1

Introduction

When I told my relatives and friends I was writing a book about measurement, the most common reaction was confusion. How hard can measurement be, they asked, that you need to write an entire book about it? Just take a ruler or thermometer and measure whatever it is you want to measure. For many physical measurements, it is indeed this simple, although the measurement devices we now take for granted, such as rulers and thermometers, took some time to develop and become accepted. In the social sciences, however, measurement is not so simple. This is because most of the things social scientists want to measure are not physical but mental attributes. That is, social scientists are interested in such things as people's intellectual abilities, attitudes, personality characteristics, and values. Such attributes do not lend themselves easily (or at all) to physical measurement. We cannot look at a person and discern his or her attitudes or values, nor is there any ruler- or thermometer-like device we can use to measure them. Instead, we hypothesize the existence of theoretical entities known as *constructs* (also called *latent constructs, factors,* or *unobserved variables)* to account for certain characteristics or behaviors. For example, a researcher might recognize, based on long experience working with people, that some seem to learn faster and adapt more quickly to new situations than others. The researcher hypothesizes a construct now known as intelligence to account for this difference. Note that the researcher cannot directly observe people's intelligence, but must infer the existence of intelligence from observations of their behavior. Other examples of constructs include creativity, anxiety, attitudes toward gun control, altruism, propensity to buy a product, and aptitude for learning. All of these are latent in the sense that we cannot measure them directly but must devise some way of getting at them indirectly.

Measurement of constructs is therefore indirect, relying on samples of behavior such as responses to test items or observations of behavior. A *test* is a way of eliciting such behaviors. Here, and throughout this book, I use the term *test* to refer to a procedure for obtaining a sample of behavior that can be used to infer a person's level or

status on a construct of interest. The terms *measure, instrument,* and *scale* are often used in the same way as "test," and although some authors make distinctions among them, the terms are generally used somewhat interchangeably, a practice I follow in this book. As an example of a test, suppose a researcher theorizes that the ability to apply knowledge learned in one context to a new context is one aspect of intelligence. To measure this ability, the researcher would have to devise a series of tasks requiring test takers to apply knowledge learned in one subject area to other subject areas. These tasks could make up a test of the ability to apply knowledge to new contexts, which could then be administered to test takers.

I note several things about this procedure. First, the researcher would likely be able to think up many tasks that could elicit the desired ability. In fact, there may be limitless tasks that can measure many constructs (consider, as an example, the ability to add two-digit numbers). This implies that the tasks included on a test are typically a sample of all the possible tasks that might have been used. Second, the researcher would have to put some limitations on the manner in which test takers are allowed to complete the tasks. For example, the researcher might stipulate that test takers cannot consult outside resources, such as websites or friends, to help them complete the tasks. The researcher might also impose a time limit so that some test takers do not have more time than others to complete the tasks. Thus, the test would likely be administered under controlled, or standardized conditions. Third, although the researcher would like to assume that correct completion of the tasks was an indication of the ability to apply knowledge in new contexts, this may not be the case. Suppose some test takers are able to complete the task in a new subject area by some other means than generalizing knowledge from the original subject area. For example, suppose some test takers are completely unable to generalize their knowledge to the new subject area but know a great deal about the new subject area and are able to answer correctly based on that knowledge. Would the test still measure the ability to apply knowledge learned in one context to another? Probably not, because the test takers did not use that ability to answer the questions. This example points to a persistent problem in the measurement of constructs: it is always possible that the tasks used do not actually elicit the construct of interest.

PROBLEMS IN SOCIAL SCIENCE MEASUREMENT

In the previous section, I discussed some of the issues inherent in measurement in the social sciences. One issue is that tests are usually based on limited samples of behavior; we cannot ask every possible question or observe every instance of behavior. A related issue is that there is no one "correct" method of measuring a construct. In the previous example, the ability to apply knowledge in new contexts could have been measured by performance assessments in which test takers are given problems to solve in different subject areas, by multiple-choice tests asking test takers to choose the most likely outcome of a theory if applied in a new context, or by interviewing test takers about how they would solve the problem in a new context, just to name a few. Use of the different

measurement methods would get at somewhat different aspects of the ability, and each method would have its own advantages and disadvantages. The researcher developing the test would therefore have to carefully consider which type of test would be best aligned with the purpose(s) of testing. For example, different methods might be appropriate for testing theories about the ability than for selecting students for an advanced educational program.

Another testing issue is that many things can (and likely will) interfere with our measurement of the construct of interest. As indicated in the previous section, test takers may be able to complete the tasks using skills or abilities other than those the test was designed to measure. Or some test takers may have the requisite abilities but may be so anxious about the test that they fail to complete any of the tasks correctly. Other test takers may have the ability to complete the tasks but have limited English proficiency, causing them to misinterpret the tasks or instructions. Other types of interference are more relevant to attitude measurement. For example, respondents to attitude items may not answer truthfully because they know their attitudes are not politically correct and they do not want to draw attention to this; this tendency is known as *socially desirable responding*. Some respondents may have a tendency to choose a "neutral" or middle response option, whereas others may tend to choose more extreme response options. Such *response styles* are ubiquitous in the measurement of attitudes, personality characteristics, and psychological disorders. Those measuring psychological disorders must also contend with *malingering*—the tendency to exaggerate one's symptoms in an effort to obtain a particular diagnosis.

