PART ONE

Technical Issues

CHAPTER

1 Fundamental Issues in Measurement

INTRODUCTION

Societies and individuals have always had to make decisions. Making decisions is a requirement of daily life for everyone. We all decide when to get up in the morning, what to have for breakfast, and what to wear; we make some kinds of decisions with such regularity that we hardly think about the process. Other decisions that we make less frequently

require careful thought and analysis: Which college should I attend? What should I major in? Should I accept a job with XYZ Enterprises? Decisions of this kind are best made based on information about the alternative choices and their consequences: Am I interested in a particular college? What job prospects will it open for me? What are the chances that I will succeed? What are the benefits and working conditions at XYZ Enterprises?

Other decisions are made by individuals acting for the larger society. Daily, teachers must make decisions about the best educational experiences to provide for students, based on an assessment of their students' current knowledge and abilities. A school psychologist may have to decide whether to recommend a special educational experience for a child who is having difficulty in reading or mathematics. School, district, and state educational administrators must make decisions about educational policy and often have to produce evidence for state and local school boards and state legislatures on the achievement of students. Employers must decide which job applicants to hire and which positions they should fill. A college counselor must decide what action to take with a student who is having difficulty adjusting to the personal freedom that a college environment provides. The list of decisions that people must make is as long as the list of human actions and interactions.

We generally assume that the more people know about the factors involved in their decisions, the better their decisions are likely to be. That is, more and better information is likely to lead to better decisions. Of course, merely having information is no guarantee that it will be used to the best advantage. The information must be appropriate for the decision to be made, and the decision maker must also know how best to use the information and what inferences it does and does not support. Our purpose in this book is to present some of the basic concepts, tools, practices, and methods that have been developed in education and psychology to aid in the decision-making process. With these tools and methods, potential decision makers will be better prepared to obtain and use the information needed to make sound decisions.

A LITTLE HISTORY

Although educational measurement has gained considerable prominence in recent decades, particularly with the No Child Left Behind Act and the rise in interest in accountability, the formal evaluation of educational achievement and the use of this information to make decisions have been going on for centuries. As far back as the dawn of the common era, the Chinese used competitive examinations to select individuals for civil service positions (DuBois, 1970; R. M. Thorndike, 1990a). Over the centuries, they developed a system of checks and controls to eliminate possible bias in their testing—procedures that in many ways resembled the best of modern practice. For example, examinees were isolated to prevent possible cheating, compositions were copied by trained scribes to eliminate the chance that differences in penmanship might affect scores, and each examination was evaluated by a pair of graders, with differences being resolved by a third judge. Testing sessions were extremely rigorous, lasting up to 3 days. Rates of passing were low, usually less than 10%. In a number of ways, Chinese practice served as a model for developing civil service examinations in western Europe and America during the 1800s.

Formal measurement procedures began to appear in Western educational practice during the 19th century. For several centuries, secondary schools and universities had been using essays and oral examinations to evaluate student achievement, but in 1897, Joseph M. Rice used some of the first uniform written examinations to test spelling achievement of students in the public

schools of Boston. Rice wanted the schools to make room in the curriculum for teaching science and argued that some of the time spent on spelling drills could be used for that purpose. He demonstrated that the amount of time devoted to spelling drills was not related to achievement in spelling and concluded that this time could be reduced, thus making time to teach science. His study represents one of the first times tests were used to help make a curricular decision.

Throughout the latter half of the 19th century, pioneering work in the infant science of psychology involved developing new ways to measure human behavior and experience. Many measurement advances came from laboratory studies such as those of Hermann Ebbinghaus, who in 1896 introduced the completion test (fill in the blanks) as a way to measure mental fatigue in students. Earlier in the century, the work of Ernst Weber and Gustav Fechner on the measurement of sensory processes had laid the logical foundation for psychological and educational measurement. Other important advances, such as the development of the correlation coefficient by Sir Francis Galton and Karl Pearson, were made in the service of research on the distribution and causes of human differences. The late 1800s were characterized by DuBois (1970) as the *laboratory period* in the history of psychological measurement. This period has also been called the *era of brass instrument psychology* because mechanical devices were often used to collect measurements of physical or sensory characteristics.

Increasing interest in measuring human characteristics in the second half of the 19th century can be traced to the need to make decisions in three contexts. First, enactment of mandatory school attendance laws resulted in a growing demand for objectivity and accountability in assessing student performance in the public schools. These laws brought into the schools for the first time a large number of students who were of middle or lower socioeconomic background and were unfamiliar with formal education. Many of these children performed poorly and were considered by some educators of the time to be "feebleminded" and unable to learn. The development of accurate measurement methods and instruments was seen as a way to differentiate children with true mental handicaps from those who suffered from disadvantaged backgrounds. Second, the medical community was in the process of refining its ideas about abnormal behavior. Behavioral and psychological measurements were seen as a way to classify and diagnose patients. Third, businesses and government agencies began to replace patronage systems for filling corporate and government jobs with competitive examinations to assess prospective employees' abilities. Tests began to be used as the basis of employee selection.

