CHAPTER 15

Testing and Individual Differences

IN THIS CHAPTER

Summary: Are you taking the AP Psychology exam in May? Have you taken the SAT or ACT? These are all standardized tests. You've already taken lots of tests in your lifetime, and you will likely take many more, but all tests are not created equal. Some tests are better than others at predicting or evaluating your potential, or measuring your achievement. Tests are important to you because they are used to make decisions that affect your life

This chapter focuses on test quality and qualities of tests, ethics in testing, intelligence and intelligence testing, and the interactions of heredity and environment on intelligence.

KEY IDEA

Key Ideas

✪ Standardization and norms
✪ Reliability and validity
✪ Types of tests
✪ Ethics and standards in testing
✪ Intelligence
✪ Intelligence testing
✪ Kinds of intelligence
✪ Heredity/environment and intelligence
✪ Human diversity

Standardization and Norms

Psychometrics is the measurement of mental traits, abilities, and processes. **Psychometricians** are involved in test development in order to measure some *construct* or behavior that distinguishes people. Constructs are ideas that help summarize a group of related phenomena or objects; they are hypothetical abstractions related to behavior and defined by groups of objects or events. For example, we can't measure happiness, honesty, or intelligence in feet or meters. If someone tells the truth in a wide variety of situations, however, we might consider that person honest. Although we cannot observe happiness, honesty, or intelligence directly, they are useful concepts for understanding, describing, and predicting behavior. Psychological tests include tests of abilities, interests, creativity, personality, and intelligence. A good test is standardized, reliable, and valid. After many questions for a test have been written, edited, and pretested, questions are thrown out if nearly everyone answers them correctly or if very few answer them right because these types of questions do not tell us anything about individual differences. Tests that differentiate among test takers and that are composed of questions that fairly test all aspects of the behavior to be assessed are assembled. They are then administered to a sample of hundreds or thousands of people who fairly represent all of the people who are likely to take the test. This sample is used to standardize the test. **Standardization** is a two-part test development procedure that first establishes test norms from the test results of the large representative sample that initially took the test and then ensures that the test is both administered and scored uniformly for all test takers. **Norms** are scores established from the test results of the representative sample, which are then used as a standard for assessing the performances of subsequent test takers; more simply, norms are standards used to compare scores of test takers. For example, the mean score for the SAT is 500 and the standard deviation is 100, whereas the mean score for the Wechsler Adult Intelligence Scale (IQ test) is 100 and the standard deviation is 15, based on the "standardization" sample. When administering a standardized test, all proctors must give the same directions and time limits and provide the same conditions as all other proctors. All scorers must use the same scoring system, applying the same standards to rate responses as all other scorers. Thus, we should earn the same test score no matter where we take the test or who scores it.

Reliability and Validity

Not only must a good test be standardized, but it must also be reliable and valid.

Reliability

If a test is reliable, we should obtain the same score no matter where, when, or how many times we take it (if other variables remain the same). Several methods are used to determine if a test is reliable. In the *test-retest* method, the same exam is administered to the same group on two different occasions, and the scores compared. The closer the correlation coefficient is to 1.0, the more reliable the test. The problem with this method of determining reliability or consistency is that performance on the second test may be better because test takers are already familiar with the questions and test procedures. In the *split-half* method, the score on one half of the test questions is correlated with the score on the other half of the questions to see if they are consistent. One way to do that might be to compare the score of all the odd-numbered questions to the score of all the even-numbered questions. In the *alternate form method* or *equivalent form method,* two different versions of a test on the same material are given to the same test takers, and the scores are correlated. The SAT given on Saturday is different from the SAT given on Sunday in October; there are different questions on each form. Although this does not happen, if the same people took both exams and the tests were highly reliable, the scores should be the same on both tests. This would also necessitate high *interrater reliability,* the extent to which two or more scorers evaluate the responses in the same way.

Validity

Tests can be very reliable, but if they are not also valid, they are useless for measuring the particular construct or behavior. Psychometricians must present data to show that a test measures what it is supposed to measure accurately and that the results can be used to make accurate decisions. Because there are no universal standards against which test scores can be compared, validation is most frequently accomplished by obtaining high correlations between the test and other assessments. **Validity** is the extent to which an instrument accurately measures or predicts what it is supposed to measure or predict. Just as there are several methods for measuring reliability, there are also several methods for measuring validity.