All of these issues in the measurement of constructs come under the broad heading of *errors of measurement*. It is important to understand that such measurement errors are part and parcel of most social science measurement. As a result, our measures are not perfect, but they should instead be thought of as approximations. Although some tests may provide quite good approximations, none are error-free. One of the tasks of those developing and using tests is therefore to be aware of the possibilities for error in test scores and to interpret and use test results with these possibilities in mind.

WHAT IS MEASUREMENT THEORY?

Another important task for those involved in social science measurement is to investigate the impact of measurement error on test results and to use the findings from these investigations to improve the tests and testing procedures. Such investigations are part of the broad field of *measurement* or *test theory*, known in psychology as *psychometrics*. These terms refer broadly to the study of methods for measuring constructs, and of their attendant problems. Measurement theory is therefore the study of how to develop tests that are as free as possible of measurement error and that yield the most appropriate measures of the desired constructs. Without good tests, social scientists would be unable to diagnose many learning disabilities and personality disorders, to study individual differences in constructs of interest, or to test theories involving these

constructs. Our measurements are the basis of our diagnoses and of our ability to test theories, which are therefore only as good as the measures underlying them. Good tests are thus crucial to both practical applications and to theory development in the social sciences.

MEASUREMENT DEFINED

So what do I mean by *measurement*? Stanley S. Stevens's (1946) definition of measurement as the "assignment of numerals to objects or events according to rules" (p. 677) is commonly cited. This definition was later amended to clarify that it is the properties of objects (usually people), such as the strength of their attitudes or their levels of altruism, and not the objects themselves that are measured. Note that, according to Stevens's definition, coding responses on a questionnaire with a "1" for male and a "2" for female constitutes measurement because this process involves assigning numerals (1 and 2) to properties of objects (male and female), according to the rule that 1 means male and 2 means female. Stevens defined four levels of measurement: nominal, ordinal, interval, and ratio. These levels are distinguished by the properties they include and are hierarchical in the sense that a higher level of measurement includes the properties of those lower in the hierarchy. In the sections that follow I describe the four levels and their properties. I also indicate the statistical operations for which Stevens felt each level was appropriate.

The Nominal Level of Measurement

Nominal measures are those for which the numbers serve only to distinguish different categories and do not have any real numerical meaning. The previous example of coding males as 1 and females as 2 exemplifies the nominal level of measurement. The only property of this type of measurement is that of distinctiveness. That is, the numbers distinguish the two categories of male and female but have no quantitative meaning. The coding of most demographic variables, such as political party or hair or eye color, constitute nominal measurement. Such measures can be transformed by applying any one-to-one substitution. In other words, we could substitute any other pairs of numbers, such as 3 and 4, or 65 and 83, for 1 and 2 because they serve to distinguish the categories equally well. The only transformation we cannot use is one in which the same number is used to represent both male and female categories because this would destroy the property of distinctiveness. Because the numbers used in nominal measurement have no numeric meaning, it is not appropriate to add, subtract, or otherwise manipulate them numerically. The only statistical indices appropriate at this level are those based on counts, such as the mode or the chi-square tests of independence and goodness of fit.

The Ordinal Level of Measurement

At the ordinal level of measurement the numbers represent a rank order of the properties of objects. The rank order could be based on size, speed, importance, correctness, or any other property capable of being ranked. Common examples of ordinal measures are the outcomes of a race (first place, second place, etc.), military ranks, and class rank. To the property of distinctiveness, ordinal measures therefore add the property of order. Although the numbers used in ordinal measurement imply rank order, the intervals between adjacent scale points are not assumed to be equal. Taking the outcomes of a race as an example, we know that the person finishing first is faster than the person finishing second, but we do not know how much faster because ordinal measurement does not tell us anything about the amounts by which scale points differ. We also do not know whether the time difference between those finishing first and second is the same as that between those finishing second and third because for ordinal measurement these intervals are not assumed to be equal. Ordinal measures can be transformed in any way that preserves the original order. Thus, we could substitute the numbers 3, 18, and 21 for the numbers 1, 2, and 3 without losing the properties of distinctiveness and order. Because ordinal measures cannot be assumed to have equal intervals, it is not meaningful or appropriate to perform arithmetic operations such as addition or subtraction on them. Statistical indices such as the mode, median or interquartile range are appropriate because these do not assume equal intervals.

Items measured on the commonly used "strongly disagree" to "strongly agree" Likert-type scale (see Chapter 5) are, strictly speaking, at the ordinal level of measurement. This is because we do not know whether respondents consider the psychological distance between "strongly agree" and "agree" to be the same as that between "strongly disagree" and "disagree," or any other adjacent scale points. Having said this, researchers differ in their willingness to treat data from Likert scales as having equal intervals. Some argue that such data probably have equal or nearly equal intervals and that little is lost, statistically speaking, by treating these data as interval. Others argue that this does not make sense unless we know that respondents do treat the intervals as equal and we generally do not have such knowledge.