Not until the first years of the 20th century did well-developed prototypes of modern educational and psychological measurements begin to appear. Although it is difficult to identify a single critical event, the 1905 publication of the Binet-Simon scales of mental ability is often considered to be the beginning of the modern era in what was at the time called mental measurement. The Binet-Simon scales, originally published in French but soon translated into English and other languages, have been hailed as the first successful attempt to measure complex mental processes with a standard set of tasks of graded complexity. These scales were designed to help educators identify students whose mental ability was insufficient for them to benefit from standard public education. On the basis of the mental measurement, a decision was then made whether to place these students in special classes. Subsequent editions of the scales, published in 1908 and 1911, contained tasks that spanned the full range of abilities for school-age children and could be used to identify students at either extreme of the ability continuum. (See R. M. Thorndike, 1990a, for a more complete description of these scales.)

At the same time that Binet and Simon were developing the first measures of intelligence, E. L. Thorndike and his students at Teachers College of Columbia University were tackling problems related to measuring school abilities. Their work ranged from theoretical developments on

the nature of the measurement process to the creation of scales to assess classroom learning in reading and arithmetic and also level of skill development in tasks such as handwriting. The era of mental testing had begun.

It is convenient to divide the history of mental testing in the 20th century into six periods: an early period, a boom period, a first period of criticism, a battery period, a second period of criticism, and a period of accountability.

The Early Period

The *early period,* which comprises the years before American entry into World War I, was a period of tentative exploration and theory development. The Binet-Simon scales, revised twice by Binet, were brought to the United States by several pioneers in measurement. The most influential of these was Lewis Terman of Stanford University. In 1916, Terman published the first version of a test that is still one of the standards by which measures of intelligence are judged: the Stanford-Binet Intelligence Scale. (The fifth edition of the Stanford-Binet was released in 2003.) Working with Terman, Arthur Otis began to explore the possibility of testing the mental ability of children and adults in groups. In Australia, S. D. Porteus prepared a maze test of intelligence for use with people with hearing or language handicaps.

In 1904 Charles Spearman published two important theories relating to the measurement of human abilities. The first was a statistical theory that proposed to describe and account for the inconsistency in measurements of human behavior. The second theory claimed to account for the fact that different measures of cognitive ability showed substantial consistency in the ways in which they ranked people. The statistical theory to describe inconsistency has developed into the concept of reliability that we will discuss in Chapter 4. Spearman's second theory, that there is a single dimension of ability underlying most human performance, played a major role in determining the direction that measures of ability took for many years and is still influential in theories of human cognitive abilities. Spearman proposed that the consistency of people's performance on different ability measures was the result of the level of general intelligence that they possessed. We will discuss modern descendants of this theory and the tests that have been developed to measure intelligence in Chapter 12.

The Boom Period

American involvement in World War I created a need to expand the army very quickly. For the first time, the new science of psychology was called on to play a part in a military situation. This event started a 15-year *boom period* during which many advances and innovations were made in the field of testing and measurement. As part of the war effort, a group of psychologists led by Robert Yerkes expanded Otis's work to develop and implement the first large-scale group testing of ability with the Army Alpha (a verbal test) and the Army Beta (a test using mazes and puzzles similar to Porteus's that required no spoken or written language). The Army Alpha was the first widely distributed test to use the multiple-choice item form. The first objective measure of personality, the Woodworth Personal Data Sheet, was also developed for the army to help identify those emotionally unfit for military service. The Alpha and Beta tests were used to select officer trainees and to remove those with intellectual handicaps from military service.

In the 12 years following the war, the variety of behaviors that were subjected to measurement continued to expand rapidly. E. K. Strong and his students began to measure vocational interests to help college students choose majors and careers consistent with their interests.

Measurements of personality and ability were developed and refined, and the use of standardized tests for educational decisions became more widespread. In 1929, L. L. Thurstone proposed ways to scale and measure attitudes and values. Many people considered it only a matter of a few years before accurate measurement and prediction of all types of human behavior would be achieved.

The period immediately following World War I was also a low point for the mental testing movement. High expectations about what test scores could tell us about people's abilities and character led test developers and users to place far too much reliance on the correctness of test scores. The results of the U.S. Army testing program revealed large score differences between White American examinees and those having different ethnic backgrounds. Low test scores for African Americans and immigrants from southern and eastern Europe were interpreted as revealing an intellectual hierarchy, with people of northern European ancestry ("Nordics") at the top. Members of the lowest scoring ethnic groups, particularly those of African ancestry, were labeled "feebleminded." A number of critics, most notably Walter Lippmann, questioned both the tests themselves and the conclusions drawn from the test scores.

