PART ONE: BASIC ISSUES

1 The Nature of Tests

AIM In this chapter we cover four basic issues. First, we focus on what is a test, not just a formal definition, but on ways of thinking about tests. Second, we try to develop a "taxonomy" of tests, that is we look at various ways in which tests can be categorized. Third, we look at the ethical aspects of psychological testing. Finally, we explore how we can obtain information about a specific test.

INTRODUCTION

Most likely you would have no difficulty identifying a psychological test, even if you met one in a dark alley. So the intent here is not to give you one more definition to memorize and repeat but rather to spark your thinking.

What is a test? Anastasi (1988), one of the best known psychologists in the field of testing, defined a test as an "objective" and "standardized" measure of a sample of behavior. This is an excellent definition that focuses our attention on three elements: (1) *objectivity*: that is, at least theoretically, most aspects of a test, such as how the test is scored and how the score is interpreted, are not a function of the subjective decision of a particular examiner but are based on objective criteria; (2) *standardization*: that is, no matter who administers, scores, and interprets the test, there is uniformity of procedure; and (3) *a sample of behavior*: a test is not a psychological X-ray, nor does it necessarily reveal hidden conflicts and forbidden wishes; it is a sample of a person's behavior, hopefully a representative sample from which we can draw some inferences and hypotheses.

There are three other ways to consider psychological tests that we find useful and we hope you will also. One way is to consider the administration of a test as an experiment. In the classical type of experiment, the experimenter studies a phenomenon and observes the results, while at the same time keeping in check all extraneous variables so that the results can be ascribed to a particular antecedent cause. In psychological testing, however, it is usually not possible to control all the extraneous variables, but the metaphor here is a useful one that forces us to focus on the standardized procedures, on the elimination of conflicting causes, on experimental control, and on the generation of hypotheses that can be further investigated. So if I administer a test of achievement to little Sandra, I want to make sure that her score reflects what she has achieved, rather than her ability to follow instructions, her degree of hunger before lunch, her uneasiness at being tested, or some other influence.

A second way to consider a test is to think of a test as an interview. When you are administered an examination in your class, you are essentially being interviewed by the instructor to determine how well you know the material. We discuss interviews in Chapter 18, but for now consider the following: in most situations we need to "talk" to each other. If I am the instructor, I need to know how much you have learned. If I am hiring an architect to design a house or a contractor to build one, I need to evaluate their competency, and so on. Thus "interviews" are necessary, but a test offers many advantages over the standard

interview. With a test I can "interview" 50 or 5,000 persons at one sitting. With a test I can be much more objective in my evaluation because for example, multiple-choice answer sheets do not discriminate on the basis of gender, ethnicity, or religion.

A third way to consider tests is as tools. Many fields of endeavor have specific tools – for example, physicians have scalpels and X-rays, chemists have Bunsen burners and retorts. Just because someone can wield a scalpel or light up a Bunsen burner does not make him or her an "expert" in that field. The best use of a tool is in the hands of a trained professional when it is simply an aid to achieve a particular goal. Tests, however, are not just psychological tools; they also have political and social repercussions. For example, the well-publicized decline in SAT scores (Wirtz & Howe, 1977) has been used as an indicator of the terrible shape our educational system is in (National Commission, 1983).

A test by any other name. . . . In this book, we use the term *psychological test* (or more briefly *test*) to cover those measuring devices, techniques, procedures, examinations, etc., that in some way assess variables relevant to psychological functioning. Some of these variables, such as intelligence, introversion-extraversion, and self-esteem are clearly "psychological" in nature. Others, such as heart rate or the amount of palmar perspiration (the galvanic skin response), are more physiological but are related to psychological functioning. Still other variables, such as socialization, delinquency, or leadership, may be somewhat more "sociological" in nature, but are of substantial interest to most social and behavioral scientists. Other variables, such as academic achievement, might be more relevant to educators or professionals working in educational settings. The point here is that we use the term *psychological* in a rather broad sense.

