Psychological Testing

AN INTRODUCTION

SECOND EDITION

GEORGE DOMINO • MARLA L. DOMINO

CAMBRIDGE www.cambridge.org/9780521861816

This page intentionally left blank

PSYCHOLOGICAL TESTING: AN INTRODUCTION

Second Edition

This book is an introductory text to the field of psychological testing primarily suitable for undergraduate students in psychology, education, business, and related fields. This book will also be of interest to graduate students who have not had prior exposure to psychological testing and to professionals such as lawyers who need to consult a useful source. *Psychological Testing* is clearly written, well organized, comprehensive, and replete with illustrative materials. In addition to the basic topics, the text covers in detail topics that are often neglected by other texts such as cross-cultural testing, the issue of faking tests, the impact of computers, and the use of tests to assess positive behaviors such as creativity.

George Domino is the former Director of Clinical Psychology and Professor of Psychology at the University of Arizona. He was also the former director of the Counseling Center and Professor of Psychology at Fordham University.

Marla L. Domino has a BA in Psychology, an MA in Criminal Law, and a PhD in Clinical Psychology specializing in Psychology and Law. She also completed a post-doctoral fellowship in Clinical-Forensic Psychology at the University of Massachusetts Medical School, Law and Psychiatry Program. She is currently the Chief Psychologist in the South Carolina Department of Mental Health's Forensic Evaluation Service and an assistant professor in the Department of Neuropsychiatry and Behavioral Sciences at the University of South Carolina. She was recently awarded by the South Carolina Department of Mental Health as Outstanding Employee of the Year in Forensics (2004).

SECOND EDITION

Psychological Testing

An Introduction

George Domino

University of Arizona

Marla L. Domino

Department of Mental Health, State of South Carolina

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521861816

First published in print format 2006

ISBN-13 978-0-511-21802-6 eBook (Adobe Reader)
ISBN-10 0-511-21802-8 eBook (Adobe Reader)

ISBN-13 978-0-521-86181-6 hardback
ISBN-10 0-521-86181-0 hardback

Contents

Preface

My first professional publication in 1963 was as a graduate student (with Harrison Gough) on a validational study of a culture-fair test. Since then, I have taught a course on psychological testing with fair regularity. At the same time, I have steadfastly refused to specialize and have had the opportunity to publish in several different areas, to work in management consulting, to be director of a counseling center and of a clinical psychology program, to establish an undergraduate honors program, and to be involved in a wide variety of projects with students in nursing, rehabilitation, education, social work, and other fields. In all of these activities, I have found psychological testing to be central and to be very challenging and exciting.

In this book, we have tried to convey the excitement associated with psychological testing and to teach basic principles through the use of concrete examples. When specific tests are mentioned, they are mentioned because they are used as an example to teach important basic principles, or in some instances, because they occupy a central/historical position. No attempt has been made to be exhaustive.

Much of what is contained in many testing textbooks is rather esoteric information, of use only to very few readers. For example, most textbooks include several formulas to compute interitem consistency. It has been our experience, however, that 99% of the students who take a course on testing will never have occasion to use such formulas, even if they enter a career in psychology or allied fields. The very few who might need to do such calculations will do them by computer or will know where to find the relevant formulas. It is the principle that is important, and that is what we have tried to emphasize.

Because of my varied experience in industry, in a counseling center, and other service-oriented settings, and also because as a clinically trained academic psychologist I have done a considerable amount of research, I have tried to cover both sides of the coin – the basic research-oriented issues and the application of tests in service-oriented settings. Thus Parts One and Two, the first eight chapters, serve as an introduction to basic concepts, issues, and approaches. Parts Three and Four, Chapters 9 through 15, have a much more applied focus. Finally, we have attempted to integrate both classical approaches and newer thinking about psychological testing.

The area of psychological testing is fairly well defined. I cannot imagine a textbook that does not discuss such topics as reliability, validity, and norms. Thus, what distinguishes one textbook from another is not so much its content but more a question of balance. For example, most textbooks continue to devote one or more chapters to projective techniques, even though their use and importance has decreased substantially. Projective techniques are important, not only from a historical perspective, but also for what they can teach us about basic issues in testing. In this text, they are discussed and illustrated, but as part of a chapter (see Chapter 15) within the broader context of testing in clinical settings. Most textbooks also have several chapters on intelligence testing, often devoting considerable space to such topics as the heritability of intelligence, theories of trait organization, longitudinal studies of intelligence, and similar topics. Such topics are of course important and fascinating,

but do they really belong in a textbook on psychological testing? If they do, then that means that some other topics more directly relevant to testing are omitted or given short shrift. In this textbook, we have chosen to focus on testing and to minimize the theoretical issues associated with intelligence, personality, etc., except where they may be needed to have a better understanding of testing approaches.

It is no surprise that computers have had (and continue to have) a major impact on psychological testing, and so an entire chapter of this book (Chapter 17) is devoted to this topic. There is also a vast body of literature and great student interest on the topic of faking, and here too an entire chapter (Chapter 16) has been devoted to this topic. Most textbooks begin with a historical chapter. We have chosen to place this chapter last, so the reader can better appreciate the historical background from a more knowledgeable point of view.