The First Period of Criticism

The 1930s saw a crash not only in the stock market but also in the confidence in and expectations for mental measurement. This time covered a *period of criticism* and *consolidation*. To be sure, new tests were published, most notably the original Kuder scales for measuring vocational interests, the Minnesota Multiphasic Personality Inventory, and the first serious competitor for the Stanford-Binet, the Wechsler-Bellevue Intelligence Scale. Major advances were also made in the mathematical theory underlying tests, particularly L. L. Thurstone's refinements of the statistical procedure known as factor analysis. However, it was becoming clear that the problems of measuring human behavior had not all been solved and were much more difficult than they had appeared to be during the heady years of the 1920s.

The rapid expansion in the variety of tests being produced, the increasing use of test scores for decision making, and the criticisms of testing in the press led a young psychologist named Oscar Buros to call on the professional psychological and educational testing community to police itself. Buros observed that many tests had little or no objective evidence to support the uses to which they were being put. In 1935, he initiated the *Mental Measurements Yearbook (MMY)* as a place where critical reviews of tests and testing practices could be published. His objective was to obtain reviews of tests from the leading experts in testing, which in turn would cause test producers to provide better tests and more evidence supporting specific uses of tests. As we shall see in Chapter 6, the *MMY* publications remain one of the best sources of information about tests.

The Battery Period

In the 1940s, psychological measurement was once again called on for use in the military service. As part of the war effort, batteries of tests were developed that measured several different abilities. Based on the theory developed by Thurstone and others that there were several distinct types or dimensions of abilities, these test batteries were used to place military recruits in the positions for which they were best suited. The success of this approach in reducing failure rates in various military training programs led the measurement field into a period of emphasis on test batteries and factor analysis. For 25 years, until about 1965, efforts were directed toward analyzing the dimensions of human behavior by developing an increasing variety of tests of ability and

personality. Taxonomies of ability, such as those of Bloom (1956) and Guilford (1985), were offered to describe the range of mental functioning.

During the 1950s, educational and psychological testing grew into a big business. The use of nationally normed, commercially prepared tests to assess student progress became a common feature of school life. The Scholastic Aptitude Tests (SAT, now called the Scholastic Assessment Tests) or the American College Testing Program (ACT Assessment) became almost universally required as part of a college admissions portfolio. Business, industry, and the civil service system made increasing use of measurements of attitudes and personality, as well as ability, in hiring and promotion decisions. The General Aptitude Test Battery (GATB) was developed by the U.S. Employment Service, and other test batteries were developed by private testing companies to assist individuals and organizations in making career and hiring decisions. Patients in mental institutions were routinely assessed through a variety of measures of personality and adjustment. In 1954, led by the American Psychological Association, the professional testing community published a set of guidelines for educational and psychological tests to provide public standards for good test development and testing practice. Testing became part of the American way of life. The widespread use—and misuse—of tests brought about a new wave of protests.

The Second Period of Criticism

The beginning of a *second period of criticism* was signaled in 1965 by a series of congressional hearings on testing as an invasion of privacy. The decade of the 1960s was also a time when the civil rights movement was in full swing and women were reacting against what they perceived to be a male-dominated society. Because the ability test scores of Blacks were generally lower than those of Whites, and the scores of women were lower than those of men in some areas (although they were higher than men's scores in other areas), tests were excoriated as biased tools of White male oppression. Since that time, debate has continued over the use of ability and personality testing in public education and employment. A major concern has been the possible use of tests to discriminate, intentionally or otherwise, against women or members of minority groups in education and employment. As a result of this concern, the tests themselves have been very carefully scrutinized for biased content, certain types of testing practices have been eliminated or changed, and much more attention has been given to the rights of individuals. The testing industry responded vigorously to the desire to make tests fair to all who take them, but this has not been sufficient to forestall both legislation and administrative and court decisions restricting the use of tests. A recent example of an administrative decision is the proposal to eliminate performance on the SAT as a tool in making admissions decisions at institutions in the University of California system. This situation is unfortunate because it deprives decision makers of some of the best information on which to base their actions. In effect, we may have thrown out the baby with the bath water in our efforts to eliminate bias from the practice of testing. In Chapter 8 we will take a closer look at the controversies surrounding educational and psychological uses of tests.

The Age of Accountability

At the same time that public criticism of testing was on the rise, governments were putting greater faith in testing as a way to determine whether government-funded programs were achieving their objectives. With passage of the 1965 Elementary and Secondary Education Act (ESEA), federally funded education initiatives began to include a requirement that programs report some form of assessment, often in the form of standardized test results. In 2002, President George

W. Bush signed the No Child Left Behind (NCLB) act into law. The primary goal of NCLB is to ensure that all public school students achieve high standards of performance in the areas of reading-language arts, mathematics, and science. To meet this goal, each state was required to generate a set of rigorous achievement standards and to develop assessments measuring student proficiency relative to those standards. States must hold schools and school districts accountable for the performance of their students. Low-performing schools are subject to a variety of interventions ranging from technical assistance to sanctions and restructuring. Many states have also enacted laws requiring students to pass standardized tests in order to earn a high school diploma. While this high-stakes use of testing is not new (Chinese practices of 1000 years ago had greater social consequences and were far more demanding), the number of states using mandatory exams as a condition of graduation is steadily increasing. Scores on the Washington Assessment of Student Learning, for example, are used to meet the accountability requirements of NCLB and to determine, beginning with the graduating class of 2008, which students meet the achievement standards necessary for a high school diploma. Nationwide, NCLB also requires new teachers to pass standardized tests to earn certification and demonstrate the subject matter competence necessary to be considered "highly qualified." Thus, at the same time that widespread criticism of tests, particularly standardized tests used to make educational decisions, has arisen, the role of such tests in ensuring that schools are accountable for the learning of their students has steadily increased.