Psychological tests can take a variety of forms. Some are true-false *inventories*, others are *rating scales*, some are actual *tests*, whereas others are *questionnaires*. Some tests consist of materials such as inkblots or pictures to which the subject responds verbally; still others consist of items such as blocks or pieces of a puzzle that the subject manipulates. A large number of tests are simply a set of printed items requiring some type of written response.

Testing vs. assessment. *Psychological assessment* is basically a judgmental process whereby a broad range of information, often including the results of psychological tests, is integrated into a meaningful understanding of a particular person. If that person is a client or patient in a psychotherapeutic setting, we call the process *clinical assessment*. Psychological testing is thus a narrower concept referring to the psychometric aspects of a test (the technical information about the test), the actual administration and scoring of the test, and the interpretation made of the scores. We could of course assess a client simply by administering a test or *battery* (group) of tests. Usually the assessing psychologist also interviews the client, obtains background information, and where appropriate and feasible, information from others about the client [see Korchin, 1976, for an excellent discussion of clinical assessment, and G. J. Meyer, Finn, Eyde, et al. (2001) for a brief overview of assessment].

Purposes of tests. Tests are used for a wide variety of purposes that can be subsumed under more general categories. Many authors identify four categories typically labeled as: *classification, self-understanding, program evaluation*, and *scientific inquiry*.

Classification involves a decision that a particular person belongs in a certain category. For example, based on test results we may assign a diagnosis to a patient, place a student in the introductory Spanish course rather than the intermediate or advanced course, or certify that a person has met the minimal qualifications to practice medicine.

Self-understanding involves using test information as a source of information about oneself. Such information may already be available to the individual, but not in a formal way. Marlene, for example, is applying to graduate studies in electrical engineering; her high GRE scores confirm what she already knows, that she has the potential abilities required for graduate work.

Program evaluation involves the use of tests to assess the effectiveness of a particular program or course of action. You have probably seen in the newspaper, tables indicating the average

achievement test scores for various schools in your geographical area, with the scores often taken, perhaps incorrectly, as evidence of the competency level of a particular school. Program evaluation may involve the assessment of the campus climate at a particular college, or the value of a drug abuse program offered by a mental health clinic, or the effectiveness of a new medication.

Tests are also used in scientific inquiry. If you glance through most professional journals in the social and behavioral sciences, you will find that a large majority of studies use psychological tests to operationally define relevant variables and to translate hypotheses into numerical statements that can be assessed statistically. Some argue that development of a field of science is, in large part, a function of the available measurement techniques (Cone & Foster, 1991; Meehl, 1978).

Tests as experimental procedure. If we accept the analogy that administering a test is very much like an experiment, then we need to make sure that the experimental procedure is followed carefully and that extraneous variables are not allowed to influence the results. This means, for example, that instructions and time limits need to be adhered to strictly. The greater the control that can be exercised on all aspects of a test situation, the lesser the influence of extraneous variables. Thus the scoring of a multiple-choice exam is less influenced by such variables as clarity of handwriting than the scoring of an essay exam; a true-false personality inventory with simple instructions is probably less influenced than an intelligence test with detailed instructions.

Masling (1960) reviewed a variety of studies of variables that can influence a testing situation, in this case "projective" testing (see Chapter 15); Sattler and Theye (1967) did the same for intelligence tests. We can identify, as Masling (1960) did, four categories of such variables:

1. *The method of administration.* Standard administration can be altered by disregarding or changing instructions, by explicitly or implicitly giving the subject a set to answer in a certain way, or by not following standard procedures. For example, Coffin (1941) had subjects read fictitious magazine articles indicating what were more socially acceptable responses to the Rorschach Inkblot test. Subsequently they were tested with the Rorschach and the responses clearly showed a suggestive influence because of the prior readings. Ironson and Davis (1979) administered a test of creativity three times, with instructions to "fake creative," "fake uncreative," or "be honest"; the obtained scores reflected the influence of the instructions. On the other hand, Sattler and Theye (1967) indicated that of twelve studies reviewed, which departed from standard administrative procedures, only five reported significant differences between standard and nonstandard administration.

2. *Situational variables.* These include a variety of aspects that presumably can alter the test situation significantly, such as a subject feeling frustrated, discouraged, hungry, being under the influence of drugs, and so on. Some of these variables can have significant effects on test scores, but the effects are not necessarily the same for all subjects. For example, Sattler and Theye (1967) report that discouragement affects the performance of children but not of college students on some intelligence tests.

