Testing and Individual Differences

CHAPTER 11

KEY TERMS

Standardized test
Norms
Standardization sample
Psychometrician
Reliability—split-half, test-retest, equivalent form
Validity—face, criterion-related (concurrent and predictive), construct
Aptitude test
Achievement test
Intelligence
Fluid intelligence
Crystallized intelligence
Spearman's intelligence theory
Gardner's intelligence theory
Sternberg's intelligence theory
Emotional intelligence
Stanford-Binet IQ test
Weschler tests (WAIS, WISC, WPPSI)
Normal distribution
Heritability
Flynn effect

OVERVIEW

We all take many standardized tests and receive scores that tell us how we perform. In this chapter, we will review what makes for a good test, how to interpret your scores on such tests, and what different kinds of tests exist. Then we will focus on one of the most tested characteristics of all, intelligence.

STANDARDIZATION AND NORMS

When we say that a test is *standardized*, we mean that the test items have been piloted on a similar population of people as those who are meant to take the test and that achievement *norms* have been established. For instance, consider the scholastic achievement test (SAT), a test with which many of you are probably all too familiar. When you take the SAT, you take an experimental section, a group of questions on which you will not be evaluated. In this case, you are helping the Educational Testing Service (ETS) to standardize its future examinations. Those people taking the SAT on a particular testing date are fairly representative of the population of people taking the SAT in general. Such a group of people is known

as the *standardization sample.* The *psychometricians* (people who make tests) at ETS use the performance of the standardization sample on the experimental sections to choose items for future tests.

The purpose of tests is to distinguish between people. Therefore, test questions that virtually everyone answers correctly as well as questions that almost no one can answer are discarded. Such items do not provide information that differentiates between the people taking the test. As you are probably aware, questions on the SAT are arranged, within a given section, in order of difficulty. The difficulty level of the questions has been predetermined by the performance of the standardization sample. Ideally, this process of standardization yields equivalent exams, allowing a fair comparison between one person's score on the November 2009 SAT with another's on the May 2010 SAT.

RELIABILITY AND VALIDITY

In order for us to have any faith in the meaning of a test score, we must believe the test is both reliable and valid. *Reliability* refers to the repeatability or consistency of the test as a means of measurement. For instance, if you were to take a test three times that purportedly determined what career you should pursue, and on each occasion you received radically different recommendations, you might question the reliability of the test. Similarly, if you scored 115, 92, and 133 on three different administrations of the same IQ (intelligence quotient) test, you would have little reason to believe your intelligence had been accurately measured.

The reliability of a test can be measured in several different ways. *Split-half reliability* involves randomly dividing a test into two different sections and then correlating people's performances on the two halves. The closer the correlation coefficient is to +1, the greater the split-half reliability of the test. Many tests are available in several equivalent forms. The correlation between performance on the different forms of the test is known as *equivalent-form reliability.* Finally, *test-retest reliability* refers to the correlation between a person's score on one administration of the test with the same person's score on a subsequent administration of the test.

A test is *valid* when it measures what it is supposed to measure. Validity is often referred to as the accuracy of a test. A personality test is valid if it truly measures an individual's personality, and the career inventory described above is valid only if it actually measures for what jobs a person is best suited. The latter example should serve to highlight an important point: a test cannot be valid if it is not reliable. If subsequent administrations of the career inventory yield grossly disparate results for the same person, it clearly does not accurately reflect a person's vocational strengths or interests. However, a test may be reliable without being valid. Even if someone's performance on the test repeatedly indicates that he or she should be a chef and thus is reliable, if the person hates to cook, the test is not a valid measure of his or her interest.

Just as several different kinds of reliability exist, a number of different kinds of validity exist. *Face validity* refers to a superficial measure of accuracy. A test of cake-baking ability has high face validity if you are looking for a chef but low face validity if you are in the market for a doctor. Face validity is a type of *content validity.*

Content validity refers to how well a measure reflects the entire range of material it is supposed to be testing. If one really wanted to design a test to find a good chef, a test that required someone to create an entrée and whip up a salad dressing in addition to baking a cake would have greater content validity.

