

Finally, even our best psychological units of measure leave much to be desired. Units are set equal by definition; the definition may have a certain plausibility, but the equality of units cannot be established in any fundamental sense. For this reason, the addition, subtraction, and comparison of scores will always be somewhat suspect. Furthermore, the precision with which the attribute is assessed—the consistency from one occasion to another or from one appraiser to another—is often discouragingly low.

In spite of the problems involved in developing educational and psychological measuring devices, the task has been proven worthwhile. The procedures now available have developed a record of usefulness in helping individuals make decisions in a wide variety of human contexts. Efficiency of education and equality of opportunity are enhanced by the proper use and interpretation of tests. Use of interest and personality inventories has led to people having greater self-understanding and reduced psychological discomfort. Measures of human abilities have been used to provide access to educational and occupational positions without regard to ethnic or racial background.

Critics of testing are quick to point out that inequalities in test performance still exist, but, in the last 25 years, the users of tests have become much more cautious in their interpretations and much more sensitive to the rights of the test taker. Generally, the information provided by tests is more accurate than that available from other sources. When we acknowledge that decisions *must* be made and that accurate information coupled with a clear understanding of our values generally leads to the best decisions, we will find continuing growth in the importance of measurement and assessment processes in psychology and education.

SOME CURRENT ISSUES IN MEASUREMENT

Educational and psychological assessments are far from perfect. Since the earliest attempts to develop measurement techniques in a systematic way, the procedures have been a target for a wide spectrum of critics. Much of this criticism has been a justified response to the naive enthusiasm of measurement proponents and some of their ill-conceived applications and interpretations of measurement results. In subsequent chapters in our discussion of test interpretation and use, we will try to be sensitive to earlier criticisms and to the more technical questions that arise concerning the reliability and validity of test results. At this point, however, we will identify and comment briefly on some of the issues that have been of special concern.

Testing Minority Individuals

For many years, the use and interpretation of tests within ethnic minority and other groups whose experiences and cultures may differ from that typical of the general population for which a test was designed have received a great deal of attention. There are, of course, all sorts of subgroups in American society, differing from one another in a variety of ways. Ethnic and linguistic minorities are probably the most clear-cut of these; they are the ones for whom the appropriateness of tests and questionnaires designed to reflect the values and experiences of typical middle-class, White Americans are most open to question. In recent years, the number of Americans for whom English is not the preferred language has increased to the point where many ability and achievement tests are available in Spanish translations. Additionally, feminist groups have complained that test material is often male oriented. Major test publishers now go to considerable lengths to ensure that their test items do not present an unfair challenge to women or to members of ethnic or linguistic minority groups. Test users, teachers, counselors, and school psychologists, for example, are also required to

have training in recognizing and, to the extent possible, ameliorating the potentially adverse consequences of using particular instruments with people for whom they may not be appropriate.

Some questions arise concerning achievement tests that attempt to assess what a student has *learned* to do. In part, these questions center on the degree to which the same educational objectives hold for groups from minority cultures. Is the ability to read standard English with understanding an important goal for African American, Hispanic, or Native American children? Is knowledge about the U.S. Constitution as relevant for these groups as for the middle-class, White eighth-grader? One senses that as far as the basic skills of dealing with language and numbers are concerned, many of the same objectives would apply, but as one moves into the content areas of history and literature, more divergence might be appropriate.

In part, the questions focus on the specific materials through which basic skills are exhibited. Is the same reading passage about the life of the Zulu appropriate in a reading test for a Hispanic or a Native American youngster as it would be for a child of White middle-class background? Or should test materials, and perhaps instructional materials as well, be specifically tailored to the life and experiences of each ethnic group? We know too little about the importance of factors of specific content on performance in areas such as reading comprehension and need to do further research to make informed decisions.