TYPES OF DECISIONS

Educational and psychological evaluation and measurement have evolved to help people make decisions related to people, individually or in groups. Teachers, counselors, school administrators, and psychologists in business and industry, for example, are continuously involved in making decisions about people or in helping people make decisions for and about themselves. The role of measurement procedures is to provide information that will permit these decisions to be as informed and appropriate as possible.

Some decisions are *instructional;* many decisions made by teachers and school psychologists are of this sort. An instructional decision may relate to a class as a whole: For example, should class time be spent reviewing "carrying" in addition? Or does most of the class have adequate competency in this skill? Other decisions relate to specific students: For example, what reading materials are likely to be suitable for Mary, in view of her interests and level of reading skill? If such decisions are to be made wisely, it is important to know, in the first case, the overall level of skill of the class in "carrying" and, in the second, how competent a reader Mary is and what her interests are.

Some decisions are *curricular*. A school may be considering a curricular change such as introducing computer-assisted instruction (CAI) to teach the principles of multiplying fractions or a Web-based module on African geography. Should the change be made? A wise decision hinges on finding out how well students progress in learning to multiply fractions using CAI or African geography from Web materials rather than the conventional approaches. The evidence of progress, and hence the quality of the decision, can only be as good as the measures of mathematics competence or geography knowledge we use to assess the outcomes of the alternative instructional programs.

Some decisions are *selection* ones made by an employer or a decision maker in an educational institution. A popular college must decide which applicants to admit to its freshman class.

Criteria for admission are likely to be complex, but one criterion will usually be that each admitted student is judged likely to be able to successfully complete the academic work that the college requires. When combined with other sources, such as high school grades, standardized tests of academic ability such as the SAT or ACT Assessment can add useful information about who is most likely to succeed at the college. Selection decisions also arise in employment. The employer, seeking to identify the potentially more effective employees from an applicant pool, may find that performance in a controlled testing situation provides information that can improve the accuracy and objectivity of hiring decisions, resulting in improved productivity and greater employee satisfaction.

Sometimes decisions are *placement*, or *classification*, decisions. A high school may have to decide whether a freshman should be put in the advanced placement section in mathematics or in the regular section. An army personnel technician may have to decide whether a recruit should be assigned to the school for electronic technicians or the school for cooks and bakers. A family doctor makes a classification decision when he or she diagnoses a backache to be the result of muscle strain or a pinched nerve. For placement decisions, the decision maker needs information to help predict how much the individual will learn from or how successful the candidate will be in each of the alternative programs. Information helps the person making a classification decision to identify the group to which the individual most likely or properly belongs.

Finally, many decisions can best be called *personal* decisions. They are the choices that each individual makes at the many crossroads of life. Should I plan to go on for a master's degree or to some other type of postcollege training? Or should I seek a job at the end of college? If a job, what kind of job? In light of this decision, what sort of program should I take in college? Guidance counselors frequently use standardized tests to help young adults make decisions like these. The more information people have about their own interests and abilities, the more informed personal decisions they can make.

MEASUREMENT AND DECISIONS

Educational and psychological measurement techniques can help people make better decisions by providing more and better information. Throughout this book, we will identify and describe properties that measurement devices must have if they are to help people make sound decisions. We will show the form in which test results should be presented if they are to be most helpful to the decision maker. As we look at each type of assessment technique, we will ask, "For what types of decisions can the particular technique contribute valuable information?" We must be concerned with a variety of factors, including poor motivation, emotional upset, inadequate schooling, or atypical linguistic or cultural background, all of which can distort the information provided by a test, questionnaire, or other assessment procedure. We will also consider precautions that need to be observed in using the information for decision making.

The Role of Values in Decision Making

Measurement procedures do not make decisions; *people* make decisions. At most, measurement procedures can provide information on some of the factors that are relevant to the decision. The SAT can provide an indication of how well Grace is likely to do in college-level work. Combined with information about how academically demanding the engineering program is at Siwash

University, the test score can be used to make a specific estimate of how well Grace is likely to do in that program. However, only Grace can decide whether she should go to Siwash and whether she should study engineering. Is she interested in engineering? Does she have a personal reason for wanting to go to Siwash rather than to some other university? Are economic factors of concern? What role does Grace aspire to play in society? Maybe she has no interest in further education and would rather be a competitive surfer.