The Interval Level of Measurement

In addition to the properties of distinctiveness and order, interval measures have the property of equal intervals. This means that the intervals between adjacent scale points are assumed to be the same across the entire scale continuum. A common example of interval level measurement is temperature as measured by the Fahrenheit or Centigrade scales, in which the difference in heat between scale points of 50° and 51° is the same as the difference between 90° and 91°. Interval measures can be transformed through any linear transformation of the form $y = a + bX$. Because of their equal-interval property,

it is appropriate to calculate nearly all parametric statistics, such as the mean, standard deviation, and correlation from interval level data.

The Ratio Level of Measurement

The ratio level is the highest of Stevens's (1946) levels of measurement. In addition to the properties of distinctiveness, order, and equal intervals, ratio-level measurement has the property of a true zero point. A true zero point is one that represents the absolute lack of the property being measured. For example, $0 indicates the absolute lack of any money. The Kelvin temperature scale is on a ratio scale because, on that scale, zero degrees indicates a complete lack of heat. This is not the case for the Fahrenheit and Centigrade scales, which is why they are relegated to the interval level of measurement. Many physical scales, such as height, weight, and time, as well as things that can be counted, such as the number of test items correct or the number of students in a classroom, are at the ratio level of measurement. Numbers on a ratio scale can be legitimately transformed only through multiplication of scale points by a constant. Adding a constant, as is permissible at the interval level of measurement, is not permissible at the ratio level because this would change the value of the zero point, rendering it nonabsolute. All parametric statistical operations are permissible for variables at the ratio level of measurement.

It may seem that test scores are at the ratio level of measurement. This depends, however, on how we want to interpret the scores. If we are content to interpret a test score as the number of points obtained on the test, the scores can be considered as ratio level. This is because the zero point can be appropriately interpreted as the absolute absence of any points obtained. However, if we want to interpret the test score as an indication of a particular level of knowledge or achievement, achieving the ratio level of measurement becomes much more problematic. This is because it is difficult to argue that a test score of zero means the absolute absence of any knowledge or achievement. A more likely interpretation is that a student earning a score of zero has some knowledge but does not have knowledge of the particular questions included on the test. It may be that if different questions had been asked, the student would have obtained a higher score. Or the student may have suffered from test anxiety, may have been unable to correctly interpret the test questions, or may have marked the answers incorrectly on the bubble sheet. As you can see, it is much more difficult to make ratio-level interpretations for abstract constructs such as achievement than for more concrete entities such as the number of points earned.

Criticisms of Stevens's Levels of Measurement

Although Stevens's (1946) levels of measurement are widely used in the social sciences, it is important to point out that not all experts agree with his conceptualizations. In particular, Joel Michell (1986, 1997) has argued forcefully that Stevens's definition of measurement does not adhere to the rules of quantitative structure. In other words,

Michell contends that Stevens's levels of measurement are not truly quantitative and that, even if they were, Stevens has provided no method for determining their quantitative properties. According to Michell (1997), measurement is defined as the "estimation or ratio of some magnitude of a quantitative attribute to a unit of the same attribute" (p. 358). Measurement, in Michell's definition, requires a quantitative structure in which the numeric relations (additivity or ratios) between points on the scale can be verified. Stevens's definition of measurement does not qualify because it is based on what Michell calls operationalism. This means that in Stevens's system a measure defined as being at a particular level of measurement is simply assumed to have the properties of that level of measurement. Stevens did not propose any methods for determining whether, for example, the equal-interval property of interval measures actually holds. Michell argues that measurement theories such as those advanced by Luce and his colleagues (Luce, Krantz, Suppes, & Tversky, 1990; Luce & Tukey, 1964), in which it is possible to test such quantitative properties of measures, are preferable because they allow for empirical verification of measurement properties.

Debates about the quantitative structure of social science measurement (or lack thereof) will likely continue in the coming years. Although I cannot predict the outcome of these debates, the history of testing may offer some insight about their origins. It is to this history that I now turn.

A BRIEF HISTORY OF TESTING

By any account, testing has had a long and storied history. Dubois (1970) traces so-called "modern" testing back to three major influences: Chinese (and, later, European and American) civil service examinations, assessment of individual differences in the early European and American psychological laboratories of the 19th and early 20th centuries, and the assessment of achievement in early European schools and universities. While it is true that educational and psychological testing as we know it today had its origins in these early testing efforts, other early societies also made use of various testing methods. For example, as we shall see, testing in early Greek society could be quite a harrowing experience. In this chapter, I will trace the history of testing from its early origins in China to the early educational and psychological tests developed in the United States, with a brief detour into the somewhat different procedures employed by the Greeks.

The Chinese Civil Service Examinations

Dubois (1970) dates the Chinese civil service examinations back to as early as 2200 B.C.E. Initially, examinations were conducted only for the purpose of evaluating civil servants to determine whether they should continue in office, a practice that took place every three years. At some point, however, Chinese rulers decided that examinations should also be used to choose candidates for civil service positions. The examinations were continued until 1905, at which time they were superseded by university credentialing.