Finally, rather than writing a textbook about testing, we have attempted to write a textbook about testing the individual. We believe that most testing applications involve an attempt to use tests as a tool to better understand an individual, whether that person is a client in therapy, a college student seeking career or academic guidance, a business executive wishing to capitalize on strengths and improve on weaknesses, or a volunteer in a scientific experiment.

Acknowledgments

In my career as a psychologist, I have had the excellent fortune to be mentored, directly and indirectly, by three giants in the psychological testing field. The first is Harrison Gough, my mentor in graduate school at Berkeley, who showed me how useful and exciting psychological tests can be when applied to real-life problems. More importantly, Gough has continued to be not only a mentor but also a genuine model to be emulated both as a psychologist and as a human being. Much of my thinking and approach to testing, as well as my major interest in students at all levels, is a direct reflection of Gough's influence.

The second was Anne Anastasi, a treasured colleague at Fordham University, a generous friend, and the best chairperson I have ever worked with. Her textbook has been truly a model of scholarship and concise writing, the product of an extremely keen mind who advanced the field of psychological testing in many ways.

The third person was Lee J. Cronbach of Stanford University. My first undergraduate exposure to testing was through his textbook. In 1975, Cronbach wrote what is now a classic paper titled, "Beyond the two disciplines of scientific psychology" (*American Psychologist*, 1975, vol. 30, pp. 116–127), in which he argued that experimental psychology and the study of individual differences should be integrated. In that paper, Cronbach was kind enough to cite at some length two of my studies on college success as examples of this integration. Subsequently I was able to invite him to give a colloquium at the University of Arizona. My contacts with him were regrettably brief, but his writings greatly influenced my own thinking.

On a personal note, I thank Valerie, my wife of 40 years, for her love and support, and for being the best companion one could hope for in this voyage we call life. Our three children have been an enormous source of love and pride: Brian, currently a professor of philosophy at Miami University of Ohio; Marisa, a professor of health economics at the University of North Carolina, Chapel Hill; and Marla, chief forensic psychologist in the Department of Mental Health of South Carolina, and co-author of this edition. Zeno and Paolo, our two grandchildren, are unbelievably smart, handsome, and adorable and make grandparenting a joy. I have also been truly blessed with exceptional friends whose love and caring have enriched my life enormously.

George Domino
Tucson, AZ

An abundance of gratitude to my father for giving me the opportunity to collaborate with one of the greatest psychologists ever known. And an immeasurable amount of love and respect to my heroes – my Dad and Mom. I would also like to thank my mentor and friend, Stan Brodsky, whose professional accomplishments are only surpassed by his warmth, kindness, and generous soul.

Marla Domino
Columbia, SC

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.

The right to *privacy* basically concerns the willingness of a person to share with others personal information, whether that information be factual or involve feelings and attitudes. In many tests, especially personality tests, the subject is asked to share what may be very personal information, occasionally without realizing that such sharing is taking place. At the same time, the subject cannot be instructed that, "if you answer true to item #17, I will take that as evidence that you are introverted."

What is or is not invasion of privacy may be a function of a number of aspects. A person seeking the help of a sex therapist may well expect and understand the need for some very personal questions about his or her sex life, while a student seeking career counseling would not expect to be questioned about such behavior (for a detailed analysis of privacy as it relates to psychological testing see Ruebhausen & Brim, 1966; for some interesting views on privacy, including Congressional hearings, see the November 1965 and May 1966 issues of the *American Psychologist*).

Mention might also be made of *feedback,* providing and explaining test results to the client. Pope (1992) suggests that feedback may be the most neglected aspect of assessment, and describes feedback as a dynamic, interactive process, rather than a passive, information-giving process.

The concern for ethical behavior is a pervasive aspect of the psychological profession, but one that lay people often are not aware of. Students, for example, at times do not realize that their requests ("can I have a copy of the XYZ intelligence test to assess my little brother") could involve unethical behavior.

In addition to the two major sets of ethical standards discussed above, there are other pertinent documents. For example, there are guidelines for providers of psychological services to members of populations whose ethnic, linguistic, or cultural background are diverse (APA, 1993), which include at least one explicit statement about the application of tests to such individuals, and there are guidelines for the disclosure of test data (APA, 1996). All of these documents are the result of hard and continuing work on the part of many professional organizations.

Test levels. If one considers tests as tools to be used by professionals trained in their use, then it becomes quite understandable why tests should not be readily available to unqualified users. In fact, the APA proposed many years ago a rating system of three categories of tests: level A tests require minimal training, level B tests require some advanced training, and level C tests require substantial professional expertise. These guidelines are followed by many test publishers who often require that prospective customers fill out a registration form indicating their level of expertise to purchase specific tests.

There is an additional reason why the availability of tests needs to be controlled and that is for *security.* A test score should reflect the dimension being measured, for example, knowledge of elementary geography, rather than some other process such as knowledge of the right answers. As indicated earlier, some tests are highly secured and their use is tightly controlled; for example tests like the SAT or the GRE are available only to those involved in their administration, and a strict accounting of each test booklet is required. Other tests are readily available, and their item content can sometimes be found in professional journals or other library documents.