Another kind of validity is *criterion-related validity.* Tests may have two kinds of criterion-related validity, concurrent and predictive. *Concurrent validity* measures how much of a characteristic a person has now; is a person a good chef now? *Predictive validity* is a measure of future performance; does a person have the qualities that would enable him or her to become a good chef?

Finally, *construct validity* is thought to be the most meaningful kind of validity. If an independent measure already exists that has been established to identify those who will make fine chefs and love their work, we can correlate prospective chefs' performance on this measure with their performance on any new measure. The higher the correlation, the more construct validity the new measure has. The limitation, of course, is the difficulty in creating any measure that we believe is perfectly valid in the first place.

HINT

Reliability and validity are important terms for you to know. The psychological meaning ascribed to these two terms may differ somewhat from how they are used by the general population. Reliability refers to a test's consistency, and validity refers to a test's accuracy.

TYPES OF TESTS

Two common types of tests are aptitude tests and achievement tests. *Aptitude tests* measure ability or potential, while *achievement tests* measure what one has learned or accomplished. For instance, any intelligence test is supposed to be an aptitude test. These tests are made to express someone's potential, not his or her current level of achievement. Conversely, most, if not all, the tests you take in school are supposed to be achievement tests. They are supposed to indicate how much you have learned in a given subject. However, making a test that exclusively measures one of these qualities is virtually impossible. Whatever one's aptitude for a particular field or skill, one's experience affects it. Someone who has had a lot of schooling will score better on a test of mathematics aptitude than someone who might have an equally great potential to be a mathematician but who has never had any formal training in math. Similarly, two people who have achieved equally in learning biology will not necessarily score the same on an achievement test. If one has far greater test-taking aptitude, she or he will likely outscore the other.

Distinguishing between speed and power tests is also possible. *Speed tests* generally consist of a large number of questions asked in a short amount of time. The goal of a speed test is to see how quickly a person can solve problems. Therefore, the amount of time allotted should be insufficient to complete the problems. The goal of a *power test* is to gauge the difficulty level of problems an individual can solve. Power tests consist of items of increasing difficulty levels. Examinees are given sufficient time to work through as

HINT

Even though it is essentially impossible to create a pure aptitude or pure achievement test, tests that purport to measure aptitude seek to measure someone's ability or potential, whereas achievement tests seek to measure how much of a body of material someone has learned.

many problems as they can since the goal is to determine the ceiling difficulty level, not their problem-solving speed.

Finally, some tests are *group tests* while others are *individual tests*. Group tests are administered to a large number of people at a time. Interaction between the examiner and the people taking the test is minimal. Generally, instructions are provided to the group, and then people are given a certain amount of time to complete the various sections of the test. Group tests are less expensive to administer and are thought to be more objective than individual tests. Individual tests involve greater interaction between the examiner and examinee. Several of the IQ tests that will be discussed later in this chapter are individual tests. The Rorschach inkblot test, discussed in the personality chapter, is also an individual test. The examiner attends not only to what the person says about the inkblots but also to the process by which he or she analyzes the stimuli.

THEORIES OF INTELLIGENCE

While *intelligence* is a commonly used term, it is an extremely difficult concept to define. Typically, intelligence is defined as the ability to gather and use information in productive ways. However, we will not present any one correct definition of intelligence because nothing that approaches a consensus has been achieved. Rather, we will present brief summaries of some of the most widely known theories of intelligence.

Many psychologists differentiate between *fluid intelligence* and *crystallized intelligence*. Fluid intelligence refers to our ability to solve abstract problems and pick up new information and skills, while crystallized intelligence involves using knowledge accumulated over time. While fluid intelligence seems to decrease as adults age, research shows that crystallized intelligence holds steady or may even increase. For instance, a 20-year-old may be able to learn a computer language more quickly than a 60-year-old, whereas the older person may well have the advantage on a vocabulary test or an exercise dependent upon wisdom.

Charles Spearman

One fundamental issue of debate is whether intelligence refers to a single ability, a small group of abilities, or a wide variety of abilities. Spearman argued that intelligence could be expressed by a single factor. He used factor analysis, a statistical technique that measures the correlations between different items, to conclude that underlying the many different specific abilities *s* that people regard as types of intelligence is a single factor that he named *g* for general.