Although the date of 2200 B.C.E. is commonly cited as the approximate origin of the Chinese examinations, there is some controversy about this date. For example, Bowman (1989) disputes Dubois's claim that these date back to 2200 B.C.E. stating that "[t]he attribution of examinations to dates of 2200 B.C.E. and 1115 B.C.E. is unsupported by evidence and at variance with what we now know about the societies of those times" (p. 577). Bowman dates these exams to between 200 and 100 B.C.E. However, Kim and Cohen (2007) defend Dubois's claim, stating that the earliest examinations were developed during the reign of Emperor Shun, which they date at 2255 B.C.E.–2205 B.C.E. (p. 5). Whatever the date, it seems safe to say that these examinations represent the earliest documented use of testing for widespread selection or evaluation purposes.

The major problem in dating the earliest civil service examinations is related to the fact that these were oral rather than written examinations. As Bowman (1989) points out, there are no written records about testing from the period covering the disputed date of 2200 B.C.E., for the very good reason that writing had not yet been developed (although some symbolic forms of communication had been used; see Kim & Cohen, 2007). The development of the Chinese character system is commonly ascribed to the Shang dynasty (c. 1500 B.C.E.). Dubois (1970) and Kim and Cohen state that during the later Chou dynasty candidates were examined in the "six arts" of music, archery, horsemanship, writing, arithmetic, and ceremonial protocol. Although some of these examinations may have been written, we can infer from the subject matter that most were oral or performance based. Dubois puts the date for these examinations at around 1115 B.C.E.

In contrast to the divergences of opinion regarding the beginnings of the early oral examination system, the advent of the largely written literary examinations is ascribed by most scholars (including Bowman) to the time of the Han dynasty (c. 206–220 B.C.E.). Many historical accounts from this period describe the use of examinations for the selection of civil service officials. Candidates were examined on the "five studies": civil law, military affairs, agriculture, revenue, and geography. These examinations were held at the district, provincial, and national levels, with candidates succeeding at a lower level going on to test at the higher levels. Students today who find the day-long procedures for the SAT tests grueling may be interested to know that the Chinese civil service examinations lasted three days and three nights!

China had no formal university or public school systems during this time. However, the civil service examinations served as a means of implementing a national curriculum. Because candidates had to be familiar with the "five studies" and, later in the period, with the Confucian classics, aspiring civil service employees from around the country were compelled to study these subjects in order to succeed. Thus, the examination system contributed to the stability of the Chinese empire both by ensuring a steady supply of qualified government administrators and by effecting a common knowledge base. Despite its success, the civil service examination system in China was abolished in 1905, superseded by university degrees and qualifications.

By this time, however, the Chinese system of civil service examinations had made its mark in Western countries, having been introduced by British diplomats and missionaries who were so impressed with the system that they suggested it be implemented

in England. Their recommendations led to the first such examination, introduced in 1833 to select trainees for the Indian Civil Service. The success of this venture led in turn to interest in the United States, where the Civil Service Act of 1883 established competitive tests for entry into government service.

Testing in Ancient Greece

In contrast to the Chinese civil service examinations, tests in ancient Greek societies were not written but were typically oral or performance based. Examinations were conducted for the purpose of determining whether men (and of course then it was only men) were qualified for various aspects of Greek life and citizenship; examinations were not, as far as we know, used to select citizens for government positions. Tests in physical skills, such as running, jumping, wrestling, and discus and javelin throwing, were very common and highly standardized in both execution and scoring. Examinations of mental achievements were also carried out but were generally less highly standardized than those in the physical arena. As an example, Doyle (1974) described a test of oral reading in which each student began reading where the last student left off, presenting the possibility for differences in text difficulty from one reader to the next. Overall, however, the Greeks were cognizant of the need for what we would now call standardization in testing.

Testing in ancient Greece was generally utilitarian and, in the mental sphere, focused on skills in the areas of rhetoric and recitation, which were considered the foundations of good citizenship. For the same reasons, tests of mental achievement tended to emphasize morality to a greater extent than intellectual prowess. Thus, students in Athens were expected to be able to speak extemporaneously on the issues of the day and were tested in this regard both formally and informally. Similarly, Spartan students were routinely asked questions such as "Who is the best man in the city" and expected to provide not only a ready answer but sound reasoning for their choice. One cannot help but think that such a requirement might serve us well in current democratic states. Students today who feel disappointed at receiving a low grade may take comfort in the following quotation from Plutarch (quoted in Doyle, 1974) regarding the assessment of Athenian students who, if "answered not to the purpose, had his thumb bit by the master" (p. 209).

For most of these tests, scoring was somewhat subjective and was typically based on the judgments of either the teacher or a larger audience of older citizens. Scoring does not appear to have been formalized and consisted mainly of either praise for good performance or punishment, usually physical, for poor performance. Although tests in early Greek societies were not systematically used for selection of civil servants as in China, testing did play a part in such selections. For example, Plato, in the *Republic*, discusses the importance of assessments of character as well as of knowledge and the ability to connect the various branches of knowledge in determining suitability for state offices. Thus, as in China, testing helped to both form and reinforce the national character.