INFORMATION ABOUT TESTS

It would be nice if there were one central source, one section of the library, that would give us all the information we needed about a particular test – but there isn't. You should realize that libraries do not ordinarily carry specimen copies of tests. Not only are there too many of them and they easily get out of date, but such a depository would raise some serious ethical questions. There may be offices on a college campus, such as the Counseling Center or the Clinical Psychology program, that have a collection of tests with scoring keys, manuals, etc., but these are not meant for public use. Information about specific tests is scattered quite widely, and often such a search is time consuming and requires patience as well as knowledge about available resources. The following steps can be of assistance:

1. The first step in obtaining information about a specific test is to consult the MMY. If the test is commercially published and has been reviewed

in the MMY, then our job will be infinitely easier; the MMY will give us the publishers' address and we can write for a catalog or information. It may also list references that we can consult, typically journal articles that are relevant. But what if the test is not listed in the MMY?

2. A second step is to check the original citation where mention of the particular test is made. For example, we may be reading a study by Jones which used the Smith Anxiety Scale; typically Jones will provide a reference for the Smith Anxiety Scale. We can locate that reference and then write to Smith for information about that scale. Smith's address will hopefully be listed in Smith's article, or we can look up Smith's address in directories such as the American Psychological Association Directory or a "Who's Who."

3. A third step is to conduct a computer literature search. If the test is well known we might obtain quite a few citations. If the test is somewhat more obscure, we might miss the available information. Keep in mind that currently most computer literature searches only go back a limited number of years.

4. If steps 2 and 3 give us some citations, we might locate these citations in the *Social Sciences Citation Index*; for example, if we locate the citation to the Smith Anxiety Scale, the Science Citation Index will tell us which articles use the Smith citation in their list of references. Presumably these articles might be of interest to us.

5. Suppose instead of a specific test we are interested in locating a scale of anxiety that we might use in our own study, or we want to see some of the various ways in which anxiety is assessed. In such a case, we would again first check the MMY to see what is available and take some or all of the following steps.

6. Search the literature for articles/studies on anxiety to see what instruments have been used. We will quickly observe that there are several instruments that seem to be quite popularly used and many others that are not.

7. We might repeat steps 2 and 3 above.

8. If the test is a major one, whether commercially published or not, we can consult the library to see what books have been written about that particular test. There are many books available on such tests as the Rorschach, the Minnesota Multiphasic Personality Inventory, and the Stanford-Binet (e.g., J. R. Graham, 1990; Knapp, 1976; Megargee, 1972; Snider & Osgood, 1969).

9. Another source of information is *Educational Testing Service* (ETS), the publisher of most of the college and professional school entrance exams. ETS has an extensive test library of more than 18,000 tests and, for a fee, can provide information. Also, ETS has published annually since 1975 *Tests in Microfiche*, sets of indices and abstracts to various research instruments; some libraries subscribe to these.

10. A number of journals such as the *Journal of Counseling and Development* and the *Journal of Psychoeducational Assessment*, routinely publish test reviews.

11. Finally, many books are collections of test reviews, test descriptions, etc., and provide useful information on a variety of tests. Some of these are listed in Table 1.1.

SUMMARY

A test can be defined as an objective and standardized measure of a sample of behavior. We can also consider a test as an experiment, an interview, or a tool. Tests can be used as part of psychological assessment, and are used for classification, self-understanding, program evaluation, and scientific inquiry. From the viewpoint of tests as an experiment, we need to pay attention to four categories of variables that can influence the outcome: the method of administration, situational variables, experimenter variables, and subject variables. Tests are used for decision making, although the content of a test need not coincide with the area of behavior that is assessed, other than to be empirically related.

Tests can be categorized according to whether they are commercially published or not administrative aspects such as group versus individual tests, the type of item, the area of assessment, the function of the test, how scores are interpreted, whether the test is a self-report or not, the age range and type of client, and the measurement properties.

Ethical standards relate to testing and the issues of informed consent, confidentiality, and privacy. There are many sources of information about tests available through libraries, associations, and other avenues of research.

Table 1–1. Sources for test information

Andrulis, R. S. (1977). *Adult assessment.* Springfield, IL: Charles C Thomas.

Six major categories of tests are listed, including aptitude and achievement, personality, attitudes, and personal performance.

Beere, C. A. (1979). *Women and women's issues: A handbook of tests and measures.* San Francisco: Jossey-Bass.

This handbook covers such topics as sex roles, gender knowledge, and attitudes toward women's issues, and gives detailed information on a variety of scales.

Chun, K. T. et al. (1975). *Measures for psychological assessment: A guide to 3000 original sources and their applications.* Ann Arbor: University of Michigan.

An old but still useful source for measures of mental health.

Compton, C. (1980). *A guide to 65 tests for special education.* Belmont, California: Fearon Education.

A review of tests relevant to special education.

Comrey, A. L., Backer, T. F., & Glaser, E. M. (1973). *A sourcebook for mental health measures.* Los Angeles: Human Interaction Research Institute.

A series of abstracts on about 1,100 lesser known measures in areas ranging from alcoholism through mental health, all the way to vocational tests.

Corcoran, K., & Fischer, J. (1987). *Measures for clinical practice: A sourcebook.* New York: Free Press.

A review of a wide variety of measures to assess various clinical problems.

Fredman, N., & Sherman, R. (1987). *Handbook of measurements for marriage and family therapy.* New York: Bruner Mazel.

A review of 31 of the more widely used paper-and-pencil instruments in the area of marriage and family therapy.