- **Face validity** is a measure of the extent to which the content of the test measures all of the knowledge or skills that are supposed to be included within the domain being tested, according to the test takers. For example, we expect the AP Psychology exam to ask between five and seven questions dealing with testing and individual differences on the multiple-choice section of the test, as defined by the content outline for the course, which sets the structure and boundaries for the content domain.
- **Content validity** is a measure of the extent to which the content of the test measures all of the knowledge or skills that are supposed to be included within the domain being tested, according to expert judges.
- **Criterion related validity** is a measure of the extent to which a test's results correlate with other accepted measures of what is being tested.
- **Predictive validity** is a measure of the extent to which the test accurately forecasts a specific future result. For example, the SAT is designed to predict how well someone will succeed in his or her freshman year in college. High scores on the SAT should predict high grades for the first year in college.
- **Construct validity,** which some psychologists consider the true measure of validity, is the extent to which the test actually measures the hypothetical construct or behavior it is designed to assess. The MMPI-2 (described in Chapter 14) has a clinical trial set of questions for schizophrenia. This test has construct validity if this subset of questions successfully discriminates people with schizophrenia from other subjects taking the MMPI-2. Many people question whether intelligence tests have construct validity for measuring intelligence.

Types of Tests

Ask different psychometricians to categorize types of tests, and they may give different answers, because tests can be categorized along many dimensions.

Performance, Observational, and Self-Report Tests

Psychological tests can be sorted into the three categories of performance tests, observational tests, and self-report tests. For a performance test, the test taker knows what he or she should do in response to questions or tasks on the test, and it is assumed that the test taker will do the best he or she can to succeed. Performance tests include the SATs, AP tests, **Wechsler intelligence tests, Stanford-Binet intelligence tests,** and most classroom tests, including finals, as well as computer tests and road tests for a driver's license. Observational tests differ from performance tests in that the person being tested does not have a single, well-defined task to perform, but rather is assessed on typical behavior or performance in a specific context. Employment interviews and formal on-the-job observations for evaluation by supervisors are examples of observational tests. Self-report tests require the test taker to describe his or her feelings, attitudes, beliefs, values, opinions, physical state, or mental state on surveys, questionnaires, or polls. The **MMPI-2** (described in Chapter 14) exemplifies the self-report test.

Performance tests in which there is a correct answer for each item can be divided into two types, *speed tests* and *power tests. Speed tests* generally include a large number of relatively easy items administered with strict time limits under which most test takers find it impossible to answer all questions. Given more time, many test takers would probably score higher, so differences in scores among test takers are at least partly a function of the speed with which they respond. This differs from power tests, which allot enough time for test takers to complete the items of varying difficulty on the test, so that differences in scores among test takers are a function of the test taker's knowledge, and possibly good guessing.

Ability, Interest, and Personality Tests

Another way tests can be categorized is into ability, interest, and personality tests, which are relevant to decision making. General mental ability is particularly important in scholastic performance and in performing cognitively demanding tasks. Interests influence a person's reactions to and satisfaction with his or her situation. Personality involves consistency in behavior over a wide range of situations. (For personality tests, see Chapter 14.) Ability tests include **aptitude tests** designed to predict a person's future performance or to assess the person's capacity to learn, and **achievement tests** are designed to assess what a person has already learned. For example, the SAT is designed to measure potential to do well in college, whereas the AP Psychology test is designed to measure your mastery of the material in this course of study. Interest tests use a person's descriptions of his or her own interests to predict vocational adjustment and satisfaction. For example, the current version of the Strong-Campbell Interest Inventory, which is the most widely used vocational interest test, is based on the assumptions that responses that are similar to a particular occupational group and different from people in general provide key information about occupational interests, and that interests can be measured.