L. L. Thurstone and J. P. Guilford

Thurstone's primary mental abilities theory states that intelligence is comprised of seven main abilities including reasoning, verbal comprehension, and memory. Guilford, on the other hand, posited the existence of well over 100 different mental abilities.

Howard Gardner

Gardner also subscribes to the idea of *multiple intelligences.* Unlike many other researchers, however, the kinds of intelligences that this contemporary researcher has named thus far encompass a large range of human behavior. Three of Gardner's multiple intelligences—linguistic, logical-mathematical, and spatial—fall within the bounds of qualities traditionally labeled as intelligences. To that list Gardner has added musical, bodily-kinesthetic, intrapersonal, interpersonal, and naturalist intelligence. He is working on naming others. Musical intelligence, as one might suspect, includes the ability to play an instrument or compose a symphony. A dancer or athlete would have a lot of bodily-kinesthetic intelligence as would a hunter. Intrapersonal intelligence refers to one's ability to understand oneself. People who are able to persevere without becoming discouraged or who can differentiate between situations in which they will be successful and those that may simply frustrate them have intrapersonal intelligence. Interpersonal intelligence, on the other hand, corresponds to a person's ability to get along with and be sensitive to others. Successful psychologists, teachers, and salespeople would have a lot of interpersonal intelligence. Finally, naturalist intelligence is found in people gifted at recognizing and organizing the things they encounter in the natural environment. Such people would be successful in fields such as biology and ecology.

Daniel Goleman

Recently there has been a lot of discussion of *EQ,* which is also known as *emotional intelligence.* One of the main proponents of EQ is Goleman. EQ roughly corresponds to Gardner's notions of interpersonal and intrapersonal intelligence. Researchers who argue for the importance of EQ point out that the people with the highest IQs are not always the most successful people. They contend that both EQ and IQ are needed to succeed.

Robert Sternberg

Sternberg is another contemporary researcher who has offered a somewhat nontraditional definition of intelligence. *Sternberg's triarchic theory* holds that three types of intelligence exist. Componential or analytic intelligence involves the skills traditionally thought of as reflecting intelligence. Most of what we are asked to do in school involves this type of intelligence: the ability to compare and contrast, explain, and analyze. The second type, experiential or creative intelligence, focuses on people's ability to use their knowledge and experiences in new and innovative ways. Rather than comparing the different definitions of intelligence that others have offered, someone with this type of intelligence might prefer to come up with his or her own theory of what constitutes intelligence. The third kind of intelligence Sternberg discusses is contextual or practical intelligence. People with this type of intelligence are what we consider street-smart, they are able to apply what they know to real-world situations.

This last aspect of Sternberg's theory, the idea of practical intelligence, raises another important and unresolved issue in the study of intelligence: does intelli-

TABLE 11.1

Theories of Intelligence

Spearman	Intelligence can be measured by a single, general ability (*g*)
Gardner	Theory of multiple intelligences—the term "intelligence" should be applied to a wide variety of abilities including kinesthetic, musical, interpersonal, intrapersonal, naturalistic, verbal, spatial, and mathematical
Sternberg	Triarchic theory of intelligence—people can be intelligent in different ways; they can evidence analytic, practical, and creative intelligence

gence depend upon context? The other theories of intelligence discussed above essentially posit that intelligence is an ability, some thing or collection of things that one has or does not have. Sternberg, on the other hand, asserts that what is intelligent behavior depends on the context or situation in which it occurs. If intelligence does, indeed, depend upon context, devising an intelligence test becomes a particularly difficult task. The most common intelligence tests used (described in the next section) are based on the view of intelligence as ability based.

INTELLIGENCE TESTS

Not surprisingly, the ongoing debate over what constitutes intelligence makes constructing an assessment particularly difficult. Two widely used individual tests of intelligence are the Stanford-Binet and the Weschler.

