look for a number that indicates how spread out cases are on variables of interest. Additionally, formulas for statistics that are written in English indicate sample statistics, or values calculated on a sample of respondents from a target population. Formulas written using Greek letters are for *parameters* of the target population, calculated from all cases, that is a *census*, from an entire population. Such datasets are extremely rare. Greek versions of formulas are also frequently used to express theoretical concepts.

Summation Operator. Many statistical formulas require summing various quantities. To express this, an upper-case sigma is used. Consider the following scores for six respondents on a test measuring self-concept:

i	Self-concept (x)
1	7
2	5
3	3
4	8
5	9
6	10

The sum of these six scores is:

$$\sum x_i = 7 + 5 + 3 + 8 + 9 + 10 = 42$$

If you wanted to add only the second and third scores, you can limit the summation operator as follows:

$$\sum_{i=2}^{3} x_i = 5 + 3 = 8$$

For some formulas, the order of operations is important. For example, summing a set of squares is not equal to the squared sum:

$$\sum x^2 = 7^2 + 5^2 + 3^2 + 8^2 + 9^2 + 10^2 = 328$$
$$\left(\sum x\right)^2 = 42^2 = 1,764$$

Central Tendency. By far, the most popular statistic for measuring central tendency is the mean, or average. The sample and population means are shown below. The formulas in this case are identical. The calculation of the sample mean for the above six self-concept scores is shown. N is the population size, and n is the sample size.

Sample	Population
$\bar{X} = \frac{\sum x_i}{n}$ $\bar{X} = \frac{7 + 5 + 3 + 8 + 9 + 10}{6} = \frac{42}{6} = 7$	$\mu_x = \frac{\sum x_i}{N}$

Another popular statistic for central tendency is the median. The median is defined as the midpoint of the scores, half below, half above. Sample and population versions of the median are also identical. If there are an odd number of scores, the median is simply the middle score. If there are an even number of scores, the median is the mean of the middle two scores. The self-concept scores are first arranged in ascending order: 3, 5, 7, 8, 9, 10. Since there is an even number of them, we locate the middle two scores, 7 and 8. The median is then 7.5. The median is preferred to the mean when scores are highly skewed. For example, the median salaries of professional athletes and home prices are often reported instead of the mean because of a small number of extremely high salaries and home prices. If Bill Gates were to enroll in a college course, then the average student is a billionaire!

Dispersion. The most frequently used statistics to characterize dispersion are the variance and standard deviation. The variance can be described as the mean of squared deviations from the mean. The standard deviation is simply the square root of the variance. Its major advantage is that it is expressed in the original score units. Sample and population formulas differ, as shown below. Additionally, if a variable is dichotomous, that is, it can take only two values, such as a test item scored correct or incorrect, the formulas for variance and standard deviation are simplified.

	Sample	Population
Variance	$s^2 = \frac{\sum(x_i - \bar{X})^2}{n-1}$ $s^2 = \frac{(7-7)^2 + (5-7)^2 + (3-7)^2 + (8-7)^2 + (9-7)^2 + (10-7)^2}{6-1}$ $= \frac{34}{5} = 6.8$ For a dichotomously scored variable: $s^2 = p(1-p)$, where p = proportion of one of the possible scores	$\sigma^2 = \frac{\sum(x_i - \mu_x)^2}{N}$
Standard Deviation	$s = \sqrt{\frac{\sum(x_i - \bar{X})^2}{n-1}}$ $s = \sqrt{6.8} = 2.61$ For a dichotomously scored variable: $s = \sqrt{p(1-p)}$,	$\sigma = \sqrt{\frac{\sum(x_i - \mu_x)^2}{N}}$

Relationship Between Variables. Relationships between two quantitative variables (i.e., variables measured at the interval or ratio levels) are usually described by two

related statistics, covariance and correlation coefficient.[1] Both are based on the cross-products between two variables. A cross-product for a single respondent multiplies the deviation from the mean on one variable with the deviation from the mean of the other variable. If these deviations are both positive (i.e., both scores are above the mean) and negative (i.e., both scores are below the mean), their cross-products will be positive, indicating a positive relationship. If one score is above the mean while the other is below the mean, the cross-product will be negative. If most cross-products are negative, a negative relationship results. The covariance is conceptually the mean of the cross-products. The correlation is the covariance divided by the standard deviations of the two variables. Formulas for the covariance and correlation are shown below. As an example, the sample covariance and correlation are computed on the following data:

i	Self-concept (*x*)	Statistics Anxiety (*y*)
1	7	10
2	5	9
3	3	13
4	8	9
5	9	11
6	10	14