Group vs. Individual Tests

Also, there are group tests and individual tests. Standardized tests that can be administered in groups are much more widely used than individual tests administered to one person by a trained professional. Whereas group tests require a test taker to work alone on a structured task and respond to questions, individual tests require social interaction between the examiner and test taker, and require test takers to respond to a person. The test taker needs to view the examiner as trustworthy, competent, and nonjudgmental. The psychologist or other trained professional must use sound professional judgment in eliciting and scoring responses to test items. The differing roles of examiners in individual versus group tests can significantly affect the responses of test takers. Group tests are better standardized and more efficient than individual tests, but individual tests provide more information on test behavior, can be given to test takers who cannot sit for group tests, and can sometimes elicit more creative responses. The most popular individual intelligence test, the Wechsler Adult Intelligence Scale-III and the Stanford-Binet Intelligence Scales, exemplify individual exams. Examples of group tests are the widely used Armed Services Vocational Aptitude Battery (ASVAB) employed by the military to screen recruits and assign them to various jobs, training programs, and career paths; and the SAT and ACT (American College Test).

Ethics and Standards in Testing

Because of the potential for abuse, ethical standards guide the development and application of tests. Numerous professional organizations, including the American Psychological

Association, have produced documents detailing appropriate technical and professional standards for construction, evaluation, interpretation, and application of psychological tests to promote the welfare and best interests of the client, guard against the misuse of assessment results, respect the client's right to know the results, and safeguard the dignity of test takers. Psychologists need to obtain informed consent and guarantee confidentiality in personnel testing, for example. Tests should be used for the purpose for which they were designed by professionals trained in their use.

Because some groups (such as African Americans) have tended to score lower on average than other groups (such as European Americans) on intelligence tests and SATs, critics argue that such tests are biased. Since these tests predict school achievement of all races equally well, the major tests are not biased with respect to predictive validity. However, they do seem biased with respect to performance differences resulting from cultural experience. Biologically oriented theorist Arthur Jensen attempted to succeed where Galton failed in developing a culture-free measure of intelligence by measuring reaction time, but his test is inadequate to represent a measure of intelligence. Several attempts at creating culture-reduced tests that measure general intelligence, such as Raven's Progressive Matrices, have not succeeded in eliminating the difference in mean scores. *Culture relevant tests* that incorporate skills and knowledge related to the cultural experiences of the test takers may be more successful.

Intelligence and Intelligence Testing

Since intelligence is a construct, it can only be defined by the behaviors that indicate intelligence, such as the ability to learn from experience, solve problems, use information to adapt to the environment, and benefit from training. Because intelligence tests are common and have been used so widely, they have influenced the definition of intelligence; sometimes a score is used to define someone's intelligence. Intelligence is sometimes reified. *Reification* occurs when a construct is treated as though it were a concrete, tangible object. Intelligence test developer David Wechsler said, "**Intelligence,** operationally defined, is the aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment."

Francis Galton's Measurement of Psychophysical Performance

Modern ability testing originated with Charles Darwin's cousin, nativist **Francis Galton**, who measured psychomotor tasks to gauge intelligence, reasoning that people with excellent physical abilities are better adapted for survival, and thus highly intelligent. **James McKeen Cattell** brought Galton's studies to the United States, measuring strength, reaction time, sensitivity to pain, and weight discrimination, using the term "mental test." Although Galton and Cattell's measurements correlated poorly with reasoning ability, they drew attention to the systematic study of measuring cognitive and behavioral differences among individuals. At about the same time, French psychologist **Alfred Binet** was hired by the French government to identify children who would not benefit from a traditional school setting and those who would benefit from special education. He thought intelligence could be measured by sampling performance of tasks that involved memory, comprehension, and judgment. He collaborated with **Theodore Simon** to create the Binet-Simon scale, which he meant to be used only for class placement.

Alfred Binet's Measurement of Judgment

Binet thought that as we age, we become more sophisticated in the ways we know about the world and that, therefore, most 6-year-olds answer questions differently from 8-year-olds.

As a result of their responses to test items, children were assigned a *mental age* or *mental level* reflecting the age at which typical children give those same responses. Although mental age differentiates between abilities of children, it can be misleading when a 6-year-old and an 8-year-old, for example, have mental ages 2 years below their actual (chronological) ages. The younger child would be proportionally further behind peers than the older child. German psychologist **William Stern** suggested using the ratio of mental age (MA) to chronological age (CA) to determine the child's level of intelligence.