The motivation of minority groups to do well on tests in school is also an issue. Some minority groups, such as recent immigrants from Southeast Asia, place great emphasis on individual academic achievement. Others, such as some Native American groups, place much more value on communal accomplishments and group identity. In many cases of academic difficulty, we must ask whether unfortunate experiences with testing in school have soured the students on the whole enterprise so that they withdraw from the task and do not try. It is perhaps a challenge to the whole pattern of education, not just to the choice of testing instruments, to provide a reasonable mixture of satisfying and success-enhancing experiences in school with all types of school tasks. As far as possible, tasks and tests should be adapted to the present capabilities and concerns of the individual student, but accountability concerns also dictate that this should not come at the expense of achievement.

Many more questions—perhaps more serious ones—are raised when tests are used as a basis for deciding what an individual *can learn to do*, that is, as aptitude measures. An inference from a test score obtained at Time 1 concerning what a person can learn to do by Time 2 is clearly a more questionable inference than one that merely states what that person can do at Time 1. There are many intervening events that can throw the prediction off—and there are correspondingly more possibilities of biasing factors coming in to systematically distort the prediction for minority group members. Present facts that imply one prediction for the majority group may imply a different prediction for a minority group member whose experiences before testing were far from typical. Remembering that decisions must be made, the problem is to learn what types of inferences *can* appropriately be made for individuals with differing backgrounds and what types of adjustments or modifications need to be built into the system to permit the most accurate and equitable inferences and decisions for all people. The best decision-making systems are likely to be those that employ a monitoring process to assess the progress of each student and contain enough flexibility that a correction can be made if the original decision appears to be in error.

Invasion of Privacy

A second concern often expressed involves invasion of privacy. What kind of information is reasonable to require individuals to give about themselves and under what circumstances? This issue arises not only in relation to testing, but also in relation to all types of information

from and about a person. What types of records should appropriately be kept? And to whom should they be made available? At one end of the spectrum are tests of job knowledge or skill, such as clerical tests, to which few would object when the skill is clearly and directly relevant to the position for which the person is applying and access to the information is limited to potential employers. It is hard to argue that a test of the ability to use a word processor should not be used to select the best candidate for a job where the chief duties will involve word processing. At the other end of the spectrum are self-descriptive instruments that lead to inferences about emotional stability or a scattering of tests that try to assess honesty under temptation to lie or steal; in these latter tests, individuals are led to give information about themselves without being aware of what information they are giving or how it will be interpreted. The use of these instruments seems most open to question. In between are instruments that involve varying degrees of self-revelation and that appear to have varying degrees of immediate relevance to a decision.

The issue is not only what information is being obtained, but also the purpose for which it is being obtained. Is the information being obtained at the individual's request to provide help with personal or vocational problems or for use in a counseling relationship? The fact that the person has come for help implies a willingness on his or her part to provide the information needed for help to be given; here, invasion of privacy becomes a matter of relatively minor concern. However, when the information is obtained to further institutional objectives—that is, those of an employer, of an educational institution, or of “science”—then concern for the individual's right to privacy mounts, and some equitable balance must be struck between individual values and rights and social ones. For example, more students were willing to allow the use of a questionnaire to verify the emotional stability of an airline pilot, who is responsible for many lives, than to verify that of a bank clerk (75% vs. 34%). The rights of the individual are not absolute, but the feeling is often expressed that these rights have received too little consideration in the past.

To an increasing extent, the courts are taking a role in deciding what information is allowable. Several court cases have required a demonstration of the validity or relevance of test scores or personality profiles to job performance before such instruments can be used for employee selection. These court decisions have affected the type of information that may be collected and who may have access to that information. As the security of information stored in computer files becomes more doubtful, privacy concerns also take on increased salience.