Goldman, B. A., & Saunders, J. L. (1974). *Directory of unpublished experimental mental measures, Vol. 1–4.* New York: Behavioral Publications.

The first volume contains a listing of 339 unpublished tests that were cited in the 1970 issues of a group of journals. Limited information is given on each one.

Hogan, J., & Hogan, R. (Eds.) (1990). *Business and industry testing.* Austin, TX: Pro-ed.

A review of tests especially pertinent to the world of work, such as intelligence, personality, biodata, and integrity tests.

Johnson, O. G. (1970; 1976). *Tests and measurements in child development.* San Francisco: Jossey-Bass.

The two volumes cover unpublished tests for use with children.

Keyser, D. J., & Sweetland, R. C. (Eds.) (1984). *Test critiques.* Kansas City: Test Corporation of America.

This is a continuing series that reviews the most frequently used tests, with reviews written by test experts, and quite detailed in their coverage. The publisher, Test Corporation of America, publishes a variety of books on testing.

Lake, D. G., Miles, M. B., & Earle, R. B., Jr. (1973). *Measuring human behavior.* New York: Teachers College Press.

A review of 84 different instruments and 20 compendia of instruments; outdated but still useful.

Mangen, D. J., & Peterson, W. A. (Eds.) (1982). *Research instruments in social gerontology; 2 volumes.* Minneapolis: University of Minnesota Press.

If you are interested in measurement of the elderly this is an excellent source. For each topic, for example death and dying, there is a brief overall discussion, some brief commentary on the various instruments, a table of the cited instruments, a detailed description of each instrument, and a copy of each instrument.

McReynolds, P. (Ed.) (1968). *Advances in psychological assessment.* Palo Alto: Science and Behavior Books.

This is an excellent series of books, the first one published in 1968, each book consisting of a series of chapters on assessment topics, ranging from reviews of specific tests like the Rorschach and the California Psychological Inventory (CPI), to topic areas like the assessment of anxiety, panic disorder, and adolescent suicide.

Newmark, C. S. (Ed.) (1985; 1989), *Major psychological assessment instruments*, volumes I and II. Boston: Allyn & Bacon.

A nice review of the most widely used tests in current psychological assessment, the volumes give detailed information about the construction, administration, interpretation, and status of these tests.

Reeder, L. G., Ramacher, L., & Gorelnik, S. (1976). *Handbook of scales and indices of health behavior.* Pacific Palisades, CA.: Goodyear Publishing.

A somewhat outdated but still useful source.

Reichelt, P. A. (1983). Location and utilization of available behavioral measurement instruments. *Professional Psychology, 14*, 341–356.

Includes an annotated bibliography of various compendia of tests.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (Eds.) (1990). *Measures of personality and social psychological attitudes.* San Diego, CA.: Academic Press.

Robinson and his colleagues at the Institute for Social Research (University of Michigan) have published a number of volumes summarizing measures of political attitudes (1968), occupational attitudes and characteristics (1969), and social-psychological attitudes (1969, 1973, & 1991).

Schutte, N. S., & Malouff, J. M. (1995). *Sourcebook of adult assessment strategies.* New York: Plenum Press.

A collection of scales, their description and evaluation, to assess psychopathology, following the diagnostic categories of the Diagnostic and Statistical Manual of Mental Disorders.

Table 1–1. (*continued*)

Shaw, M. E., & Wright, J. M. (1967). *Scales for the measurement of attitudes.* New York: McGraw-Hill.

An old but still useful reference for attitude scales. Each scale is reviewed in some detail, with the actual scale items given.

Southworth, L. E., Burr, R. L., & Cox, A. E. (1981). *Screening and evaluating the young child: A handbook of instruments to use from infancy to six years.* Springfield, IL: Charles C Thomas.

A compendium of preschool screening instruments, but without any evaluation of these instruments.

Straus, M. A. (1969). *Family measurement techniques.* Minneapolis: University of Minnesota Press.

A review of instruments reported in the psychological and sociological literature from 1935 to 1965.

Sweetland, R. C., & Keyser, D. J. (Eds.) (1983). *Tests: A comprehensive reference for assessments in psychology, education, and business.* Kansas City: Test Corporation of America.

This is the first edition of what has become a continuing series. In this particular volume, over 3,000 tests, both commercially available and unpublished, are given a brief thumbnail sketches.

Walker, D. K. (1973). *Socioemotional measures for preschool and kindergarten children.* San Francisco: Jossey-Bass.

A review of 143 measures covering such areas as personality, self-concept, attitudes, and social skills.

Woody, R. H. (Ed.) (1980). *Encyclopedia of clinical assessment. 2 vols.* San Francisco: Jossey-Bass.

This is an excellent, though now outdated, overview of clinical assessment; The 91 chapters cover a wide variety of tests ranging from measures of normality to moral reasoning, anxiety, and pain.

SUGGESTED READINGS

Dailey, C. A. (1953). The practical utility of the clinical report. *Journal of Consulting Psychology, 17*, 297–302.

An interesting study that tried to quantify how clinical procedures, based on tests, contribute to the decisions made about patients.

Fremer, J., Diamond, E. E., & Camara, W. J. (1989). Developing a code of fair testing practices in education. *American Psychologist, 44*, 1062–1067.