Alfred Binet was a Frenchman who wanted to design a test that would identify which children needed special attention in schools. His purpose was not to rank or track children but, rather, to improve the children's education by finding a way to tailor it better to their specific needs. Binet came up with the concept of *mental age*, an idea that presupposes that intelligence increases as one gets older. The average 10-year-old child has a mental age of 10. When this average child grows to age 12, she or he will seem more intelligent and will have a mental age of 12. By using this method, Binet created a test that would identify children who lagged behind most of their peers, were in step with their peer group, and were ahead of their peers. Binet created a standardized test using the method described earlier in this chapter. He administered questions to a standardization sample and constructed a test that would differentiate between children functioning at different levels.

Louis Terman, a Stanford professor, used this system to create the measure we know as IQ and the test known as the *Stanford-Binet IQ* test. *IQ* stands for intelligence quotient. A person's IQ score on this test is computed by dividing the person's mental age by his or her chronological age and multiplying by 100. Thus, the child described above has an IQ of 100 because $10/10 \times 100 = 100$. A child who has a mental age of 15 at age 10 would have an IQ of 150, $15/10 \times 100 = 150$. A commonly asked question about this system is how it deals with adults. While

talking about a mental age of 8 or 11 or 17 makes sense, what does having a mental age of 25 or 33 or 58 mean? To address this problem, Terman assigned all adults an arbitrary age of 20.

David Weschler used a different way to measure intelligence. Although it does not involve finding a quotient, it is still known as an IQ test. Three different Weschler tests actually exist. The *Weschler adult intelligence scale* (*WAIS*) is used in testing adults, the *Weschler intelligence scale for children* (*WISC*) is given to children between the ages of six and 16, and the *Weschler preschool and primary scale of intelligence* (*WPPSI*) can be administered to children as young as four. The Weschler tests yield IQ scores based on what is known as *deviation IQ*. The tests are standardized so that the mean is 100, the standard deviation is 15, and the scores form a normal distribution. Remember that in a normal distribution, the percentages of scores that fall under each part of the normal curve are predetermined (see Fig. 11.1).

For instance, approximately 68 percent of scores fall within one standard deviation of the mean, approximately 95 percent fall within two standard deviations of the mean, and 98 to 99 percent of scores fall within three standard deviations of the mean. People's scores are determined by how many standard deviations they fall away from the mean. Thus, Peter who scores at the 15.87th percentile falls at one standard deviation below the mean and is assigned a score of 85, while Juanita who scores at the 97.72nd percentile has scored two standard deviations above the mean and has scored 130. Of course, most people do not fall exactly one or two