The Use of Normative Comparisons

A somewhat different type of issue has been raised concerning the emphasis, in test interpretation, on comparing the performance of one person with norms representing the typical performance of a national, or sometimes a local, sample of people. The point being made with increasing fervor is that many of the decisions for which tests are used, especially instructional ones, do not call for—and are only confused by—comparisons with other people. The essential information is whether the person can perform a specified task; this information should guide decisions on what should be taught or what tasks should be undertaken next. Of course, there are settings in which comparison with others is essential to sound judgment and decision making: Is the latest applicant at the personnel office a good keyboard operator? How do we define “good keyboard operator” except in terms of the performance of other job applicants? A rate of 60 words per minute with two errors per minute is meaningless unless we know that the average graduate from a commercial school can enter about 70 words per minute with perhaps one error per

minute. The employer wants an employee who comes up to a standard of typical performance, and that standard is set by the performance of others. In the same way, a school system trying to evaluate the reading achievement of its sixth-graders as “good,” “satisfactory,” or “needing improvement” needs some benchmark against which to compare its own performance. No absolute standard of reading performance exists. Whether it is reasonable to expect the sixth-graders in Centerville to read and understand a particular article in *Reader's Digest* can only be judged by knowing whether representative groups of sixth-graders nationwide are able to read and understand the same or similar articles.

In the past, normative comparisons with an outside reference group have often been used to guide decisions for which they had little or no relevance. Greater emphasis is now being given to criterion-referenced and mastery tests and to performance assessments. These procedures have an important place in making some sorts of decisions, but not all. We need to know which type of comparison is useful for which situation. When should we ask if a student can satisfactorily perform a specific task? And when should we ask how that student compares in relation to other students? This issue is discussed in more detail in Chapter 3.

Other Factors That Influence Scores

An issue that has concerned measurement professionals and consumers for many years is the effect of extraneous factors on performance. One such factor that has received considerable study is the effect of anxiety. Does anxiety raise or lower performance on achievement tests? Are some groups more prone to this effect than others? If test anxiety does have a systematic effect, what can or should the teacher or examiner do to minimize the effect?

Other factors, such as the nutritional status of students or their ability to concentrate on the tasks at hand, may also affect scores. Two particular factors that have received considerable theoretical attention are (1) the racial, ethnic, or gender relationship between examiner and examinee and (2) the effect of coaching on test performance. Although the effects of these extraneous factors are not clear, public concern over their possible effects has prompted educators and measurement professionals to give them increasing attention.

Rights and Responsibilities of Test Takers

Researchers in all areas of science where living organisms are used in research, including psychologists and educational researchers, have grown increasingly sensitive to the rights of their research participants. Psychologists and counselors in professional practice have also developed a greater awareness of their responsibilities toward their clients. On the research side, this trend has led to the development of institutional research review boards (IRBs) at universities and funding agencies that review planned research on humans and animals to make sure that the rights of the study participants are respected.

In the area of educational and psychological measurement, the awareness of the rights of examinees has found voice in several publications by the professional organizations most concerned with tests and testing practice. The most influential publication has been a series of sets of guidelines for educational and psychological tests that was first developed by the American Psychological Association in 1954. Since then, the guidelines have been revised four times. The most recent edition, published in 1999 as a joint project of the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), titled *Standards for Educational and Psychological Testing*

(Standards), explains the current standards for the practice of test construction, administration, and interpretation to protect the rights of test takers.

The APA maintains a Web page devoted to issues in the fair use and interpretation of tests at <http://www.apa.org/science/testing.html>. Here one can read the APA statement on "Rights and Responsibilities of Test Takers," a statement that points out that both the person giving the test and the test taker have obligations in a testing situation. For example, the test taker has the right to an explanation of the procedures he or she will encounter and how they will be used, but test takers also have a responsibility to ask questions about aspects of the testing session they do not understand and to participate responsibly in the testing enterprise. This site has links to a wide array of resources about testing and provides an order form to obtain a copy of the current version of the *Standards*.

The APA has published a number of articles on high-stakes testing (testing in which the outcome will have important life consequences for the examinees) and related issues in all levels of American public education that may be accessed at the APA Web site (<http://www.apa.org>). Other organizations, such as the American Counseling Association, the National Association of School Psychologists, the NCME, and the Society for Industrial and Organizational Psychology also maintain Web sites that may from time to time contain information about tests and testing. Links to all of these Web sites from the APA Web site mentioned in the preceding paragraph can be found under the heading "links to other testing-related sites." A particularly useful site that we will visit in more detail in Chapter 6 is maintained by the Buros Institute of Mental Measurements.