This example should make it clear that decisions about courses of action to take involve *values* as well as facts. The SAT produces a score that is a *fact,* and that fact may lead to a prediction that Grace has five chances in six of being admitted to Siwash and only one chance in six of being admitted to Stanford. But, if she considers Stanford to be 10 times more desirable than Siwash, it still might be a sensible decision for her to apply to Stanford despite her radically lower chance of being admitted. The test score provides no information about the domain of values. This information must be supplied from other sources before Grace can make a sensible decision. One of the important roles that counselors play is to make people aware of the values they bring to any decision.

The issue of values affects institutional decision makers as well as individuals. An aptitude test may permit an estimate of the probability of success for a Black or a Hispanic or a Native American student, in comparison with the probability of success for a White or Asian student, in some types of academic or professional training. However, an admission decision would have to include, explicitly or implicitly, some judgment about the relative value to society of adding more White or Asian individuals to a profession in comparison with the value of having increased Black, Hispanic, or Native American representation. Concerns about social justice and the role of education in promoting equality, both of opportunity and of outcome, have assumed an increasing importance in decision making over the last 35 years. However, in recent years three states, California, Washington, and Michigan, have passed laws prohibiting the use of race in educational decision making.

Issues of value are always complex and often controversial, but they are frequently deeply involved in decision making. Very few decisions are value neutral. It is important that this fact be recognized, and it is also important that assessment procedures that can supply better information not be blamed for the ambiguities or conflicts that may be found in our value systems. We should not kill the messenger who brings us news we do not want to hear, nor should we cover our eyes or ears and refuse to receive the news. Rather, we should consider policies and procedures that might change the unwelcome facts.

As we suggested at the beginning of this chapter, all aspects of human behavior involve making decisions. We *must* make decisions. Even taking no action in a situation is a decision. In most cases, people weigh the evidence on the likelihood of various outcomes and the positive and negative consequences and value implications of each possible outcome. The role of educational and psychological assessment procedures can be no more than to provide some of the information on which certain kinds of decisions may be based. The evidence suggests that, when properly used, these procedures can provide useful information that is more accurate than that provided by alternate approaches. Your study of educational and psychological assessment should give you an understanding of the tools and techniques available for obtaining the kinds of information about people that these measures can yield. Beyond that, such study should provide criteria for evaluating the information that these tools offer, for judging the degree of confidence that can be placed in the information, and for sensing the limitations inherent in that information.

After voters in the state of California voted to end the use of ethnic identity as a factor in university admissions, state education officials proposed to eliminate the use of SAT scores in

admissions decisions. The apparent reason for the decision not to use the test scores is that White and Asian American students typically earn higher scores on these tests than do Black and Hispanic students, thereby giving them a better chance of admission to the state's universities. If proportional representation by all ethnic groups is a valued objective for system administrators, then using test scores has negative value because it would tend to produce unequal admissions rates. On the other hand, legislators in Washington State wished to be able to reward schools and districts whose students showed higher than average achievement of the state's educational objectives, as measured by the state's assessment tests. Like the SAT, the Washington State tests also show different levels of achievement for different ethnic groups, but the high value placed on rewarding achievement outweighed the negative outcome of revealing ethnic differences. The two state education establishments reached contradictory conclusions about the use of standardized tests due to the different values each was trying to satisfy.

So far, we have considered practical decisions leading to action. Measurement is also important in providing information to guide theoretical decisions. In these cases, the desired result is not action but, instead, understanding. Do girls of a certain age read at a higher level than boys? A reading test is needed to obtain the information on which to base a decision. Do students who are anxious about tests perform less well on them than students who are not anxious? A questionnaire on "test anxiety" and a test of academic achievement could be used to obtain information helpful in reaching a decision on this issue. Even a question as basic as whether the size of reward a rat receives for running through a maze affects the rat's running speed requires that the investigator make measurements. Measurement is fundamental to answering nearly all the questions that science asks, not only in the physical sciences but also in the behavioral and biological sciences. The questions we choose to ask and how we ask them, however, are guided and limited by our assumptions about the nature of reality and the values we bring to the task.

STEPS IN THE MEASUREMENT PROCESS

In this book, we discuss the measurement of human abilities, interests, and personality traits. We need to pause for a moment to look at what is implied by measurement and what requirements must be met if we are legitimately to claim that a measurement has been made. We also need to ask how well the available techniques for measuring the human characteristics of interest do in fact meet these requirements.

Measurement in any field involves three common steps: (1) identifying and defining the quality or the attribute that is to be measured, (2) determining the set of operations by which the attribute may be isolated and displayed for observation, and (3) establishing a set of procedures or definitions for translating our observations into quantitative statements of degree or amount. An understanding of these steps, and of the difficulties that each presents, provides a sound foundation for understanding the procedures and problems of measurement in psychology and education.