3. *Experimenter variables.* The testing situation is a social situation, and even when the test is administered by computer, there is clearly an experimenter, a person in charge. That person may exhibit characteristics (such as age, gender, and skin color) that differ from those of the subject. The person may appear more or less sympathetic, warm or cold, more or less authoritarian, aloof, more adept at establishing rapport, etc. These aspects may or may not affect the subject's test performance; the results of the available experimental evidence are quite complex and not easily summarized. We can agree with Sattler and Theye (1967), who concluded that the experimenter-subject relationship is important and that (perhaps) less qualified experimenters do not obtain appreciably different results than more qualified experimenters. Whether the race, ethnicity, physical characteristics, etc., of the experimenter significantly affect the testing situation seems to depend on a lot of other variables and, in general, do not seem to be as powerful an influence as many might think.

4. *Subject variables.* Do aspects of the subject, such as level of anxiety, physical attractiveness, etc., affect the testing situation? Masling (1960)

used attractive female accomplices who, as test subjects, acted "warm" or "cold" toward the examiners (graduate students). The test results were interpreted by the graduate students more favorably when the subject acted warm than when she acted cold.

In general what can we conclude? Aside from the fact that most studies in this area seem to have major design flaws and that many specific variables have not been explored consistently, Masling (1960) concluded that there is strong evidence of situational and interpersonal influences in projective testing, while Sattler and Theye (1967) concluded that:

1. Departures from standard procedures are more likely to affect "specialized" groups, such as children, schizophrenics, and juvenile delinquents than "normal" groups such as college students;
2. Children seem to be more susceptible to situational factors, especially discouragement, than are college-aged adults;
3. Rapport seems to be a crucial variable, while degree of experience of the examiner is not;
4. Racial differences, specifically a white examiner and a black subject, may be important, but the evidence is not definitive.

Tests in decision making. In the real world, decisions need to be made. To allow every person who applies to medical school to be admitted would not only create huge logistical problems, but would result in chaos and in a situation that would be unfair to the candidates themselves, some of whom would not have the intellectual and other competencies required to be physicians, to the medical school faculty whose teaching efforts would be diluted by the presence of unqualified candidates, and eventually to the public who might be faced with incompetent physicians.

Given that decisions need to be made, we must ask what role psychological tests can play in such decision making. Most psychologists agree that major decisions should not be based on the results of a single test administration, that whether or not state university admits Sandra should not be based solely on her SAT scores. In fact, despite a stereotype to the contrary, it is rare for such decisions to be based solely on test data. Yet in many situations, test data represent the only source of objective data standard for all candidates; other sources of data such as interviews, grades, and letters of recommendation are all "variable" – grades from different schools or different instructors are not comparable, nor are letters written by different evaluators. Finally, as scientists, we should ask what is the empirical evidence for the accuracy of predicting future behavior. That is, if we are admitting college students to a particular institution, which sources of data, singly or in combination, such as interviewers' opinions, test scores, high school GPA, etc., would be most accurate in making relevant predictions, such as, "Let's admit Marlene because she will do quite well academically." We will return to this issue, but for now let me indicate a general psychological principle that past behavior is the best predictor of future behavior, and a corollary that the results of psychological tests can provide very useful information on which to make more accurate future predictions.

Relation of test content to predicted behavior. Rebecca is enrolled in an introductory Spanish course and is given a Spanish vocabulary test by the instructor. Is the instructor interested in whether Rebecca knows the meaning of the specific words on the test? Yes indeed, because the test is designed to assess Rebecca's mastery of the vocabulary covered in class and in homework assignments. Consider now a test such as the SAT, given for college admission purposes. The test may contain a vocabulary section, but the concern is not whether an individual knows the particular words; knowledge of this sample of words is related to something else, namely doing well academically in college. Finally, consider a third test, the XYZ scale of depression. Although the scale contains no items about suicide ideation, it has been discovered empirically that high scorers on this scale are likely to attempt suicide. These three examples illustrate an important point: In psychological tests, the content of the test items may or may not cover the behavior that is of interest – there may be a lack of correspondence between test items and the predicted behavior. But a test can be quite useful if an empirical correspondence between test scores and real-life behavior can be shown.