Mental Age and the Intelligence Quotient

In adapting Binet's test for Americans, **Lewis Terman** developed the Stanford-Binet Intelligence Scale reporting results as an IQ, **intelligence quotient,** which is the child's mental age divided by his or her chronological age, multiplied by 100; or MA/CA × 100. A 10-year-old who answers questions typical of most 12-year-olds has an IQ score of 120. Another 10-year-old who answers questions typical of an 8-year-old scores 80. With the development of intelligence tests for adults, the ratio IQ becomes meaningless and has been replaced by the deviation IQ determined as a result of the standardizing process for a particular test. For the fifth edition of the Stanford-Binet Intelligence Scale for Adults, the test has been standardized with a representative sample of test takers up to age 90. Fluid reasoning, visual-spatial processing, working memory, and quantitative reasoning seem to peak in the 30s, whereas knowledge seems to peak in the 50s.

The newest version assesses each of five ability areas, such as knowledge, fluid reasoning, and quantitative reasoning, both nonverbally and verbally. By combining these subtest scores, one IQ score is determined.

The Wechsler Intelligence Scales

David Wechsler developed another set of age-based intelligence tests: the *Wechsler Preschool and Primary Scale of Intelligence* (*WPPSI*) for preschool children, the *Wechsler Intelligence Scale for Children* (*WISC*) for ages 6 to 16, and the *Wechsler Adult Intelligence Scale* (*WAIS*) for older adolescents and adults. The latest edition, the WAIS-III, has a verbal scale including items on comprehension, vocabulary, information, similarities, arithmetic, and digit span; and a performance scale including items dealing with object assembly, block design, picture completion, picture arrangement, and digit symbols. Wechsler based his measures on deviation IQs or how spread out the scores were from the mean of 100 (Figure 15.1). Since intelligence has a bell curve distribution, 68 percent of the population will have an IQ between 85 and 115. These test takers are considered to be low normal through high normal. Test takers who fall two deviations below the mean have a score of 70, typically considered the borderline for **intellectual disability,** while test takers two standard deviations above the mean have scores of 130, sometimes considered intellectually gifted, and those three standard deviations above the mean have scores of 145, sometimes considered geniuses. The Wechsler tests are judged more helpful for determining the extremes of intelligence at the intellectual disability and the genius level than the Stanford-Binet. They also help indicate possible learning disabilities when a child's performance IQ is very different from his or her verbal score.

Intellectual Disability

Over the past two decades, the term *mental retardation* has been replaced by **intellectual disability** (intellectual developmental disorder). To be considered intellectually disabled, an individual must earn a score at or below 70 on an IQ test and also show difficulty adapting in everyday life. Adaptive behavior is expressed in conceptual skills, social skills, and practical skills. Severity is determined by adaptive functioning rather than IQ score. Typically individuals with mild intellectual disability (about 85 percent) can care for