Identifying and Defining the Attribute

We never measure a thing or a person. We always measure a **quality** or an **attribute** of the thing or person. We measure, for example, the *length* of a table, the *temperature* of a blast furnace, the *durability* of an automobile tire, the *flavor* of a soft drink, the *intelligence* of a schoolchild, or the

emotional maturity of an adolescent. Psychologists and educators frequently use the term **construct** to refer to the more abstract and difficult-to-observe properties of people, such as their intelligence or personality.

When we deal with simple physical attributes, such as length, it rarely occurs to us to wonder about the meaning or definition of the attribute. A clear meaning for length was established long ago in the history of both the species and the individual. The units for expressing length and the operations for making the property manifest have changed over the years (we no longer speak of palms or cubits); however, the underlying concepts have not. Although mastery of the concepts of *long* and *short* may represent significant accomplishments in the life of a preschool child, they are automatic in adult society. We all know what we mean by *length*.

The construct of length is one about which there is little disagreement, and the operations by which length can be determined are well known. However, this level of construct agreement and clarity of definition do not exist for all physical attributes. What do we mean by durability in an automobile tire? Do we mean resistance to wear and abrasion from contact with the road? Do we mean resistance to puncture by pointed objects? Do we mean resistance to deterioration or decay with the passage of time or exposure to sunlight? Or do we mean some combination of these three and possibly other factors? Until we can reach some agreement on what we mean by durability, we can make no progress toward measuring it. To the extent that we disagree on what durability means (i.e., on a definition of the construct), we will disagree on what procedures are appropriate for measuring it. If we use different procedures, we are likely to get different results from our measurements, and we will disagree on the value that we obtain to represent the durability of a particular brand of tire.

The problem of agreeing on what a given construct means is even more acute when we start to consider those attributes of concern to the psychologist or educator. What do we mean by *intelligence?* What kinds of behavior shall we characterize as intelligent? Shall we define the construct primarily in terms of dealing with ideas and abstract concepts? Or will it include dealing with things—with concrete objects? Will it refer primarily to behavior in novel situations? Or will it include responses in familiar and habitual settings? Will it refer to speed and fluency of response or to level of complexity of the response without regard to time? Will it include skill in social interactions? What kinds of products result from the exercise of intelligence: a theory about atomic structures, a ballet, or a snowman? We all have a general idea of what we mean when we characterize behavior as intelligent, but there are many specific points on which we may disagree as we try to make our definition sufficiently precise to allow measurement. This problem of precisely defining the attribute is present for all psychological constructs—more for some than for others. The first problem that psychologists or educators face as they try to measure attributes is arriving at clear, precise, and generally accepted definitions of those attributes.

Of course, we must answer another question even before we face the problem of defining the attribute. We must decide which attributes are relevant and important to measure if our description is to be useful. A description may fail to be useful for the need at hand because the chosen features are irrelevant. For example, in describing a painting, we might report its height, breadth, and weight with great precision and reach high agreement about the amount of each property the painting possesses. If our concern were to crate the picture for shipment, these might be just the items of information that we would need. However, if our purpose were to characterize the painting as a work of art, our description would be useless; the attributes of the painting we just described would be irrelevant to its quality as a work of art.

Similarly, a description of a person may be of little value for our purpose if we choose the wrong attributes to describe. A company selecting employees to become truck drivers might test

their verbal comprehension and ability to solve quantitative problems, getting very accurate measures of these functions. Information on these factors, however, is likely to be of little help in identifying people who have low accident records and would be steady and dependable on the job. Other factors, such as eye–hand coordination, depth perception, and freedom from uncontrolled aggressive impulses, might prove much more relevant to the tasks and pressures that a truck driver faces. The usefulness of a measuring procedure for its intended purpose is called its *validity*.

Consider a high school music teacher who thoroughly tested the pupils' knowledge of such facts as who wrote the "Emperor Concerto" and whether andante is faster than allegro. The teacher would obtain a dependable appraisal of the students' knowledge about music and musicians without presenting them with a single note of actual music, a single theme or melody, a single interpretation or appraisal of living music. As an appraisal of musical appreciation, such a test seems almost worthless because it uses bits of factual knowledge *about* music and composers in place of information that would indicate progress in appreciation of the music itself. One of the pitfalls to which psychologists and educators occasionally are prone is to elect to measure some attribute because it is easy to measure rather than because it provides the most relevant information for making the decision at hand. It is important to measure traits that are relevant to the decisions to be made rather than merely to measure traits that are easy to assess. We will discuss the issue of relevance again when we cover validity in Chapter 5.