CATEGORIES OF TESTS

Because there are thousands of tests, it would be helpful to be able to classify tests into categories, just as a bookstore might list its books under different headings. Because tests differ from each other in a variety of ways, there is no uniformly accepted system of classification. Therefore, we will invent our own based on a series of questions that can be asked of any test. I should point out that despite a variety of advances in both theory and technique, standardized tests have changed relatively little over the years (Linn, 1986), so while new tests are continually published, a classificatory system should be fairly stable, i.e., applicable today as well as 20 years from now.

Commercially published? The first question is whether a test is commercially published (sometimes called a proprietary test) or not. Major tests like the Stanford-Binet and the Minnesota Multiphasic Personality Inventory are available for purchase by qualified users through commercial companies. The commercial publisher advertises primarily through its catalog, and for many tests makes available, for a fee, a *specimen set*, usually the test booklet and answer sheet, a scoring key to score the test, and a test manual that contains information about the test. If a test is not commercially published, then a copy is ordinarily available from the test author, and there may be some accompanying information, or perhaps just the journal article where the test was first introduced. Sometimes journal articles include the original test, particularly if it is quite short, but often they will not. (Examples of articles that contain test items are R. L. Baker, Mednick & Hocevar, 1991; L. R. Good & K. C. Good, 1974; McLain, 1993; Rehfisch, 1958a; Snell, 1989; Vodanovich & Kass, 1990). Keep in mind that the contents of journal articles are copyright and permission to use a test must be obtained from both the author and the publisher.

If you are interested in learning more about a specific test, first you must determine if the test is commercially published. If it is, then you will want to consult the *Mental Measurements Yearbook* (MMY), available in most university libraries. Despite its name, the MMY is published at irregular intervals rather than yearly. However, it is an invaluable guide. For many commercially published tests, the MMY will provide a brief description of the test (its purpose, applicable age range, type of score generated, price, administration time, and name and address of publisher), a bibliography of citations relevant to the test, and one or more reviews of the test by test experts. Tests that are reviewed in one edition of the MMY may or may not be reviewed in subsequent editions, so locating information about a specific test may involve browsing through a number of editions. MMY reviews of specific tests are also available through a computer service called the Bibliographic Retrieval Services.

If the test you are interested in learning about is not commercially published, it will probably have an author(s) who published an article about the test in a professional journal. The journal article will most likely give the author's address at the time of publication. If you are a "legitimate" test user, for example a graduate student doing a doctoral dissertation or a psychologist engaged in research work, a letter to the author will usually result in a reply with a copy of the test and permission to use it. If the author has moved from the original address, you may locate the current address through various directories and "Who's Who" type of books, or through computer generated literature searches.

Administrative aspects. Tests can also be distinguished by various aspects of their administration. For example, there are *group* vs. *individual* tests; group tests can be administered to a group of subjects at the same time and individual tests to one person only at one time. The Stanford-Binet test of intelligence is an individual test, whereas the SAT is a group test. Clinicians who deal with one client at a time generally prefer individual tests because these often yield observational data in addition to a test score; researchers often need to test large groups of subjects in minimum time and may prefer group tests (there are of course, many exceptions to this statement). A group test can be administered to one individual; sometimes, an individual test can be modified so it can be administered to a group.

Tests can also be classified as *speed* vs. *power* tests. Speed tests have a time limit that affects performance; for example, you might be given a page of printed text and asked to cross out all the "e's" in 25 seconds. How many you cross out will

be a function of how fast you respond. A power test, on the other hand, is designed to measure how well you can do and so either may have no time limit or a time limit of convenience (a 50-minute hour) that ordinarily does not affect performance. The time limits on speed tests are usually set so that only 50% of the applicants are able to attempt every item. Time limits on power tests are set so that about 90% of the applicants can attempt all items.