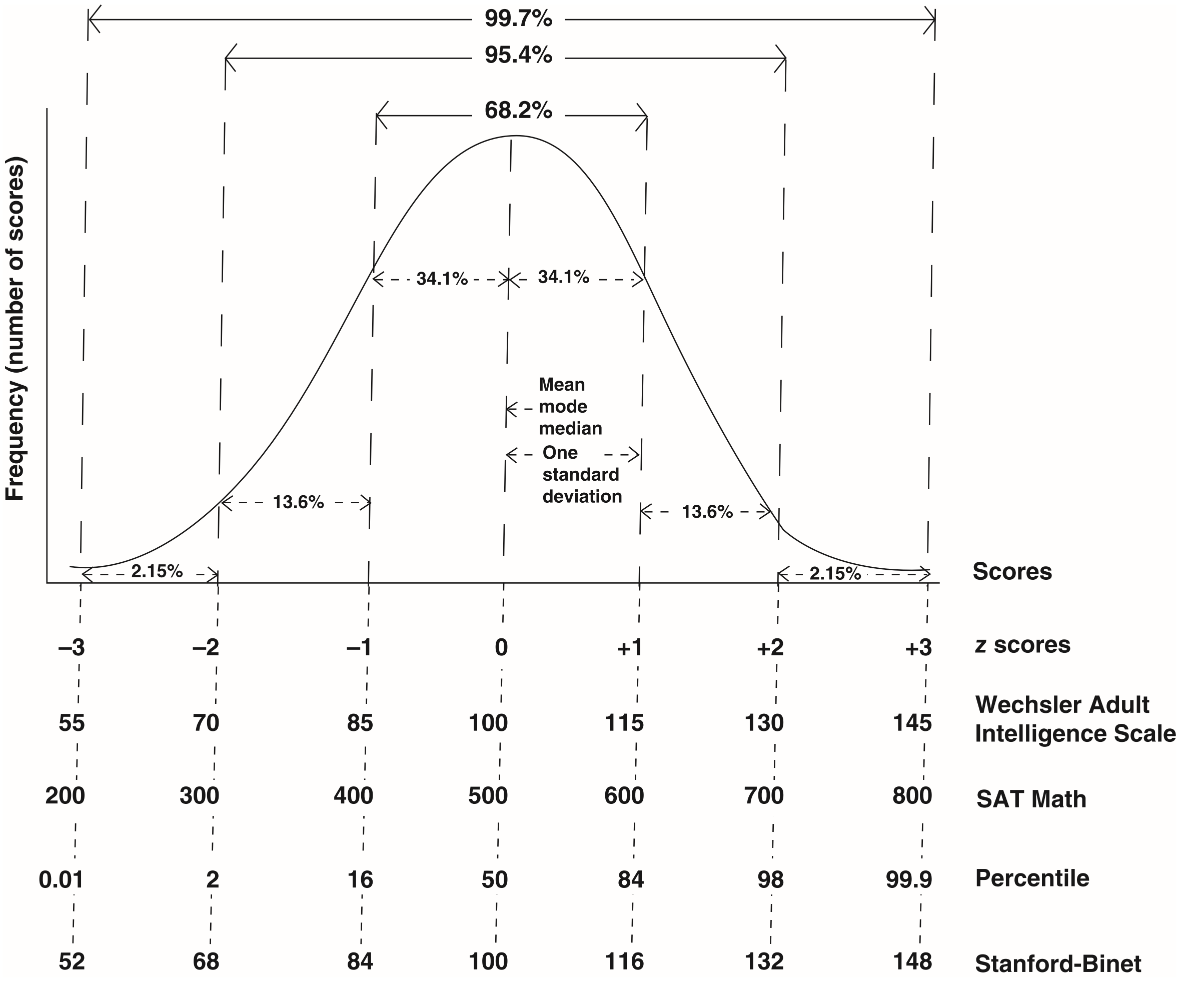

Figure 15.1 The normal curve.

themselves, can care for a home, achieve a sixth-grade education, hold a job, get married, and become an adequate parent. In schools, they are often *mainstreamed*, or integrated into regular education classes.

Individuals with moderate intellectual disability (about 10 percent) may achieve a second-grade education; may be given training in skills such as eating, toileting, hygiene, dressing, and grooming so that they can care for themselves; and may be given basic training in home management, consumer, and community mobility skills so that they can hold menial jobs and live successfully in a group home. Individuals with severe intellectual disability (about 5 percent) typically develop a very limited vocabulary and learn limited self-care skills. Usually they are unable to care for themselves adequately and do not develop enduring friendships. Individuals with profound intellectual disability (1–2 percent) require custodial care. Communities have been housing a greater proportion of cognitively disabled people than in the past. These people live with their own families or in group homes when possible. This deinstitutionalization is termed *normalization.*

Kinds of Intelligence

Is there one underlying capacity for intelligence, or do we have different, distinct ways of being intelligent? A contemporary of Alfred Binet, **Charles Spearman**, tested a large number of people on a number of different types of mental tasks. He used **factor analysis,** a statistical procedure that identifies closely related clusters of factors among groups of items by determining which variables have a high degree of correlation. Because all of the mental tasks had a high degree of correlation, he concluded that one important factor, which he called ***g,*** underlies all intelligence. Because the correlation wasn't a perfect 1.0 between all pairs of factors, he also concluded the existence of the less important *s,* or specialized abilities. **Louis Thurstone** disagreed with Spearman's concept of *g.* Based on factor analysis of tests of college students, Thurstone identified seven distinct factors he called *primary mental abilities,* including inductive reasoning, word fluency, perceptual speed, verbal comprehension, spatial visualization, numerical ability, and associative memory. J. P. Guilford divided intelligence into 150 different intelligence sets.