Determining Operations to Isolate and Display the Attribute

The second step in developing a measurement procedure is finding or inventing a set of operations that will isolate the attribute of interest and display it. The operations for measuring the length of an object such as a table have been essentially unchanged for many centuries. We convey them to the child early in elementary school. The ruler, the meter stick, and the tape measure are uniformly accepted as appropriate instruments, and laying one of them along an object is an appropriate procedure for displaying to the eye the length of the table, desk, or other object. But the operations for measuring length, or distance, are not always that simple. By what operations do we measure the distance from New York to Chicago? From the earth to the sun? From our solar system to the giant spiral galaxy in Andromeda? How shall we measure the length of a tuberculosis bacillus or the diameter of a neutron? Physical science has progressed by developing both instruments that extend the capabilities of our senses and indirect procedures that make accessible to us amounts too great or too small for the simple, direct approach of laying a measuring stick along the object. Some operations for measuring length, or distance, have become indirect, elaborate, and increasingly precise. These less intuitive methods (such as the shift of certain wavelengths of light toward the red end of the spectrum) are accepted because they give results that are consistent, verifiable, and useful, but their relationship to the property of interest is far from obvious.

Returning to the example of the durability of the automobile tire, we can see that the operations for eliciting or displaying that attribute will depend on and interact with the definition that we have accepted for the construct. If our definition is in terms of resistance to abrasion, we need to develop some standard and uniform procedure for applying an abrasive force to a specimen and gauging the rate at which the rubber wears away, that is, a standardized simulated road test. If we have indicated puncture resistance as the central concept, we need a way of applying graduated puncturing forces. If our definition is in terms of resistance to deterioration from sun, oil, and other destructive agents, our procedure must expose the samples to these agents and provide

some index of the resulting loss of strength or resilience. A definition of durability that incorporates more than one aspect will require an assessment of each, with appropriate weight, and combine the aspects in an appropriate way; that is, if our definition of durability, for example, includes resistance to abrasion, punctures, and deterioration, then a measure of durability must assess all of these properties and combine them in some way to give a single index of durability. Many of the constructs we wish to measure in education and psychology are similar to our construct of durability in that the global construct includes a combination of more or less independent simpler constructs. How to combine a person's ability to answer questions that require reasoning with unfamiliar material, their short-term memory, their knowledge of culturally salient facts, and many other relatively narrowly defined constructs into a global construct of intelligence, or whether to combine them at all, is a hotly debated topic in both psychology and education.

The definition of an attribute and the operations for eliciting it interact. On the one hand, the definition we have set up determines what we will accept as relevant and reasonable operations. Conversely, the operations that we can devise to elicit or display the attribute constitute in a very practical sense the definition of that attribute. An attribute defined by how it is measured is said to have an **operational definition.** The set of procedures we are willing (or forced by our lack of ingenuity) to accept as showing the durability of an automobile tire become the operational definition of durability for us and may limit what we can say about it.

The history of psychological and educational measurement during the 20th century and continuing into the 21st century has largely been the history of the invention of instruments and procedures for eliciting, in a standard way and under uniform conditions, the behaviors that serve as indicators of the relevant attributes of people. The series of tasks devised by Binet and his successors constitute operations for eliciting behavior that is indicative of intelligence, and the Stanford-Binet and other tests have come to provide operational definitions of intelligence. The fact that there is no single, universally accepted test and that different tests vary somewhat in the tasks they include and the order in which they rank people on the trait are evidence that we do not have complete consensus on what intelligence is or on what the appropriate procedures are for eliciting it. This lack of consensus is generally characteristic of the state of the art in psychological and educational measurement. There is enough ambiguity in our definitions and enough variety in the instruments we have devised to elicit the relevant behaviors that different measures of what is alleged to be the same trait may rank people quite differently. This fact requires that we be very careful not to overgeneralize or overinterpret the results of our measurements.

The problem of developing an operational definition for the characteristic of interest is also present in the classroom. Teachers regularly face the problem of assessing student performance, but what we will call *performance* is closely linked with the way we assess it. Only to the extent that teachers agree on their operational definitions of student achievement will their assessments have comparable meaning. If one teacher emphasizes quick recall of facts in his assessments and another looks for application of principles in hers, they are, to some undetermined extent, evaluating different traits, and their assessments of their students' accomplishments will mean different things. Here, we do not suggest that this difference in emphasis is inappropriate, only that it is a fact of which educators should be aware. This variability in definitions also provides a major impetus for standardized achievement tests, because such tests are seen as providing a common definition of the attribute to be measured and the method of measurement. The definition provided by the test may not exactly represent what either teacher means by achievement in Subject X, but such definitions usually are developed to provide an adequate fit to the definitions most teachers espouse.

Quantifying the Attribute

The third step in the measurement process, once we have accepted a set of operations for eliciting an attribute, is to express the result of those operations in quantitative terms. Measurement has sometimes been defined as *assigning numbers to objects or people according to a set of rules.* The numbers represent how much of the attribute is present in the person or thing.

Using numbers has several advantages, two of which concern us. First, quantification makes communication more efficient and precise. We know much more from the statement that Ralph is 6 feet tall and weighs 175 pounds than we could learn from an attempt to describe Ralph's size in nonquantitative terms. We will see in Chapter 3 that much of the meaning that numbers have comes from the context in which the measurements are made (i.e., the rules used to guide the assignment), but, given that context, information in quantitative form is more compact, more easily understood, and generally more accurate than is the same information in other forms, such as verbal descriptions or photographs. In fact, we are so accustomed to communicating some types of information, such as temperature or age, in quantitative terms that we would have difficulty using another framework.