Another administrative distinction is whether a test is a *secure* test or not. For example, the SAT is commercially published but is ordinarily not made available even to researchers. Many tests that are used in industry for personnel selection are secure tests whose utility could be compromised if they were made public. Sometimes only the scoring key is confidential, rather than the items themselves.

A final distinction from an administrative point of view is how *invasive* a test is. A questionnaire that asks about one's sexual behaviors is ordinarily more invasive than a test of arithmetic; a test completed by the subject is usually more invasive than a report of an observer, who may report the observations without even the subject's awareness.

The medium. Tests differ widely in the materials used, and so we can distinguish tests on this basis. Probably, the majority of tests are *paper-and-pencil* tests that involve some set of printed questions and require a written response, such as marking a multiple answer sheet. Other tests are *performance* tests that perhaps require the manipulation of wooden blocks or the placement of puzzle pieces in correct juxtaposition. Still other tests involve *physiological* measures such as the galvanic skin response, the basis of the polygraph (lie detector) machine. Increasing numbers of tests are now available for computer administration and this may become a popular category.

Item structure. Another way to classify tests, which overlaps with the approaches already mentioned, is through their item structure. Test items can be placed on a continuum from objective to subjective. At the objective end, we have multiple-choice items; at the subjective end, we have the type of open-ended questions that clinical psychologists and psychiatrists ask, such as "tell me more," "how do you feel about that?" and "tell me about yourself." In between, we have countless variations such as matching items (closer to the objective pole) and essay questions (closer to the subjective pole). Objective items are easy to score and to manipulate statistically, but individually reveal little other than that the person answered correctly or incorrectly. Subjective items are difficult and sometimes impossible to quantify, but can be quite a revealing and rich source of information.

Another possible distinction in item structure is whether the items are *verbal* in nature or require *performance*. Vocabulary and math items are labeled verbal because they are composed of verbal elements; building a block tower is a performance item.

Area of assessment. Tests can also be classified according to the area of assessment. For example, there are intelligence tests, personality questionnaires, tests of achievement, career-interest tests, tests of reading, tests of neuropsychological functioning, and so on. The MMY uses 16 such categories. These are not necessarily mutually exclusive categories, and many of them can be further subdivided. For example, tests of personality could be further categorized into introversion-extraversion, leadership, masculinity-femininity, and so on.

In this textbook, we look at five major categories of tests:

1. Personality tests, which have played a major role in the development of psychological testing, both in its acceptance and criticism. Personality represents a major area of human functioning for social-behavioral scientists and lay persons alike;
2. Tests of cognitive abilities, not only traditional intelligence tests, but other dimensions of cognitive or intellectual functioning. In some ways, cognitive psychology represents a major new emphasis in psychology which has had a significant impact on all aspects of psychology both as a science and as an applied field;
3. Tests of attitudes, values, and interests, three areas that psychometrically overlap, and also offer lots of basic testing lessons;
4. Tests of psychopathology, primarily those used by clinicians and researchers to study the field of mental illness; and

5. Tests that assess normal and positive functioning, such as creativity, competence, and self-esteem.

Test function. Tests can also be categorized depending upon their function. Some tests are used to *diagnose* present conditions. (Does the client have a character disorder? Is the client depressed?) Other tests are used to make *predictions.* (Will this person do well in college? Is this client likely to attempt suicide?) Other tests are used in *selection* procedures, which basically involve accepting or not accepting a candidate, as in admission to graduate school. Some tests are used for *placement* purposes – candidates who have been accepted are placed in a particular "treatment." For example, entering students at a university may be placed in different level writing courses depending upon their performance in a writing exam. A battery of tests may be used to make such a placement decision or to assess which of several alternatives is most appropriate for the particular client – here the term typically used is *classification* (note that this term has both a broader meaning and a narrower meaning). Some tests are used for *screening* purposes; the term screening implies a rapid and rough procedure. Some tests are used for *certification*, usually related to some legal standard; thus passing a driving test certifies that the person has, at the very least, a minimum proficiency and is allowed to drive an automobile.

