I vividly remember my first genuine IQ test. I was 17 at the time. The youth director at my church was in graduate school, working on an advanced degree in psychology, and as part of a course in intelligence testing, he was required to administer an IQ test to several subjects. I was one of his selected “volunteers,” although I was also a friend. I remember wondering later about whether or not he had given me an unfair advantage on the test. He often responded to my asking for clarification by going into great detail while explaining a particular kind of question. I wondered if my score would be comparable to that of another person who was tested by someone who was not so generous about clarifying items. —from Psychology: Science, Behavior, and Life by Robert L. Crooks and Jean Stein, 1988

All tests have one characteristic that makes them both fascinating and remarkably practical: they try to make it possible to find out a great deal about a person in a short time. Tests can be somewhat useful with the following: in predicting how well a person might do in a particular career; in assessing an individual’s desires, interests, and attitudes; and in revealing psychological problems. One virtue of standardized tests is that they can provide comparable data about many individuals. Tests can show how an individual compares to others. Further, psychologists can use some tests to help people understand things about themselves more clearly. Using tests to predict behavior, though, is controversial. It is important to keep in mind what the test is measuring. One of the great dangers of testing is that we tend to forget that tests are merely tools for measuring and predicting human behavior. We start to think of test results (for example, an IQ) as an end in itself. The justification for using a test to make decisions about a person’s future depends on whether a decision based on test scores would be fairer and more accurate than one based on other criteria. The fairness and usefulness of a test depend on several factors: its reliability, its validity, and standardization.

**TEST RELIABILITY**

The term reliability refers to a test’s consistency—its ability to yield the same result under a variety of different circumstances. There are three basic ways of determining a test’s reliability. First, if a person retakes the test or takes a similar test within a short time after the first testing, does he or she receive approximately the same score? If, for example, you take a mechanical aptitude test three times in the space of six months and score 65 in January, a perfect score of 90 in March, and 70 in June, then the test is unreliable because it does not produce a measurement that is stable over time. The scores vary too much. This is assessing the measure’s test-retest reliability. The second measure of reliability is whether the test yields the same results when scored at different times by different people. If both your teacher and another teacher critique an essay test that you have taken, and one gives you a B while the other gives you a D, then you have reason to complain about the test’s reliability. The score you receive depends more on the grader than on you. This is called inter-scorer reliability. If the same teacher grades papers at different times, he or she may score the same essay differently. This is scorer reliability. On a reliable test, your score would be the same no matter who graded it and when it was graded. One final way of determining a test’s reliability is to find out whether, if you divide the test in half and score each half separately, the two scores are approximately the same. This is called split-half reliability. If a test is supposed to measure one quality in a person—for example, reading comprehension or administrative ability—it should not have some sections on which the person scores
high and others on which he or she scores low. In checking tests for reliability, psychologists try to prevent variables from influencing a person’s score. All kinds of irrelevant matters can interfere with a test. No test can screen out all interferences, but a highly reliable test can eliminate a good part of them.

**TEST VALIDITY**

A test may be reliable but still not valid. **Validity** is the ability of a test to measure what it is intended to measure. For example, a test that consists primarily of vocabulary lists will not measure ability for engineering. A history test will not measure general learning ability. A test you take in physical education class may not measure your knowledge of grammar, or a history test that asks questions that were not covered in class does not measure what you learned in class. Determining the validity of a test is more complex than assessing its reliability. One of the chief methods for measuring validity is to find out how well a test predicts performance—its predictive validity. For example, a group of psychologists designs a test to measure management ability. They ask questions about management systems, attitudes toward employees, and other relevant information. Will the people who score high on this test really make good managers? Suppose the test makers decide that a good way to check the validity of the test is to find out how much a manager’s staff improves in productivity in one year. If the staffs of those managers who scored high on the test produce more than the staffs of those managers who scored low on the test, the test may be considered valid. Corporations may then adopt it as one tool to use in deciding whom to hire as managers, assuming the test is also valid in their situation. What if managers who are good at raising productivity are poor at decision making? It may be that this test measures talent for improving productivity, not general management ability. This is the kind of difficulty psychologists encounter in trying to assess the validity of a test. As the example shows, nothing can be said about a test’s validity unless the purpose of the test is absolutely clear.

**STANDARDIZATION**

Tests must be standardized. Standardization refers to two things. First, standardized tests must be administered and scored the same way every time. Test administrators are trained to follow the same procedures and to ask the same questions the same way. If test administrators give instructions in an inconsistent manner or provide hints, errors in assessing the test taker would result. Second, standardization refers to establishing the norm, or average score, made by a large group of people.

**Establishing Norms**

Once a test result is obtained, the examiner must translate the score into something useful. Suppose a child answers 32 of 50 questions on a vocabulary test correctly. What does this score mean? If the test is reliable and valid, it means that the child can be expected to understand a certain percentage of the words in a book at the reading level being tested. In other words, the score predicts how the child will perform at a given level. Yet a “raw” score does not tell us where the child stands in relation to other children at his or her age and grade level. If most children answered 45 or more questions correctly, 32 is a low score. If most answered only 20 questions correctly, however, 32 is a very high score. When psychologists design a test to be used in a variety of settings, they usually set up a scale for comparison by establishing norms. This is usually done by transforming raw test scores into a percentile system, which resembles what is called “grading on the curve.” In the percentile system,
the scores actually achieved on the test are placed in order, ranging from the highest to the lowest. Each score is then compared with this list and assigned a percentile according to the percentage of scores that fall at or below this point. For example, if half the children in the above example scored 32 or below, then a score of 32 is at the 50th percentile. If 32 were the top score, it would be at the 100th percentile. In the example given in Figure 13.4, a score of 32 puts the child in the 75th percentile, because only 25 percent of the children scored higher than she did. The test is given to a large representative sample of the group to be measured—for example, sixth graders or army privates or engineers. Percentiles are then established on the basis of the scores achieved by this standardization group. These percentiles are called the test’s norms. Most of the intelligence, aptitude, and personality tests you will encounter have been provided with norms in this way. Your percentile on an aptitude test such as the Scholastic Assessment Test (SAT), for example, reflects your standing among people of your age and grade who have taken these exams. You should remember, however, that norms are not really standards—even though a norm group is sometimes misleadingly referred to as a “standardization group.” Norms refer only to what has been found to be average for a particular group. If John can read at the 50th percentile level, that does not mean that he has met some absolute standard for ability to read. It means only that he reads better than half the population and worse than the other half in his particular group. In summary, when you take a test and obtain your score, you should consider the following questions in evaluating the results. (1) Do you think that if you took the same test again, you would receive a similar score? (2) Does your performance on this test reflect your normal performance in the subject? (3) If you were to compare your score with those of your classmates, would it reflect your general standing within that group?
Write a definition, using a complete sentence, for the following terms:

reliability; validity; percentile-system; norms

Write a paragraph about what you think the author's meaning is in the following primary source and include your opinion:

Not Fair!

I vividly remember my first genuine IQ test. I was 17 at the time. The youth director at my church was in graduate school, working on an advanced degree in psychology, and as part of a course in intelligence testing, he was required to administer an IQ test to several subjects. I was one of his selected “volunteers,” although I was also a friend. I remember wondering later about whether or not he had given me an unfair advantage on the test. He often responded to my asking for clarification by going into great detail while explaining a particular kind of question. I wondered if my score would be comparable to that of another person who was tested by someone who was not so generous about clarifying items. —from Psychology: Science, Behavior, and Life by Robert L. Crooks and Jean Stein, 1988