A second major advantage of quantification is that we can apply the power of mathematics to our observations to reveal broader levels of meaning. Consider, for example, trying to describe the performance of a class on a reading test or the accomplishments of a batter in baseball. In either case, we are accustomed to using the average as a summary of several individual performances. For many purposes, it is useful to be able to add, subtract, multiply, or divide to bring out the full meaning that a set of information may have. Some of the mathematical operations that are useful to educators and psychologists in summarizing their quantitative information are described in Chapter 2.

The critical initial step in quantification is to use a set of rules for assigning numbers that allows us to answer the question "How many?" or "How much?" The set of rules is called a **scale.** In the case of the length of a table the question becomes "How many inches?" or "How many meters?" The inch, or the meter, represents the basic unit, and the set of rules includes the measuring instrument itself and the act of laying the instrument along the object to be measured.

We can demonstrate that any inch equals any other inch by laying two inch-long objects side by side and seeing their equality. Such a demonstration is direct and straightforward proof of equality that is sufficient for some of the simplest physical measures. For measuring devices such as the thermometer, units are equal by *definition.* Thus, we define equal increases in temperature to correspond to equal amounts of expansion of a volume of mercury. One degree centigrade is *defined as* 1/100 of the difference between the freezing and boiling points of water. Long experience with this definition has shown it to be useful because it gives results that relate in an orderly and meaningful way to many other physical measures. (Beyond a certain point—the boiling point of mercury—this particular definition breaks down. However, other procedures that can be used outside this range can be shown to yield results equal to those of the mercury thermometer. The same principle allows educators to use a graded series of tests to assess student progress over several years.)

None of our psychological attributes have units whose equality can be demonstrated by direct comparison in the way that the equality of inches or pounds can. How will we demonstrate that arithmetic Problem X is equal, in amount of arithmetic ability that it represents, to arithmetic Problem Y? How can we show that one symptom of anxiety is equal to another anxiety indicator? For the qualities of concern to the psychologist or educator, we always have to fall back on a somewhat arbitrary definition to provide units and quantification. Most frequently, we

consider one task successfully completed—a word defined, an arithmetic problem solved, an analogy completed, or an attitude statement endorsed—equal to any other task in the series and then use the total number of successes or endorsements for an individual as the value representing the person on the particular attribute. This count of tasks successfully completed, or of choices of a certain type, provides a plausible and manageable definition of amount, but we have no adequate evidence of the equivalence of different test tasks or different questionnaire responses. By what right or evidence do we treat a number series item such as "1, 3, 6, 10, 15, ___?___, ___?___" as showing the same amount of intellectual ability as, for example, a verbal analogies item such as "Hot is to cold as wet is to ___?___?"

The definition of equivalent tasks and, consequently, of units for psychological tests is shaky at best. When we have to deal with a teacher's rating of a student's cooperativeness or a supervisor's evaluation of an employee's initiative, for example, where a set of categories such as "superior," "very good," "good," "satisfactory," and "unsatisfactory" is used, the meaningfulness of the units in which these ratings are expressed is even more suspect. The problem of arbitrary metrics in psychology has been discussed recently by Blanton and Jaccard (2006). We explore ways to report the results of measurements that yield approximately equal units in Chapter 3, but, as we shall see, even when units are equal they may only convey information about relative position, not absolute amount.

Problems Relating to the Measurement Process

In psychological and educational measurement, we encounter problems in relation to each of the three steps just described.

First, we have problems in selecting the attributes of concern and in defining them clearly, unequivocally, and in mutually agreeable terms. Even for something as straightforward as reading ability, we can get a range of interpretations. To what extent should a definition include each of the following abilities?

1. Reads quickly
2. Converts visual symbols to sounds
3. Obtains direct literal meanings from the text
4. Draws inferences that go beyond what is directly stated
5. Is aware of the author's bias or point of view.

As we deal with more complex and intangible concepts, such as cooperativeness, anxiety, adjustment, or rigidity, we may expect even more diversity in definition.

Second, we encounter problems in devising procedures to elicit the relevant attributes. For some attributes, we have been fairly successful in setting up operations that call on the individual to display the attribute and permit us to observe it under uniform and standardized conditions. We have had this success primarily in the domain of abilities, where standardized tests have been assembled in which the examinee is called on, for instance, to read with understanding, to perceive quantitative relationships, or to identify correct forms of expression. However, with many attributes we clearly have been less successful. By what standard operations can we elicit, in a form in which we can assess it, a potential employee's initiative, a client's anxiety, or a soldier's suitability for combat duty? With continued research and with more ingenuity, we may hope to devise improved operations for making certain of these attributes or qualities manifest, but identifying suitable measurement operations for many psychological attributes will remain a problem.