John Horn and **Raymond Cattell** determined that Spearman's *g* should be divided into two factors of intelligences: **fluid intelligence,** those cognitive abilities requiring speed or rapid learning that tend to diminish with adult aging; and **crystallized intelligence,** learned knowledge and skills such as vocabulary that tend to increase with age.

Multiple Intelligences

Howard Gardner is one of the many critics of the *g* or single factor intelligence theory. **Savants,** individuals otherwise considered mentally retarded, have a specific exceptional skill, typically in calculating, music, or art. To Howard Gardner, this is one indication that a single factor *g* does not underlie all intelligence. He has proposed a **theory of multiple intelligences.** Three of his intelligences are measured on traditional intelligence tests: logical-mathematical, verbal-linguistic, and spatial. Five of his intelligences are not usually tested on standardized tests: musical, bodily-kinesthetic, naturalistic, intrapersonal, and interpersonal. Gardner has also introduced the possibility of a ninth intelligence—existential—which would be seen in those who ask questions about our existence, life, death, and how we got here. According to Gardner, these abilities also represent ways that people process information differently in the world, which has led to changes in how some school systems classify gifted and talented children for special programs. Peter Salovey and John Mayer labeled the ability to perceive, express, understand, and regulate emotions as **emotional intelligence.**

Salovey's and Mayer's emotional intelligence combines Gardner's intrapersonal and interpersonal intelligences. Salovey, Mayer, and David Caruso developed the Multifactor Emotional Intelligence Scale (MEIS) to measure emotional intelligence. The items test the test taker's ability to perceive, understand, and regulate emotions.

Robert Sternberg also believes that intelligence is more than what is typically measured by traditional IQ tests, and has described three distinct types of intelligence in his **triarchic theory of intelligence:** analytic, creative, and practical. *Analytical* thinking is what is tested by traditional IQ test and what we are asked to do in school—compare, contrast, analyze, and figure out cause and effect relationships. *Creative* intelligence is evidenced by adaptive reactions to novel situations, showing insight, and being able to see more than one way to solve a problem. *Practical* intelligence is what some people consider "street smarts." This would include the ability to read people, knowing how to put together a bake sale, or being able to get to a distant location. Whether it is labeled as emotional intelligence, interpersonal intelligence, or practical intelligence, such emotionally smart people can often succeed in careers, marriages, and parenting, where people with higher IQ scores, but less emotional intelligence, fail.

Creativity

Creativity, the ability to generate ideas and solutions that are original, novel, and useful, is not usually measured by intelligence tests. According to the **threshold theory,** a certain level of intelligence is necessary, but not sufficient for creative work. Although many tests of creativity have been developed, such as the Torrance Test of Creative Thinking, the Christensen-Guilford Test, the Remote Associates Test, and the Wallach and Kogan Creative Battery, they do not have high criterion-related validity.

Because tests are used to make decisions, these are criticized for their shortcomings. Although psychometricians, other psychologists, educators, and ethicists agree that intelligence tests measure the ability to take tests well, they do not agree that intelligence tests actually measure intelligence. Since results of intelligence tests correlate highly with academic achievement, they do have predictive validity.

Heredity/Environment and Intelligence

A continuing theme of psychology known as the nature–nurture controversy asks to what extent intelligence is hereditary and to what extent it is learned. Intellectual disability resulting from genetic defects, such as **Down syndrome** (see Genetics and Behavior in Chapter 7), is primarily hereditary, whereas intellectual disability resulting from prenatal exposure to alcohol, **fetal alcohol syndrome (FAS)** (see Physical Development in Chapter 13), is primarily environmental. **Phenylketonuria (PKU)** results from the interaction of nature and nurture (see Genetics and Behavior in Chapter 7). About 75 percent of all cases of intellectual disability result from nurture, from sociocultural deprivation in an impoverished environment. This illustrates that both nature and nurture contribute to intelligence. Theorists continue to argue about the relative contributions of heredity/genes and environment/experience to intelligence because of the important implications. If intelligence is inherited, then special educational programs for disadvantaged groups are unnecessary. If, on the other hand, intelligence can be affected by better education and an enriched environment, special programs are warranted. For example, the Head Start program was designed to provide economically disadvantaged children with preschool opportunities to ready them for elementary school. Research shows that, compared to matched control groups, children who had the Head Start experience did better in the first two grades, thus supporting the nurture position. The program reduced the likelihood that these students would have to repeat a grade or be placed in a special education class. Opponents of the program say that this advantage is short-lived. Continuing disadvantages experienced by these youngsters are not being addressed, according to the defenders.