Early European Testing

As in early Greece, the earliest testing in Europe was concerned with what today would be called student assessment. University exams are known to have been used in Europe as early as 1219 (at the University of Bologna) for determining eligibility for degrees. These exams were exclusively oral; written tests were not used until much later when the Jesuits pioneered the use of this format beginning in the sixteenth century. The parallels between the rules for these examinations, established in 1599, with those used today are striking. For example, examinees were enjoined to "be present in the classroom in good time," "come supplied with books and writing materials," "diligently look over what he has written, correct and improve it as much as he may wish," and "clearly know how much time is granted for writing, rewriting, and revision" (McGucken, 1932, quoted in DuBois, 1970, pp. 9–10).

At Oxford University, oral examinations for both the BA and MA degrees were introduced in 1636, and written exams can be dated back to at least 1803. The success of such an examination system has been credited with leading, at least in part, to easier acceptance of later civil service exams based on the Chinese model.

Beginnings of Psychological Measurement

As noted by Goodenough (1949), early research in experimental psychology was focused on the study of physical sensation. German psychophysical laboratories of the early 1800s, such as that of Wilhelm Wundt, were concerned with obtaining precise estimates of reaction time, visual and auditory perception, and other physical sensations under various conditions. Because many early students of psychology were also trained in such "hard" sciences as biology and physiology, it is not surprising that these researchers turned to such physical measures in their attempts to understand mental functioning. Unlike today's testing efforts, however, individual differences were not the focus of these studies. On the contrary, such differences were generally considered to be the result of imperfect control of experimental conditions, and every effort was made to design studies in which such differences were minimized.

Another influence on psychological research at this time was the publication of Darwin's *Origin of Species* in 1859. The possibility of an evolutionary basis for variation in mental abilities, suggested by Darwin's theories, was not lost on psychologists. Indeed, Francis Galton, a cousin of Darwin, wrote several works on the hereditability of scientific aptitude. Galton is well known for the establishment of laboratories for the collection of physical measurements such as height, weight, strength of pull, and discrimination of colors. The first of these laboratories was established at the 1884 International Health Exhibition in London, where over 9000 people paid three pence each to have their measurements taken and immortalized in the form of normal curves, the idea for which Galton adopted from the Belgian scientist Quetelet. In fact, it is Galton's development of the ideas for many of the statistical techniques used today for which he is best known. From his studies of variation in physical measurements, Galton realized

the power of the normal curve for organizing the vast amounts of data he had collected. He went on to develop the concept of correlation, although the actual mathematics was worked out by his protegé and friend Karl Pearson.

Galton's interests were wide ranging, and he soon became interested in expanding his anthropometric measures to include measures of a more psychological nature. He became aware of the work of James McKeen Cattell, who was in Germany at the time, studying psychology in Wundt's laboratory. According to Sokal (1987), Galton contacted Cattell to find out more about his apparatus for measuring reaction time. Sokal reports that it was through this contact with Galton (whom Cattell later called "the greatest man I have ever known") that Cattell became interested in the measurement of individual differences, which, as noted previously, were not of much interest to European psychologists at that time. After receiving his doctorate, Cattell served as an assistant in Galton's Anthropometric Laboratory, where he developed further tests of perception and memory. He continued his work in this area at the University of Pennsylvania, where he was appointed to a professorship in psychology, and later at Columbia University. It is in the series of tests he developed in the United States that we begin to see precursors of those commonly used today to study mental functioning. In 1890, these included tests of strength of squeeze, rate of arm movement, acuity of sensation, amount of pressure causing pain, least noticeable difference in weight, reaction time for sound, time taken to name colors, accuracy in bisecting a line, accuracy of judging time elapsed, and number of letters that could be repeated on one hearing. Although many of these tasks were clearly influenced by the psychophysical measures common at the time, some movement toward the measurement of mental processes can be detected.

The tension between the use of anthropometric measures and tests focused on the measurement of mental processes became more evident as psychologists continued their studies in the United States. In 1895, the American Psychological Association appointed a committee to study the feasibility of combining efforts across the different psychological laboratories in collecting data on mental and physical characteristics. While most members of the committee favored an emphasis on anthropometric measures, this view was not unanimous. James Baldwin of Princeton University argued that anthropometric tests were given too much weight and that tests of the higher mental processes were also necessary to more fully understand mental functioning. The proponents of anthropometric measures, however, prevailed. Arguments for the use of such measures, as put forth by Cattell, were largely utilitarian, emphasizing that psychological traits were simply too difficult to measure.

Unfortunately for their proponents, however, the utilitarianism of the anthropometric approach to measuring mental processes did not carry over into the applications envisioned for these measures. For example, Cattell felt that his tests could be used to evaluate academic abilities, and to this end he administered these tests to his students at Columbia. He also collected data on the course grades of these students and had his student Clark Wissler use the newly developed formula for the correlation coefficient to determine the degree of relationship of the tests with each other as well as with course grades. The results were disappointing, to say the least. Wissler reported a correlation

of –.05 between tests of reaction time and the ability to identify occurrences of the letter A in a grid of randomly arranged letters—two tasks that had been thought to be fairly similar. Similarly small correlations were found between scores on Cattell's tests and his students' course grades. After other researchers across the country found similar results, interest in the use of anthropometric measures began to wane.