In this case, statistics anxiety has a sample mean of 11.0 and standard deviation of 2.10. The covariance can range from very large negative values to very large positive values. Its main deficiency as a descriptive statistic is that it is difficult to judge how large a covariance is indicative of a strong relationship. The correlation coefficient solves this problem by dividing the covariance by the product of the two standard deviations. This limits correlations to values between −1.00 and +1.00. In this example, the covariance coincidentally is 1.0, but the correlation is 0.11, indicating a positive but weak relationship between self-concept and statistics anxiety.

[1]The correlation coefficient presented here is officially the Pearson product-moment correlation coefficient.

<table>
<tr><th></th><th>Sample</th><th>Population</th></tr>
<tr><td>Covariance</td><td>$S_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$
$S_{xy} = \frac{(7-7)(10-10)+(5-7)(8-10)+(3-7)(13-10)+(8-7)(9-10)+(9-7)(11-10)+(10-7)(14-10)}{6-1}$
$= \frac{5}{5} = 1.0$</td><td>$\sigma_{XY} = \frac{\sum(X-\mu_X)(Y-\mu_Y)}{N}$</td></tr>
<tr><td>Correlation Coefficient</td><td>$r_{xy} = \frac{s_{xy}}{s_x s_y}$
$r_{xy} = \frac{1.0}{(2.61)(2.10)} = 0.11$</td><td>$\rho_{XT} = \frac{\sigma_{XT}}{\sigma_X \sigma_T}$</td></tr>
<tr><td>Point biserial correlation</td><td>If x is the dichotomous variable, then:
$r_{pbis} = \frac{\bar{y}_1 - \bar{y}_0}{s_y}\sqrt{\frac{n_1 n_0}{n(n-1)}}$</td><td></td></tr>
<tr><td>Phi coefficient</td><td>Let the sample sizes of each value of each variable are noted as:
<table>
<tr><td></td><td>$y = 1$</td><td>$y = 0$</td><td>Total</td></tr>
<tr><td>$x = 1$</td><td>n_{11}</td><td>n_{10}</td><td>$n_{1.}$</td></tr>
<tr><td>$x = 0$</td><td>n_{01}</td><td>n_{00}</td><td>$n_{0.}$</td></tr>
<tr><td>Total</td><td>$n_{.1}$</td><td>$n_{.0}$</td><td>n</td></tr>
</table>
Then,
$\varnothing = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1.}n_{0.}n_{.0}n_{.1}}}$</td><td></td></tr>
</table>

As you will see later in this book, there are several versions of the correlation coefficient that simplify the formula for its calculation. If one of the variables being correlated can take only two possible values, often 0 and 1, then the correlation simplifies to the *point biserial correlation coefficient*, as shown above. If both variables can take only two values, then the correlation reduces to the *phi* coefficient. The important thing to note about the phi coefficient is that, even though it is still a Pearson correlation, its value depends entirely on sample sizes, or proportions, of each variable that are 1's and 0's, and not on means, standard deviations, or covariances.

1.7 SAMPLE DATASETS: TIMSS AND CIRP

To do the data analyses suggested in this book requires specialized software (actually, much of it could be done with standard applications such as EXCEL, but it would be quite laborious). In keeping with the assumption that you are either working alone or in a small group and that your resources are limited, I will use only free and open-sourced software. One application is jMetrik (https://itemanalysis.com/), which carries out a wide variety of psychometric analyses necessary for test development. I will also use several R packages (https://www.r-project.org/): psych, mirt, and lavaan. While jMetrik is easy to use through its graphical user interface, R packages can be more challenging. As a result, I have placed the code for all the analyses in Appendix B.