A brief historical introduction to a series of conferences that eventuated into a code of fair testing practices, and the code itself.

Lorge, I. (1951). The fundamental nature of measurement. In. E. F. Lindquist (Ed.), *Educational Measurement*, pp. 533–559. Washington, D.C.: American Council on Education.

An excellent overview of measurement, including the NOIR system.

Willingham, W. W. (Ed.). (1967). Invasion of privacy in research and testing. *Journal of Educational Measurement, 4*, No. 1 supplement.

An interesting series of papers reflecting the long standing ethical concerns involved in testing.

Wolfle, D. (1960). Diversity of Talent. *American Psychologist, 15*, 535–545.

An old but still interesting article that illustrates the need for broader use of tests.

DISCUSSION QUESTIONS

1. What has been your experience with tests?
2. How would you design a study to assess whether a situational variable can alter test performance?
3. Why not admit everyone who wants to enter medical school, graduate programs in business, law school, etc.?
4. After you have looked at the MMY in the library, discuss ways in which it could be improved.
5. If you were to go to the University's Counseling Center to take a career interest test, how would you expect the results to be handled? (e.g., should your parents receive a copy?).

2 Test Construction, Administration, and Interpretation

AIM This chapter looks at three basic questions: (1) How are tests constructed? (2) What are the basic principles involved in administering a test? and (3) How can we make sense of a test score?

CONSTRUCTING A TEST

How does one go about constructing a test? Because there are all sorts of tests, there are also all sorts of ways to construct such tests, and there is no one approved or sure-fire method of doing this. In general, however, test construction involves a sequence of 8 steps, with lots of exceptions to this sequence.

1. Identify a need. The first step is the identification of a need that a test may be able to fulfill. A school system may require an intelligence test that can be administered to children of various ethnic backgrounds in a group setting; a literature search may indicate that what is available doesn't fit the particular situation. A doctoral student may need a scale to measure "depth of emotion" and may not find such a scale. A researcher may want to translate some of Freud's insights about "ego defense" mechanisms into a scale that measures their use. A psychologist may want to improve current measures of leadership by incorporating new theoretical insights, and therefore develops a new scale. Another psychologist likes a currently available scale of depression, but thinks it is too long and decides to develop a shorter version. A test company decides to come out with a new career interest test to compete with what is already available on the market. So the need may be a very practical one (we need a scale to evaluate patients' improvement in psychotherapy), or it may be very theoretical (a scale to assess "anomie" or "ego-strength"). Often, the need may be simply a desire to improve what is already available or to come up with one's own creation.

2. The role of theory. Every test that is developed is implicitly or explicitly influenced or guided by the theory or theories held by the test constructor. The theory may be very explicit and formal. Sigmund Freud, Carl Rogers, Emile Durkheim, Erik Erikson, and others have all developed detailed theories about human behavior or some aspect of it, and a practitioner of one of these theories would be heavily and knowingly influenced by that theory in constructing a test. For example, most probably only a Freudian would construct a scale to measure "id, ego, and superego functioning" and only a "Durkheimite" would develop a scale to measure "anomie." These concepts are embedded in their respective theories and their meaning as measurement variables derives from the theoretical framework in which they are embedded.

A theory might also yield some very specific guidelines. For example, a theory of depression might suggest that depression is a disturbance in four areas of functioning: self-esteem, social support, disturbances in sleep, and negative affect. Such a schema would then dictate that the measure of depression assess each of these areas.

The theory may also be less explicit and not well formalized. The test constructor may, for example, view depression as a troublesome state composed of negative feelings toward oneself, a reduction in such activities as eating and talking with friends, and an increase in negative thoughts and suicide ideation. The point is that a test is not created in a vacuum, nor is it produced by a machine as a yardstick might be. The creation of a test is intrinsically related to the person doing the creating and, more specifically, to that person's theoretical views. Even a test that is said to be "empirically" developed, that is, developed on the basis of observation or real-life behavior (how do depressed people answer a questionnaire about depression), is still influenced by theory.

Not all psychologists agree. R. B. Cattell (1986), for example, argues that most tests lack a true theoretical basis, that their validity is due to work done *after* their construction rather than before, and that they lack good initial theoretical construction. Embretson (1985b) similarly argues that although current efforts have produced tests that do well at predicting behavior, the link between these tests and psychological theory is weak and often nonexistent.

3. Practical choices. Let's assume that I have identified as a need the development of a scale designed to assess the eight stages of life that Erik Erikson discusses (Erikson, 1963; 1982; see G. Domino & Affonso, 1990, for the actual scale). There are a number of practical choices that now need to be made. For example, what format will the items have? Will they be true-false, multiple choice, 7-point rating scales, etc.? Will there be a time limit or not? Will the responses be given on a separate answer sheet? Will the response sheet be machine scored? Will my instrument be a quick "screening" instrument or will it give comprehensive coverage for each life stage? Will I need to incorporate some mechanism to assess honesty of response? Will my instrument be designed for group administration?