Binet's Contributions to Intelligence Testing

Meanwhile, Alfred Binet was pursuing his well-known work in France with a focus that was much more in sympathy with those favoring an emphasis on higher mental abilities. Binet believed that differences in abilities to think and reason, to adapt to new conditions, and to solve problems were critical to mental functioning. The problem was to find suitable tasks to elicit these abilities, a problem to which Binet devoted much of the remainder of his short life. As early as 1896, Binet published an article in which he described tests of memory, mental imagery, imagination, attention, comprehension, suggestibility, aesthetic appreciation, force of will (operationalized as amount of effort in muscular tasks), moral sentiments, motor skill, and judgment of visual space; many of these survive in some form in today's "modern" aptitude tests.

During the period in which he was developing these tests, Binet served as a member and later president of a subgroup of the Société Libre pour l'Étude de l'Enfant (or the Free Society for the Study of the Education of Children, loosely translated). As part of his involvement with this group, he organized a working group concerned with the best way to educate what were then called "mentally retarded" children. This group recommended that such children should not be sent to special schools for the retarded unless an examination showed that the child would be unable to profit from regular education. The obvious problem was that no such examinations were then available. Although both medical and educational examinations were proposed as possibilities, Binet developed a new procedure, based on his ideas about measuring mental capacity, which he referred to as the psychological method. This became the first intelligence scale, developed by Binet and his colleague Theodore Simon in 1905. They argued that this was a better diagnostic tool than either medical tests for mental capacity, which were not diagnostic for all cases, or the educational tests of the time, which were primarily tests of memory.

Binet and Simon administered their early tests to both "normal" and retarded children in order to determine whether the tests would distinguish between these groups, as well as among children of different ages and levels of retardation. Tests were evaluated by giving them to children of different ages, based on the idea that the ability to answer more difficult questions should increase with age. Items were chosen systematically to cover different difficulty levels and, foreshadowing today's standardized ability and achievement tests, had highly standardized instructions for administration. The 1905 test contained many item formats still in use today, such as identifying parts of the body, obeying simple commands, identifying objects in pictures, repeating sentences and digit sequences, and identifying similarities.

The next version of Binet and Simon's scale, published in 1908, featured several new tests, including those requiring children to name as many words as possible in a given timeframe, naming the days of week, unscrambling sentences, executing three verbal commands sequentially, and detecting absurdities. Instead of being ordered by overall difficulty level, as in the previous version, these tests were ordered by age, based on the results obtained from administration to 303 Parisian school children ages 3–12. Age levels were determined by placing each test into levels based on the age at which the majority of students were able to pass it. For example, if the majority of 4-year-old children passed a test, it was assigned to level 4. This allowed for children's scores to be expressed as "mental levels" or, later, as "mental ages." Although Binet himself disliked the latter term, its use of a familiar reference point may have helped to sell the concept of ability testing to the general public (Anastasi & Urbina, 1997).

The third revision of the Binet–Simon Scale was published in 1911. The main revision at this point was the extension of the scale to the adult level. During this year, Binet also reported on a great many extensions of his scales, including plans for group in addition to individual testing, testing of military applicants and criminals, and studies of the relations between the Binet–Simon scales and school success. Unfortunately, he died in 1911 before these applications could be fully developed.

Testing in the United States

The Stanford–Binet Scale

Binet published his work widely, and soon psychologists in other countries began to adapt his work for their own use. Among these psychologists was Robert Yerkes, who would later go on to direct the development of the Army Alpha test for military recruits during World War I. However, the most successful American version of Binet's scale, still in use today, was developed by Lewis Terman of Stanford University; this is the well-known Stanford–Binet scale. Although over half of the tests on Terman's scale were based on the 1911 Binet–Simon scale, others were developed as part of his doctoral dissertation. These tests were administered to over 2,500 children and adults of various levels of ability. Terman put a heavy emphasis on clear and consistent instructions for administering his scale, spending six months in training the examiners. Students who have scored the current version of the Stanford–Binet will be impressed by the fact that Terman scored every one of the over 2,500 tests himself to assure reliability. Goodenough (1949) tells us that Terman's devotion to training and scoring seems to have paid off, as higher correlations with school success and other indicators of ability were found in Terman's study than in the earlier unsuccessful studies by Cattell and Wissler.

Another of Terman's contributions was the development of the intelligence quotient. Although the concept was originally developed by William Stern, Terman was the first to fully develop this type of scoring in his 1916 scale. To implement the idea, he moved tests from one age level to another until the median mental age at each age level was equal to its corresponding chronological age level. Thus, children at each age would have an average intelligence quotient of 1, or 100%.