I have selected two datasets to illustrate the various analyses. The first dataset comes from the grade 8 mathematics assessment from the *Trends in International Mathematics and Science Study (TIMSS) Assessment 2003 for Grade 8 Mathematics*. This dataset consists of item response data for 20 multiple-choice items from one booklet of the US sample. The second dataset comes from a series of 18 items measuring political viewpoint from the *1999 CIRP Freshman Survey* (Higher Education Research Institute, 1999). The questionnaire and data come from a random sample of students preparing to attend a large public university. These datasets were selected to illustrate the development of a cognitive construct (mathematics skill) and a noncognitive construct (political viewpoint) and data from items scored dichotomously (correct versus incorrect) and polytomously (varying degrees of agreement). These datasets can be found on the accompanying website.

1.8 CHAPTER SUMMARY

This chapter introduced several terms used throughout the book, the most important being measurement. The measuring devices covered in the book go by a variety of names, including tests, examinations, inventories, and scales, but I will refer to them more broadly as tests. I outlined a brief history of test development in five areas: personnel selection, credentialing, intelligence, educational achievement, and personality. Some of the key principles in test development, such as standardization, go back thousands of years. Despite the vast differences in purpose and design, all of these types of tests use much the same methodology in their development. Finally, although it is likely that many readers have had some familiarity with statistics, I introduced some of the descriptive statistics that are widely used in the context of test development.

1.9 EXERCISES AND ACTIVITIES

Test Development Project

In the chapters to follow, this book will guide you in the development of a new test. For now, begin to think about a construct you would like to measure and how a test measuring that construct would be used. Write a concise definition of this construct.

Questions for Discussion

1. Why is standardization a key characteristic of test development?
2. What is the key difference between a test and a survey questionnaire?
3. How does Stevens's definition of measurement differ from the classical concept of measurement?

FURTHER READING

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Elman, B. A. (2013). *Civil examinations and meritocracy in late imperial China*. Cambridge, MA: Harvard University Press.

Emre, M. (2018). *The personality brokers: The strange history of Myers-Briggs and the birth of personality testing*. New York, NY: Doubleday.

Gibby, R. E., & Zichar, M. J. (2008). A history of the early days of personality testing in American industry: An obsession with adjustment. *History of Psychology*, 11(3), 164–184.

Gierl, M. J., & Lai, H. (2013). Using automated processes to generate test items. *Educational Measurement: Issues and Practice*, 32(3), 36–50. NCME ITEM http://ncme.org/publications/items/

Gould, S. J. (1996). *The mismeasure of man: Revised and expanded*. New York, NY: W. W. Norton.

Joint Committee on Testing Practices (1988). *Code of fair testing practices in education*. Washington, DC: Author.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge, UK: Cambridge University Press.

National Council on Measurement in Education. (2006). In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: Praeger.

National Council on Measurement in Education (2020). *Digital module 18: Automated scoring*. Mt. Royal, NJ: Author.

Reese, W. J. (2013). *Testing wars in the public schools: A forgotten history*. Cambridge, MA: Harvard University Press.

REFERENCES

Crocker, L., & Algina, J. (1986). *An introduction to classical and modern test theory*. Belmont, CA: Wadsworth Publishing.

Guilford, J. P. (1936). *Psychometric methods*. New York, NY: McGraw-Hill.

Heath, T. L. (1908). *The thirteen books of Euclid's Elements* (Vol. 2). Cambridge, UK: Cambridge University Press.

Hernstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York, NY: The Free Press.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 39, 1–123.

Kelley, T. L. (1927). *Interpretation of educational measurements*. Yonkers-on-Hudson, NY: World Book Company.

Merriam-Webster. (2014). Measurement. In *Merriam-Webster's Collegiate Dictionary* (11th ed., p. 769).

Rao, C. R., & Sinharay, S. (Eds.) (2007). *Handbook of statistics, volume 26: Psychometrics*. Amsterdam: Elsevier North-Holland.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677–680.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurement of educational products. In G. M. Whipple (Ed.), *Seventeenth Yearbook of the National Society for the Study of Education* (Vol. 2, pp. 16–24). Bloomington, IL: Public School Publishing.

Woodworth, R. S. (1919). Examination of emotional fitness for warfare. *Psychological Bulletin*, 16, 59–60.

Yerkes, R. M. (Ed.). (1921). *Psychological examining in the United States army: Memoirs of the National Academy of Science* (Vol. 15). Washington, DC: U.S. Government Printing Office.