4. Pool of items. The next step is to develop a *table of specifications*, much like the blueprint needed to construct a house. This table of specifications would indicate the subtopics to be covered by the proposed test (in our example, the eight life stages), perhaps their relative importance (are they all of equal importance?), and how many items each subtopic will contribute to the overall test (I might decide, for example, that each of the eight stages should be assessed by 15 items, thus yielding a total test of 120 items). This table of specifications may reflect not only my own thinking, but the theoretical notions present in the literature, other tests that are available on this topic, and the thinking of colleagues and experts. Test companies that develop educational tests such as achievement batteries often go to great lengths in developing such a table of specifications by consulting experts, either individually or in group conferences; the construction of these tests often represent major efforts of many individuals, at a high cost beyond the reach of any one person.

The table of specifications may be very formal or very informal, or sometimes absent, but leads to the writing or assembling of potential items. These items may be the result of the test constructor's own creativity, they may be obtained from experts, from other measures already available, from a reading of the pertinent literature, from observations and interviews with clients, and many other sources. Writing good test items is both an art and a science and is not easily achieved. I suspect you have taken many instructor made tests where the items were not clear, the correct answers were quite obvious, or the items focused on some insignificant aspects of your coursework. Usually, the classroom instructor writes items and uses most of them. The professional test constructor knows that the initial pool of items needs to be at a minimum four or five times as large as the number of items actually needed.

5. Tryouts and refinement. The initial pool of items will probably be large and rather unrefined. Items may be near duplications of each other, perhaps not clearly written or understood. The intent of this step is to refine the pool of items to a smaller but usable pool. To do this, we might ask colleagues (and/or enemies) to criticize, the items, or we might administer them to a captive class of psychology majors to review and identify items that may not be clearly written. Sometimes, *pilot testing* is used where a preliminary form is administered to a sample of subjects to determine

whether there are any glitches, etc. Such pilot testing might involve asking the subjects to think aloud as they answer each item or to provide feedback as to whether the instructions are clear, the items interesting, and so on. We may also do some preliminary statistical work and assemble the test for a trial run called a *pretest.* For example, if I were developing a scale to measure depression, I might administer my pool of items (say 250) to groups of depressed and nondepressed people and then carry out *item analyses* to see which items in fact differentiate the two groups. For example, to the item "I am feeling blue" I might expect significantly more depressed people to answer "true" than nondepressed people. I might then retain the 100 items that seem to work best statistically, write each item on a 3×5 card, and sort these cards into categories according to their content; such as all the items dealing with sleep disturbances in one pile, all the items dealing with feelings in a separate pile, and so on. This sorting might indicate that we have too many items of one kind and not enough of another, so I might remove some of the excess items and write some new ones for the underrepresented category. Incidentally, this process is known as *content analysis* (see Gottschalk & Gleser, 1969). This step then, consists of a series of procedures, some requiring logical analysis, others statistical analysis, that are often repeated several times, until the initial pool of items has been reduced to manageable size, and all the evidence indicates that our test is working the way we wish it to.

6. Reliability and validity. Once we have refined our pool of items to manageable size, and have done the preliminary work of the above steps, we need to establish that our measuring instrument is *reliable*, that is, consistent, and measures what we set out to measure, that is, the test is *valid.* These two concepts are so basic and important that we devote an entire chapter to them (see Chapter 3). If we do not have reliability and validity, then our pool of items is not a measuring instrument, and it is precisely this that distinguishes the instruments psychologists use from those "questionnaires" that are published in popular magazines to determine whether a person is a "good lover," "financially responsible," or a "born leader."

7. Standardization and norms. Once we have established that our instrument is both reliable and valid, we need to standardize the instrument and develop norms. To *standardize* means that the administration, time limits, scoring procedures, and so on are all carefully spelled out so that no matter who administers the test, the procedure is the same. Obviously, if I administer an intelligence test and use a 30-minute time limit, and you administer the same test with a 2-hour time limit, the results will not be comparable. It might surprise you to know that there are some tests both commercially published and not that are not well standardized and may even lack instructions for administration.

Let's assume that you answer my vocabulary test, and you obtain a score of 86. What does that 86 mean? You might be tempted to conclude that 86 out of 100 is fairly good, until I tell you that second graders average 95 out of 100. You'll recall that 86 and 95 are called *raw scores*, which in psychology are often meaningless. We need to give meaning to raw scores by changing them into *derived* scores; but that may not be enough. We also need to be able to compare an individual's performance on a test with the performance of a group of individuals; that information is what we mean by *norms.* The information may be limited to the mean and standard deviation for a particular group or for many different groups, or it may be sufficiently detailed to allow the translation of a specific raw score into a derived score such as percentiles, T scores, z scores, IQ units, and so on.

The test constructor then administers the test to one or more groups, and computes some basic descriptive statistics to be used as norms, or normative information. Obviously, whether the normative group consists of 10 students from a community college, 600 psychiatric patients, or 8,000 sixth graders, will make quite a difference; test norms are not absolute but simply represent the performance of a particular sample at a particular point in time. The sample should be large enough that we feel comfortable with its size, although "large enough" cannot be answered by a specific number; simply because a sample is large, does not guarantee that it is representative. The sample should be representative of the population to which we generalize, so that an achievement test for use by fifth graders should have norms based

on fifth graders. It is not unusual for achievement tests used in school systems to have normative samples in the tens of thousands, chosen to be representative on the basis of census data or other guiding principles, but for most tests the sample size is often in the hundreds or smaller. The sample should be clearly defined also so that the test user can assess its adequacy – was the sample a captive group of introductory psychology students, or a "random" sample representative of many majors? Was the sample selected on specific characteristics such as income and age, to be representative of the national population? How were the subjects selected?