Score interpretation. Yet another classification can be developed on the basis of how scores on a test are interpreted. We can compare the score that an individual obtains with the scores of a group of individuals who also took the same test. This is called a *norm-reference* because we refer to norms to give a particular score meaning; for most tests, scores are interpreted in this manner. We can also give meaning to a score by comparing that score to a decision rule called a *criterion*, so this would be a *criterion-reference*. For example, when you took a driving test (either written and/or road), the examiner did not say, "Congratulations your score is two standard deviations above the mean." You either passed or failed based upon some predetermined criterion that may or may not have been explicitly stated. Note that norm-reference and criterion-reference refer not to the test but to how the score or performance is interpreted. The same test could yield either or both score interpretations.

Another distinction that can be made is whether the measurement provided by the test is *normative* or *ipsative*, that is, whether the standard of comparison reflects the behavior of others or of the client. Consider a 100-item vocabulary test that we administer to Marisa, and she obtains a score of 82. To make sense of that score, we compare her score with some normative data – for example, the average score of similar-aged college students. Now consider a questionnaire that asks Marisa to decide which of two values is more important to her: "Is it more important for you to have (1) a good paying job, or (2) freedom to do what you wish." We could compare her choice with that of others, but in effect we have simply asked her to rank two items in terms of her own preferences or her own behavior; in most cases it would not be legitimate to compare her ranking with those of others. She may prefer choice number 2, but not by much, whereas for me choice number 2 is a very strong preference.

One way of defining ipsative is that the scores on the scale must sum to a constant. For example, if you are presented with a set of six ice cream flavors to rank order as to preference, no matter whether your first preference is "crunchy caramel" or "Bohemian tutti-frutti," the sum of your six preferences will be 21 (1+2+3+4+5+6). On the other hand, if you were asked to rate each flavor independently on a 6-point scale, you could rate all of them high or all of them low; this would be a normative scale. Another way to define ipsative is to focus on the idea that in ipsative measurement, the mean is that of the individual, whereas in normative measurement the mean is that of the group. Ipsative measurement is found in personality assessment; we look at a technique called Q sort in Chapter 18. Block (1957) found that ipsative and normative ratings of personality were quite equivalent.

Another classificatory approach involves whether the responses made to the test are interpreted *psychometrically* or *impressionistically.* If the responses are scored and the scores interpreted on the basis of available norms and/or research data, then the process is a psychometric one. If instead the tester looks at the responses carefully on the basis of his/her expertise and

creates a psychological portrait of the client, that process is called impressionistic. Sometimes the two are combined; for example, clinicians who use the Minnesota Multiphasic Personality Inventory (MMPI), score the test and plot the scores on a profile, and then use the profile to translate their impressions into diagnostic and characterological statements. Impressionistic testing is more prevalent in clinical diagnosis and the assessment of psychodynamic functioning than, say, in assessing academic achievement or mechanical aptitude.

Self-report versus observer. Many tests are *self-report* tests where the client answers questions about his/her own behavior, preferences, values, etc. However, some tests require judging someone else; for example, a manager might rate each of several subordinates on promptness, independence, good working habits, and so on.

Maximal vs. typical performance. Yet another distinction is whether a test assesses *maximal performance* (how well a person can do) or *typical performance* (how well the person typically does) (Cronbach, 1970). Tests of maximal performance usually include achievement and aptitude tests and typically based on items that have a correct answer. Typical performance tests include personality inventories, attitude scales, and opinion questionnaires, for which there are no correct answers.

Age range. We can classify tests according to the age range for which they are most appropriate. The Stanford-Binet, for example, is appropriate for children but less so for adults; the SAT is appropriate for adolescents and young adults but not for children. Tests are used with a wide variety of clients and we focus particularly on children (Chapter 9), the elderly (Chapter 10), minorities and individuals in different cultures (Chapter 11), and the handicapped (Chapter 12).

Type of setting. Finally, we can classify tests according to the setting in which they are primarily used. Tests are used in a wide variety of settings, but the most prevalent are school settings (Chapter 13), occupational and military settings (Chapter 14), and "mental health" settings such as clinics, courts of law, and prisons (Chapter 15).