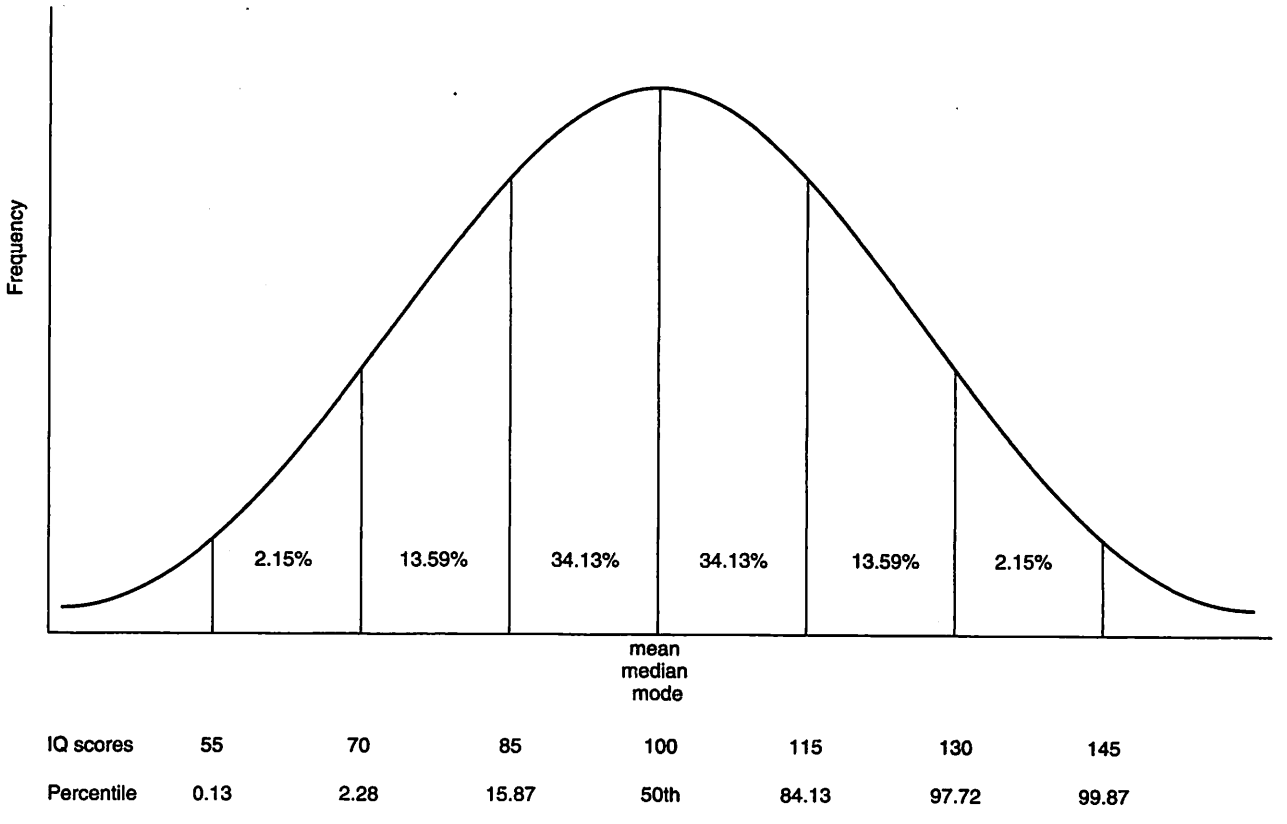

Figure 11.1. The normal distribution on an IQ test.

standard deviations above the mean. However, using such an example would necessitate less obvious mathematical calculations. For more information on the normal curve, you might want to refer to the "Statistics" section in Chapter 2.

Whereas the Stanford-Binet IQ test utilizes a variety of different kinds of questions to yield a single IQ score, the Weschler tests result in scores on a number of subscales as well as a total IQ score. For instance, the WAIS has 11 subscales. Six of them are combined to produce a verbal IQ score. Five are used to indicate performance IQ. The kinds of questions used to measure verbal IQ ask people to define words, solve mathematical word problems, and explain ways in which different items are similar. The items on the performance section involve tasks like duplicating a pattern with blocks, correctly ordering pictures so they tell a story, and identifying missing elements in pictures. Differences between a person's score on the verbal and performance sections of this exam can be used to identify learning disabilities.

BIAS IN TESTING

Much discussion has centered on whether widely used IQ tests or the SAT are biased against certain groups. Interestingly, researchers seem to agree that although different races and sexes may score differently on these tests, the tests have the same predictive validity for all groups. In other words, SAT scores are equally good predictors of college grades for both sexes and for different racial groups and thus, in a sense, the test is clearly not biased. However, other researchers have argued that both the tests and the college grades are biased in a far more fundamental way. Advantages seem to accrue to the white, middle and upper class. The experiences of other cultural groups seem to work to their detriment both on these tests and in college. Members of these groups may not have been exposed to the vocabulary and range of experiences that the writers of the test assume they have or believe they should have been. To the extent that the tests are supposed to identify academic potential, they may then be both flawed and biased.

NATURE VERSUS NURTURE: INTELLIGENCE

One of the most difficult and controversial issues in psychology involves sorting out the relative effects of nature and nurture. Keep in mind that nature refers to the influence of genetics, while nurture stresses the importance of the environment and learning. One of the more hotly contested aspects of the nature-nurture debate is intelligence. Human intelligence is clearly affected by both nature and nurture. Research suggests that both genetic and environmental factors play a role in molding intelligence.

An important term that researchers use in discussing the effects of nature and nurture is *heritability*. Heritability is a measure of how much of a trait's variation is explained by genetic factors. Heritability can range from 0 to 1, where 0 indicates that the environment is totally responsible for differences in the trait and 1 means that all of the variation in the trait can be accounted for genetically. Thus, the question is how heritable is intelligence? That heritability does not apply to an individual but rather to a population is important to point out. Whatever the heri-

tability ratio for intelligence, it will not tell us how much of any particular person's intelligence was determined by nature or nurture.