8. Further refinements. Once a test is made available, either commercially or to other researchers, it often undergoes refinements and revisions. Well-known tests such as the Stanford-Binet have undergone several revisions, sometimes quite major and sometimes minor. Sometimes the changes reflect additional scientific knowledge, and sometimes societal changes, as in our greater awareness of gender bias in language.

One type of revision that often occurs is the development of a *short form* of the original test. Typically, a different author takes the original test, administers it to a group of subjects, and shows by various statistical procedures that the test can be shortened without any substantial loss in reliability and validity. Psychologists and others are always on the lookout for brief instruments, and so short forms often become popular, although as a general rule, the shorter the test the less reliable and valid it is. (For some examples of short forms see Burger, 1975; Fischer & Fick, 1993; Kaufman, 1972; Silverstein, 1967.)

Still another type of revision that occurs fairly frequently comes about by *factor analysis.* Let's say I develop a questionnaire on depression that assesses what I consider are four aspects of depression. A factor analysis might indeed indicate that there are four basic dimensions to my test, and so perhaps each should be scored separately, in effect, yielding four scales. Or perhaps, the results of the factor analysis indicate that there is only one factor and that the four subscales I thought were separate are not. Therefore, only one score should be generated. Or the factor analysis might indicate that of the 31 items on the test, 28 are working appropriately, but 3 should be thrown out since their contribution is minimal. (For some examples of factor analysis applied to tests, see Arthur & Woehr, 1993; Carraher, 1993; Casey, Kingery, Bowden & Corbett, 1993; Cornwell, Manfredo, & Dunlap, 1991; W. L. Johnson & A. M. Johnson, 1993).

Finally, there are a number of tests that are *multivariate*, that is the test is composed of many scales, such as in the MMPI and the CPI. The pool of items that comprises the entire test is considered to be an "open system" and additional scales are developed based upon arising needs. For example, when the MMPI was first developed it contained nine different clinical scales; subsequently hundreds of scales have been developed by different authors. (For some examples, see Barron, 1953; Beaver, 1953; Giedt & Downing, 1961; J. C. Gowan & M. S. Gowan, 1955; Kleinmuntz, 1961; MacAndrew, 1965; Panton, 1958.)

TEST ITEMS

Writing test items. Because the total test is no better than its components, we need to take a closer look at test *items.* In general, items should be clear and unambiguous, so that responses do not reflect a misunderstanding of the item. Items should not be double-barreled. For example, "I enjoy swimming and tennis" is a poor item because you would not know whether the response of "true" really means that the person enjoys both of them, only one of them, or outdoor activities in general. Items should not use words such as "sometimes" or "frequently" because these words might mean different things to different people. An item such as, "Do you have headaches frequently?" is better written as, "Do you have a headache at least once a week?" (For more detailed advice on writing test items see Gronlund, 1993; Kline, 1986; Osterlind, 1989; Thorndike & Hagen, 1977; for a bibliography of citations on test construction, see O'Brien, 1988).

Categories of items. There are two basic categories of items: (1) *constructed*-response items where the subject is presented with a stimulus and produces a response – essay exams and sentence-completion tests are two examples; (2) *selected*-response items where the subject selects the correct or best response from a list of options – the

typical multiple-choice question is a good example.

There is a rather extensive body of literature on which approach is better under what circumstances, with different authors taking different sides of the argument (see Arrasmith, Sheehan, & Applebaum, 1984, for a representative study).

Types of items. There are many types of items (see Jensen, 1980; Wesman, 1971). Some of the more common ones:

1. Multiple-choice items. These are a common type, composed of a *stem* that has the question and the *response options* or choices, usually four or five, which are the possible answers. Multiple-choice items should assess the particular content area, rather than vocabulary or general intelligence. The incorrect options, called *distractors*, should be equally attractive to the test taker, and should differentiate between those who know the correct answer and those who don't. The correct response is called the *keyed* response. Sometimes, multiple-choice items are used in tests that assess psychological functioning such as depression or personality aspects, in which case there are no incorrect answers, but the keyed response is the one that reflects what the test assesses. When properly written, multiple-choice items are excellent. There are available guidelines to write good multiple-choice items. Haladyna and Downing (1989a; 1989b) surveyed some 46 textbooks and came up with 43 rules on how to write multiple-choice items; they found that some rules had been extensively researched but others had not. Properly constructed multiple-choice items can measure not only factual knowledge, but also theoretical understanding and problem-solving skills. At the same time, it is not easy to write good multiple-choice items with no extraneous cues that might point to the correct answer (such as the phrase "all of the above") and with content that assesses complex thinking skills rather than just recognition of rote memory material.

Although most multiple-choice items are written with four or five options, a number of writers have presented evidence that three option items may be better (Ebel, 1969; Haladyna & Downing, 1994; Lord, 1944; Sidick, Barrett, & Doverspike, 1994).

Multiple-choice items have a number of advantages. They can be answered quickly, so a particular test can include more items and therefore a broader coverage. They can also be scored quickly and inexpensively, so that results are obtained rapidly and feedback provided without much delay. There is also available computerized statistical technology that allows the rapid computation of item difficulty and other useful indices.