Group Testing

The Army Alpha

The Stanford–Binet was then, as it is now, individually administered. Group testing was not unknown at the time that scale was developed, but it was not commonly used. However, one of Terman's students, Arthur Otis, was interested in the idea and worked with Terman to develop a group intelligence test. Otis developed what was then the first intelligence test that could be objectively scored, relying heavily on use of an early form of the multiple-choice item. The development of this test proved to be serendipitous because at about the same time Robert Yerkes, as president of the American Psychological Association, had focused his attention on determining how that organization could best help the country in its preparations for World War I. Yerkes organized several committees to study this issue, and because of his previous work in ability testing, he chaired the committee concerned with psychological examinations for military recruits. Terman, among other testing pioneers in the United States, was asked to be on the committee and brought with him the group testing materials developed by his student Otis. The committee quickly decided that the development of psychological tests would provide the most benefit to the war effort and that all recruits should be tested. Given the decision to test all, a group test was deemed most practical, and many of the materials developed by Otis were either used outright or adapted for use in the new test.

The speed with which the committee developed the resulting test has probably not been equaled since and is surely the envy of those working in today's testing companies. In a brief seven days, the committee reviewed different types of items, chose 10 of these types for development of subtests, wrote enough items to create 10 different forms of the test, and prepared one operational test for tryout. The operational test was tried out on a variety of groups, including inmates in a reformatory, high school students, and aviation recruits. Eight months after the initial meeting of the committee, the Army Alpha, as it was called, was ready for military service. Only 8 of the original 10 subtests survived the tryouts: oral directions, arithmetical reasoning, practical judgment, synonym/antonym, disarranged sentences, number series completion, analogies, and information. The Army Alpha represented the first wide-scale use of the multiple-choice item. Although difficult to imagine today, such items were not in common usage at that time. An example, as reported by DuBois (1970), was the following:

> Why ought a grocer to own an automobile? Because:
>
> _____ it looks pretty.
>
> _____ it is useful in his business.
>
> _____ it uses rubber tires.
>
> _____ it saves railroad fare.

Overall, at least 1,250,000 recruits were tested with Army Alpha. For recruits who were unable to read and write in English, the Army Beta, a nonverbal version of the test,

was developed and used. As with Terman's Stanford–Binet, strict attention was paid to the clarity and consistency of instructions.

The development and use of the Army Alpha (and Beta) is widely recognized today as the impetus for the development of a wide variety of group tests, as well as for greater involvement of psychologists and psychological testing in areas ranging from education to industry. After World War I, group testing became all the rage, and group ability and achievement tests, as well as interest inventories, occupational aptitude tests, and personality inventories, began to proliferate.

Group Achievement Tests

Joseph Rice was an early proponent of using scientific means to improve learning in schools. He reasoned that before determining whether students were learning at an adequate rate, he must first know how much students could be expected to learn in a given time period. To this end, Rice administered standardized tests of spelling and arithmetic to thousands of children. He used the results to determine the average level of achievement that could be expected at different grades. Although Rice was not involved in test development per se, the development of these early normative data paved the way for their use in later standardized achievement tests. Rice's work caught the interest of E. L. Thorndike who, along with his students at Columbia University's Teachers College, developed many of the early educational achievement tests. Based on this early work, the first achievement battery, which included tests in several school subjects, was developed by Thorndike's student Truman Kelley, along with Terman and Giles Ruch. This was the Stanford Achievement Test, still in use today in its 10th edition.

Occupational Interest Inventories

Thorndike did not confine his interests to measures of achievement. He was involved in a wide variety of testing applications, one of which was the Thorndike Intelligence Examination for High School Graduates (1919), an early college admissions test. Thorndike also began work on the measurement of academic interests, work that was followed up by Truman Kelley. Kelley's interest inventory expanded on Thorndike's by asking students to rate their level of interest in various magazines, books, and activities. K. M. Cowdery, in his doctoral dissertation under Kelley, made further refinements in this area, including items measuring interest in sports and amusements, types of pets, reading material, and types of people. Cowdery was among the first to use an empirical keying system in which items were selected on the basis of their ability to discriminate between members of different professions and demonstrated for the first time that people in different professions have different patterns of interests. Edward Strong further refined Cowdery's scale, publishing the first version of the Strong Vocational Interest Blank in 1927. He continued to work on this scale until his death 36 years later, developing new items and conducting studies of the scale's reliability and validity. Finally, G. Frederic Küder (1934) developed a new form of item in which test takers were forced to choose among three alternatives rather than simply indicating whether or not they liked something.

This is the so-called forced-choice item that remains popular today in scales such as the Myers–Briggs Type Indicator.

Testing in Business and Industry

In addition to the Army Alpha, psychologists involved in test development for the armed forces created and evaluated tests of specific aptitudes for jobs ranging from stenography to piloting an airplane. Those involved in the development of these tests included Yerkes and Thorndike, as well as such notables as L. L. Thurstone. Thurstone developed a test for telegraphers and in so doing demonstrated the basic procedures, still in use today, for validating a test against an external criterion (which, for the telegrapher's test, was the highest receiving speed). Thurstone also developed a test for selecting office workers for nonmilitary occupations, which included tests of such abilities as checking errors in arithmetic, finding misspelled words, and alphabetizing. The tests developed during and after World War I were not the first to be developed, however. DuBois (1970) notes that as early as the 1880s the U.S. Civil Service Commission had begun development of a series of tests for applicants for government jobs, an idea borrowed from the successful use of such procedures in Great Britain (which, in turn, as already noted, got the idea from the Chinese civil service examinations). After World War I, the Civil Service Commission contacted some of the psychologists involved with the Army Alpha and arranged for an experimental administration of that test to clerical workers. This was followed up with the development and validation of various occupational aptitude tests by researchers at the Commission.