Solving this controversy once and for all is essentially impossible because we cannot ethically set up the kind of controlled experiment necessary to provide definitive answers to this question. However, many researchers have studied this issue, and some of their findings are presented below:

- Performance on intelligence tests has been increasing steadily throughout the century, a finding known as the *Flynn effect.* Since the gene pool has remained relatively stable, this finding suggests that environmental factors such as nutrition, education, and, perhaps, television and video games play a role in intelligence.
- Monozygotic (identical) twins, who share 100 percent of their genetic material, score much more similarly on intelligence tests than do dizygotic (fraternal) twins, who have, on average, only 50 percent of their genes in common. Nonetheless, some researchers have suggested that monozygotic twins tend to be treated more similarly than dizygotic twins, thus confounding the effects of nature with those of nurture.
- Research on identical twins separated at birth has found strong correlations in intelligence scores. However, researchers advocating more of an environmental influence point out that usually the twins are placed into similar environments, again making it difficult to sift out the relative effects of nature and nurture. For instance, if each of the twins is placed into a white, middle-class, suburban home, concluding that all their similarities are genetically based does not make sense.
- Some researchers have argued that racial differences in IQ scores provide evidence that intelligence is largely genetically determined. The majority of psychologists disagree, arguing that these racial differences are more likely explained by differences in environments, particularly by socioeconomic factors. For example, African Americans, as a group, tend to score 10–15 points lower on IQ tests than do whites. Many researchers argue that the greater poverty level in many minority populations, an environmental factor, is the main cause of the disparity in test scores and not a difference in genetics. Test bias is an additional factor that may contribute to the gap in test scores.
- Participation in government programs such as Head Start, meant to redress some of the disadvantages faced by impoverished groups, has been shown to correlate with higher scores on intelligence tests. However, opponents of such programs assert that these gains are limited and of short duration. Advocates of such interventions respond that expecting the gains to outlast the programs is unreasonable.

After putting the issue of cause aside, when comparing groups of people on any characteristic, keep in mind that differences within groups generally dwarf differences between groups. In other words, within any one group will be more diversity than between any two groups. Practically speaking, if we find that boys perform better on a certain test than girls do, more of a difference will exist between the

HINT

Within-group differences are typically larger than between-group differences.

highest scoring boy and the lowest scoring boy than between the average boy and the average girl. Furthermore, knowing that boys generally outperform girls on this test tells us nothing about the performance of any particular girl compared with the performance of any particular boy. Therefore, we need to be careful about how we use information about differences between groups. Essentially, we should not use it. We should ignore it and evaluate each person, regardless of group membership, as an individual.

A CAUTIONARY NOTE

It is often said that we live in a testing society. We like to be able to measure things and assign them a number. Therefore, keeping in mind the limitations and extraordinary labeling power of these instruments is particularly important. As we have discussed, the definition of intelligence (and many other concepts) remains hotly debated and many factors affect people's performances on tests. Thus, we need to take care not to ascribe too great a meaning to a test score. Many schools that used to measure all their students' IQs periodically have abandoned that practice. Schools that used to base admission to programs for exceptional children solely on these tests now frequently gather information in other ways as well. When IQ tests are given, the results remain confidential so as not to create expectations about how people *ought* to perform (see the information on self-fulfilling prophecy in Chapter 14). While well-designed tests can be extremely useful, we must recognize their limitations.

Practice Questions

Directions: Each of the questions or incomplete statements below is followed by five suggested answers or completions. Select the one that is best in each case.

1. Paul takes a test in the army to see if he would make a good pilot. Such a test is

   (A) a standardized test.
   (B) an aptitude test.
   (C) an intelligence test.
   (D) an achievement test.
   (E) a biased test.

2. If a test is reliable, it means that

   (A) it is given in the same way every time.
   (B) it tests what it is supposed to test.
   (C) it is a fair assessment.
   (D) it yields consistent results.
   (E) it is also valid.