Studies of Twins

Additional studies to gauge the influence of genes on intelligence include comparing the intelligence test scores of identical twins (who share all of the same genes) reared together with the scores of fraternal twins (who share about half of the same genes). Identical twins have much more similar scores. Intelligence scores of adoptees are more like those of their biological parents than their adopted parents, and get even more similar with age. Comparing the intelligence test scores of identical twins reared apart reveals that they are very similar, and get even more similar with age. Brain scans of identical twins reveal similar brain volume and anatomy. Experiments with other animals, such as mice, indicate that genetic engineering can produce more intelligent animals.

Environmental Influences on Intelligence

On the other hand, some studies support the influence of the environment on intelligence. During childhood, siblings raised together are more similar in IQ than siblings raised apart. The IQs of children from deprived environments who have been moved into middle- and upper-class foster or adoptive families tend to increase. School attendance seems to result in increased IQ scores. Performance on IQ tests has been increasing steadily over the past three generations. This trend was noticed by James Flynn, who observed that every time tests were renormed, more questions needed to be answered correctly to earn the same score, yet the same proportion of the population was earning that score. In other words, a score of 100 on a present test is equivalent to a score of about 120 on a test from 70 years ago.

This *Flynn effect* cannot be attributed to a change in the human gene pool because that would take hundreds of years. Theorists attribute the Flynn effect to a number of environmental factors, including better nutrition, better health care, advances in technology, smaller families, better parenting, and increased access to educational opportunities.

Heritability is the proportion of variation among individuals in a population that results from genetic causes. Heritability for intelligence estimates range from 50 to 75 percent. Heritability deals with differences on the population level, not on the individual level. According to the *reaction range model*, genetic makeup determines the upper limit for an individual's IQ, which can be attained in an ideal environment, and the lower limit, which would result in an impoverished environment.

Human Diversity

Racial differences in IQ scores show African Americans, Native Americans, and Hispanic Americans typically scoring 10 to 15 points below the mean for white children. When comparing groups of people on any construct, such as intelligence, it is important to keep in mind the concept of **within-group differences** and **between-group differences.** The range of scores *within* a particular group, such as Hispanic Americans, is much greater than the difference between the mean scores of two different groups, such as Hispanic Americans and Asian Americans. According to Leon Kamin, even if heritability is high, differences in average IQ between groups could be caused entirely by environmental factors. Neither of these statistics tells us how any one individual will score. The difference between the mean scores could result from socioeconomic differences.

Stereotype Threat

Groups such as African Americans, Native Americans, Latinos, women, the elderly, and the economically disadvantaged are often stereotyped. **Stereotypes** are overgeneralized beliefs about the characteristics of members of a particular group, schema that are used to quickly judge others. Claude Steele hypothesized that members of stereotyped groups begin to doubt themselves and fear they will fulfill their group's negative stereotype. This anxiety interferes with their performance on tests, lowering their scores. This **stereotype threat,** anxiety that influences members of a group concerned that their performance on a test will confirm a negative stereotype, has been evidenced in studies by Steele, Joshua Aronson, and many others.

❯ Review Questions

Directions: For each question, choose the letter of the choice that best completes the statement or answers the question.

1. Aptitude tests are designed to measure
(A) previously learned facts
(B) future performance
(C) previously learned skills
(D) current competence
(E) your IQ score

2. A standardization sample for developing a test
(A) should be representative of all the types of people for whom the test is designed
(B) is an early version of the test to determine questions that differentiate individuals
(C) is a set of norms that will determine what score should be considered passing
(D) should include people from all different age groups, ethnic groups, and genders
(E) must include a standard set of directions for administering the test that all students will receive