The NOIR system. One classificatory schema that has found wide acceptance is to classify tests according to their measurement properties. All measuring instruments, whether a psychological test, an automobile speedometer, a yardstick, or a bathroom scale, can be classified into one of four types based on the numerical properties of the instrument:

1. *Nominal scales.* Here the numbers are used merely as labels, without any inherent numerical property. For example, the numbers on the uniforms of football players represent such a use, with the numbers useful to distinguish one player from another, but not indicative of any numerical property – number 26 is not necessarily twice as good as number 13, and number 92 is not necessarily better or worse than number 91. In psychological testing, we sometimes code such variables as religious preference by assigning numbers to preferences, such as 1 to Protestant, 2 to Catholic, 3 to Jewish, and so on. This does not imply that being a Protestant is twice as good as being a Catholic, or that a Protestant plus a Catholic equal a Jew. Clearly, nominal scales represent a rather low level of measurement, and we should not apply to these scales statistical procedures such as computing a mean.
2. *Ordinal scales.* These are the result of ranking. Thus if you are presented with a list of ten cities and asked to rank them as to favorite vacation site, you have an ordinal scale. Note that the results of an ordinal scale indicate rankings but not differences in such rankings. Mazatlan in Mexico may be your first choice, with Palm Springs a close second; but Toledo, your third choice, may be a "distant" third choice.
3. *Interval scales.* These use numbers in such a way that the distance among different scores are based on equal units, but the zero point is arbitrary. Let's translate that into English by considering the measurement of temperature. The difference between 70 and 75 degrees is five units, which is the same difference as between 48 and 53 degrees. Each degree on our thermometer is equal in size. Note however that the zero point, although very meaningful, is in fact arbitrary; zero refers to the freezing of water at sea level – we could have chosen the freezing point of soda on top of Mount McKinley or some other standard. Because the zero point is arbitrary we

cannot make ratios, and we cannot say that a temperature of 100 degrees is twice as hot as a temperature of 50 degrees.

Let's consider a more psychological example. We have a 100-item multiple-choice vocabulary test composed of items such as:

cat = (a) feline, (b) canine, (c) aquiline, (d) asinine

Each item is worth 1 point and we find that Susan obtains a score of 80 and Barbara, a score of 40. Clearly, Susan's performance on the test is better than Barbara's, but is it twice as good? What if the vocabulary test had contained ten additional easy items that both Susan and Barbara had answered correctly; now Susan's score would have been 90 and Barbara's score 50, and clearly 90 is not twice 50. A zero score on this test does not mean that the person has zero vocabulary, but simply that they did not answer any of the items correctly – thus the zero is arbitrary and we cannot arrive at any conclusions that are based on ratios.

In this connection, I should point out that we might question whether our vocabulary test is in fact an interval scale. We score it as if it were, by assigning equal weights to each item, but are the items really equal? Most likely no, since some of the vocabulary items might be easier and some might be more difficult. I could, of course, empirically determine their difficulty level (we discuss this in Chapter 2) and score them appropriately (a real difficult item might receive 9 points, a medium difficulty item 5, and so on), or I could use only items that are of approximately equal difficulty or, as is often done, I can assume (typically incorrectly) that I have an interval scale.

4. *Ratio scales.* Finally, we have ratio scales that not only have equal intervals but also have a true zero. The Kelvin scale of temperature, which chemists use, is a ratio scale and on that scale a temperature of 200 is indeed twice as hot as a temperature of 100. There are probably no psychological tests that are true ratio scales, but most approximate interval scales; that is, they really are ordinal scales but we treat them as if they were interval scales. However, newer theoretical models known as item-response theory (e.g., Lord, 1980; Lord & Novick, 1968; Rasch, 1966; D. J. Weiss & Davison, 1981) have resulted in ways of developing tests said to be ratio scales.

ETHICAL STANDARDS

Tests are tools used by professionals to make what may possibly be some serious decisions about a client; thus both tests and the decision process involve a variety of ethical considerations to make sure that the decisions made are in the best interest of all concerned and that the process is carried out in a professional manner. There are serious concerns, on the part of both psychologists and lay people, about the nature of psychological testing and its potential misuse, as well as demands for increased use of tests.