Personality Assessment

As Anastasi and Urbina (1997) note, the term *personality testing* is typically used to refer to nonintellectual facets of human behavior, such as attitudes, values, beliefs, emotional states, and relationships with others. Test development in this area paralleled that in the area of intelligence testing, to some extent, beginning with early work with free association tests by Galton and the early German psychophysical laboratories, and later developing into more standardized measures due to the impetus of World War I. Although Galton experimented with free association techniques by trying them out on himself, the most systematic early attempts at using these methods were made in the early German psychophysical laboratories of Wundt by his student Emil Kraepelin, who experimented with a type of free association technique in which examinees were instructed to reply to word prompts with the first word that came to mind. Kraepelin's interest was in determining the effects of physical factors such as fatigue and hunger on mental processes, but the technique was seized upon as a way to study mental illness by such notables as Carl Jung. In 1921, Jung's student Hermann Rorschach famously adapted the procedures by substituting inkblots for the words that had previously been

used as a stimulus. (Interestingly, this technique had earlier been proposed by Alfred Binet and Theodore Simon.)

Robert Woodworth is generally credited with being one of the first to develop a group, self-report test for personality. His measure, later known as the Woodworth Personal Data Sheet, was motivated by the need to assess the susceptibility of military recruits to shell shock, or what might now be called posttraumatic stress disorder. Although the war ended before Woodworth could thoroughly test his instrument, its development set the stage for the use of group testing of personality.

SUMMARY

A construct is a theoretical entity hypothesized to account for particular behaviors or characteristics of people. Examples of constructs abound in the social sciences and include creativity, intelligence, various abilities and attitudes, personality characteristics, and value systems. Such constructs are latent in the sense that they are not directly observable but must be measured indirectly, relying on samples of behavior that are thought to characterize them. These samples of behavior can take the form of responses to test items or other questions, performances on physical or other tasks, behavioral observations, or any other methods thought to elicit the construct. A test is a method for eliciting these samples of behavior, such as paper and pencil (or, increasingly, computerized) questions to which test takers respond, physical tasks test takers perform, or behaviors that are elicited through overt or covert means and observed and coded by others. The fact that measurement of constructs is necessarily indirect leads to problems with their measurement. Perhaps most important among these problems is that the behavior sample thought to elicit the construct does not do so. This can happen for a variety of reasons, but perhaps the most common one is that the behavior sample was based on insufficient knowledge of the construct, of the method of measurement, or of the match between the two. Measurement of constructs is also based, in most cases, on a limited sample of the possible test questions, performances, or behaviors that are possible. As a result, our measurement of constructs is, to some extent, incomplete. In addition to these issues, test anxiety, inability to understand test questions or instructions, malingering, and response styles such as socially desirable responding can result in inaccurate measurement of the intended construct. Such errors of measurement are endemic in social science measurement, rendering the jobs of those who develop, administer, and interpret tests difficult. One purpose of this book is therefore to raise awareness of the potential pitfalls in testing and to provide those involved in testing with the background necessary to develop, use, and interpret test results appropriately.

This chapter also reviewed the history of testing, pointing out that testing developments in areas as widely varied as employment, intelligence, achievement, and personality have developed in much the same way. In many cases, the same researchers were involved in work in all of these areas, and improvements were freely shared across both substantive

and geographical areas. Given this fact, it is not surprising that the procedures for test development in different areas share many of the same procedures for item development and evaluation. It is on these procedures that we will focus in the remaining chapters.

EXERCISES

1. Define the following terms:
   a. *Test*
   b. *Construct*
2. Discuss two problems inherent in social science measurement.
3. According to Stevens's definitions, what are the levels of measurement of the following measurements? Justify your answers.
   a. Recycling behavior as measured by the following scale:

      1 = "never recycle"
      2 = "sometimes recycle"
      3 = "usually recycle"
      4 = "always recycle"

   b. Assigning the numbers 1, 2, and 3 to survey respondents living in houses, apartments, and condominiums, respectively.
   c. Distance traveled to work.
4. A researcher assigns the numbers 1, 2, and 3 to distinguish urban, suburban, and rural areas, respectively.
   a. Does this constitute measurement according to Stevens's definition? Why or why not?
   b. Does this constitute measurement according to Michell's definition? Why or why not?
5. In what ways did testing in ancient Greece differ from testing in ancient China?
6. Many early psychological tests were based on anthropometric measures such as reaction time and ability to perceive differences in weights. Why did early tests rely so much on this type of measurement? What is one reason that such measures were eventually abandoned?
7. What was the impetus behind Binet's development of the first intelligence test in 1905?
8. How did multiple-choice items come to be used on the Army Alpha test?