SUMMARY

The objective of this book is to improve the knowledge and understanding of decision makers by giving them a better basis for evaluating different measurement procedures and for knowing what each signifies and how much confidence can be placed on them. To this end, we will describe the process of preparing test exercises, develop the general criteria of validity and reliability by which all

types of measures are evaluated, provide a familiarity with the different ways of reporting test scores, and describe and evaluate a number of the techniques and instruments commonly used for appraising human characteristics. Our success will be measured by the extent to which our readers use measurement results with wisdom and restraint in their decision making.

QUESTIONS AND EXERCISES

1. List some instances, preferably recent, of decisions you made about yourself or others made about you in which results from some kind of educational or psychological measurement played a part. Classify each decision as (1) instructional, (2) selection, (3) placement or classification, or (4) personal.
2. From your personal experience of a decision made, describe one or more instances for which an educational or psychological measurement could have been helpful but was unavailable. On what basis was the decision actually made?
3. What are some alternatives to educational or psychological measurements as guides for each of the following decisions?
 - a. How much time should be spent on phonics in the first-grade reading program?
 - b. Which 5 of 15 applicants should be hired as computer programmers?

- c. Should Charles Turner be encouraged to realize his desire to go to college and law school and to become a lawyer?
4. What are the advantages of the alternatives you proposed in Question 3, in comparison with some type of test or questionnaire? What are the disadvantages?
5. To what extent and in what way might *values* be involved in each of the decisions stated in Question 3?
6. Give an example of each of the following types of tests:
 - a. A criterion-referenced achievement test
 - b. A norm-referenced achievement test
 - c. An aptitude test
 - d. A measure of likes and preferences
 - e. A measure of personality or adjustment
 - f. A measure of a trait or construct
7. For one of the attributes listed here, describe how you might (1) define the attribute, (2) set up procedures to make that attribute observable, and (3) quantify the attribute.
 - a. Critical thinking
 - b. Friendliness
 - c. "Good citizenship" in an elementary school pupil
 - d. Competence as an automobile driver
8. The usefulness of tests for making decisions involving minority group members depends on the type of decision involved. For what sorts of decisions would a standardized test be most defensible? For what sorts would one be most questionable?
9. Which of the following practices would you consider acceptable? Which would you consider to be an invasion of privacy? What factors influence your opinion?
 - a. Requiring medical school applicants (1) to take an achievement test in chemistry, (2) to fill out a questionnaire designed to assess emotional stability, or (3) to take a scale of attitudes toward socialized medicine
 - b. Requiring applicants for a secretarial job (1) to take a test of general intelligence, (2) to take a test of keyboarding speed and accuracy, or (3) to fill out a questionnaire designed to appraise dependability
 - c. Giving a 10-year-old boy whose reading achievement is at the 8-year-old level (1) a nonverbal test of intellectual ability, (2) an interview focused on the conditions in his home, or (3) a series of diagnostic tests of specific reading skills
10. Why is it important for test users to know and adhere to professional recommendations for appropriate test use?
11. Write down some rights or responsibilities that you believe test users and test takers have. Then go to the American Psychological Association Web site (<http://www.apa.org/science/testing.html>) and read the rights and responsibilities statement mentioned earlier in this chapter. Are factors listed that were not on your list? Did you include any factors that the statement did not mention?

SUGGESTED READINGS

- Alexander, L., & James, H. T. (1987). *The nation's report card: Improving the assessment of student achievement*. Washington, DC: National Academy of Education.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.
- Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, 61, 27-41.
- Brennan, R. L. (2006). Perspectives on the evolution and future of measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1-16). Westport, CT: Praeger.

Measurement and Evaluation in Psychology and Education

Eighth Edition



Robert M. Thorndike | Tracy Thorndike-Christ