APA ethics code. The American Psychological Association has since 1953 published and revised ethical standards, with the most recent publication of *Ethical Principles of Psychologists and Code of Conduct* in 1992. This code of ethics also governs, both implicitly and explicitly, a psychologist's use of psychological tests.

The Ethics Code contains six general principles:

1. Competence: Psychologists maintain high standards of competence, including knowing their own limits of expertise. Applied to testing, this might suggest that it is unethical for the psychologist to use a test with which he or she is not familiar to make decisions about clients.

2. Integrity: Psychologists seek to act with integrity in all aspects of their professional roles. As a test author for example, a psychologist should not make unwarranted claims about a particular test.

3. Professional and scientific responsibility: Psychologists uphold professional standards of conduct. In psychological testing this might require knowing when test data can be useful and when it cannot. This means, in effect, that a practitioner using a test needs to be familiar with the research literature on that test.

4. Respect for people's rights and dignity: Psychologists respect the privacy and confidentiality of clients and have an awareness of cultural, religious, and other sources of individual differences. In psychological testing, this might include an awareness of when a test is appropriate for use with individuals who are from different cultures.

5. Concern for others' welfare: Psychologists are aware of situations where specific tests (for

example, ordered by the courts) may be detrimental to a particular client. How can these situations be resolved so that both the needs of society and the welfare of the individual are protected?

6. Social responsibility: Psychologists have professional and scientific responsibilities to community and society. With regard to psychological testing, this might cover counseling against the misuse of tests by the local school.

In addition to these six principles, there are specific ethical standards that cover eight categories, ranging from "General standards" to "Resolving ethical issues." The second category is titled, "Evaluation, assessment, or intervention" and is thus the area most explicitly related to testing; this category covers 10 specific standards:

1. Psychological procedures such as testing, evaluation, diagnosis, etc., should occur only within the context of a defined professional relationship.
2. Psychologists only use tests in appropriate ways.
3. Tests are to be developed using acceptable scientific procedures.
4. When tests are used, there should be familiarity with and awareness of the limitations imposed by psychometric issues, such as those discussed in this textbook.
5. Assessment results are to be interpreted in light of the limitations inherent in such procedures.
6. Unqualified persons should not use psychological assessment techniques.
7. Tests that are obsolete and outdated should not be used.
8. The purpose, norms, and other aspects of a test should be described accurately.
9. Appropriate explanations of test results should be given.
10. The integrity and security of tests should be maintained.

Standards for educational and psychological tests. In addition to the more general ethical standards discussed above, there are also specific standards for educational and psychological tests (American Educational Research Association, 1999), first published in 1954, and subsequently revised a number of times.

These standards are quite comprehensive and cover (1) technical issues of validity, reliability, norms, etc.; (2) professional standards for test use, such as in clinical and educational settings; (3) standards for particular applications such as testing linguistic minorities; and (4) standards that cover aspects of test administration, the rights of the test taker and so on.

In considering the ethical issues involved in psychological testing, three areas seem to be of paramount importance: informed consent, confidentiality, and privacy.

Informed consent means that the subject has been given the relevant information about the testing situation and, based on that information, consents to being tested. Obviously this is a theoretical standard that in practice requires careful and thoughtful application. Clearly, to inform a subject that the test to be taken is a measure of "interpersonal leadership" may result in a set to respond in a way that can distort and perhaps invalidate the test results. Similarly, most subjects would not understand the kind of technical information needed to scientifically evaluate a particular test. So typically, informed consent means that the subject has been told in general terms what the purpose of the test is, how the results will be used, and who will have access to the test protocol.

The issue of *confidentiality* is perhaps even more complex. Test results are typically considered *privileged communication* and are shared only with appropriate parties. But what is appropriate? Should the client have access to the actual test results elucidated in a test report? If the client is a minor, should parents or legal guardians have access to the information? What about the school principal? What if the client was tested unwillingly, when a court orders such testing for determination of psychological sanity, pathology that may pose a threat to others, or the risk of suicide, etc. When clients seek psychological testing on their own, for example a college student requesting career counseling at the college counseling center, the guidelines are fairly clear. Only the client and the professional have access to the test results, and any transmission of test results to a third party requires written consent on the part of the client. But real-life issues often have a way of becoming more complex.