At the same time, multiple-choice items have been severely criticized. One area of criticism is that multiple-choice items are much easier to create for isolated facts than for conceptual understanding, and thus they promote rote learning rather than problem-solving skills. Currently, there seems to be substantial pressure to focus on constructed-response tasks; however, such an approach has multiple problems and may in fact turn out to be even more problematic (Bennet & Ward, 1993).

2. True-false items. Usually, these consist of a statement that the subject identifies as true or false, correct, or incorrect, and so on. For example:

Los Angeles is the capital of California.

I enjoy social gatherings.

Note that in the first example, a factual statement, there is a correct answer. In the second example there is not, but the keyed response would be determined theoretically or empirically; if the item were part of a scale of introversion-extraversion, a true answer might be scored for extraversion.

From a psychometric point of view, factual true-false statements are not very useful. Guessing is a major factor because there is a 50% probability of answering correctly by guessing, and it may be difficult to write meaningful items that indeed are true or false under all circumstances. Los Angeles is not the capital of California but there was a period when it was. Often the item writer needs to include words like *usually*, *never*, and *always* that can give away the correct answer. Personality- or opinion-type true-false items, on the other hand, are used quite frequently and found in many major instruments.

Most textbooks argue that true-false items, as used in achievement tests, are the least satisfactory item format. Other textbooks argue that

the limitations are more the fault of the item writer than with the item format itself. Frisbie and Becker (1991) reviewed the literature and formulated some 21 rules to writing true-false items.

3. Analogies. These are commonly found in tests of intelligence, although they can be used with almost any subject matter. Analogies can be quite easy or difficult and can use words, numbers, designs, and other formats. An example is:

46 is to 24 as 19 is to

(a) 9, (b) 13, (c) 38, (d) 106

(in this case, the answer is 9, because $4 \times 6 = 24$, $1 \times 9 = 9$).

Analogies may or may not be in a multiple-choice format, although providing the choices is a better strategy psychometrically. Like any good multiple choice item, an analogy item has only one correct answer.

4. Odd-man-out. These items are composed of words, numbers, etc., in which one component does *not* belong. For example:

donkey, camel, llama, ostrich

(Here ostrich does not belong because all the other animals have four legs, whereas ostriches have two.)

These items can also be quite varied in their difficulty level and are not limited to words. The danger here is that the dimension underlying the item (leggedness in the above example) may not be the only dimension, may not be necessarily meaningful, and may not be related to the variable being measured.

5. Sequences. This consists of a series of components, related to each other, with the last missing item to be generated by the subject or to be identified from a multiple-choice set. For example:

6, 13, 17, 24, 28, ___

(a) 32, (b) 35, (c) 39, (d) 46

(Here the answer is 35 because the series of numbers increases alternately by 7 points and 4 points: $6 + 7 = 13$; $13 + 4 = 17$; $17 + 7 = 24$; etc.)

6. Matching items. These typically consists of two lists of items to be matched, usually of unequal length to counteract guessing. For example:

Cities	States
A. Toledo	1. California
B. Sacramento	2. Michigan
C. Phoenix	3. North Carolina
D. Ann Arbor	4. Ohio
E. Helena	5. Montana
	6. Arizona
	7. South Dakota
	8. Idaho

Matching items can be useful in assessing specific *factual* knowledge such as names of authors and their novels, dates and historical events, and so on. One problem with matching items is that mismatching one component can result in mismatching other components; thus the components are not independent.

7. Completion items. These provide a *stem* and require the subject to supply an answer. If potential answers are given, this becomes a multiple-choice item. Examples of completion items are:

Wundt established his laboratory in the year ___.

I am always _____.

Note that the response possibilities in the first example are quite limited; the respondent gives either a correct or an incorrect answer. In the second example, different respondents can supply quite different responses. Sentence completion items are used in some tests of personality and psychological functioning.

8. Fill in the blank. This can be considered a variant of the completion item, with the required response coming in a variety of positions. For example:

_____ established the first psychological laboratory.

Wundt established a laboratory at the University of _____ in the year _____.

9. Forced choice items. Forced choice items consist of two or more options, equated as to attractiveness or other qualities, where the

subject must choose one. This type of item is used in some personality tests. For example:

Which item best characterizes you:

(a) I would rather go fishing by myself.

(b) I would rather go fishing with friends.

Presumably, choice (a) would reflect introversion, while choice (b) would reflect extraversion; whether the item works as intended would need to be determined empirically.

10. Vignettes. A vignette is a brief scenario, like the synopsis of a play or novel. The subject is asked to react in some way to the vignette, perhaps by providing a story completion, choosing from a set of alternatives, or making some type of judgment. Examples of studies that have used vignettes are those of G. Domino and Hannah (1987), who asked American and Chinese children to complete brief stories; of DeLuty (1988–1989), who had students assess the acceptability of suicide; of Wagner and Sternberg (1986), who used vignettes to assess what they called "tacit" knowledge; and of Iwao and Triandis (1993), who assessed Japanese and American stereotypes.

11. Rearrangement or continuity items. This is one type of item that is relatively rare but has potential. These items measure a person's knowledge about the order of a series of items. For example, we might list a set of names, such as Wilhelm Wundt, Lewis Terman, Arthur Jensen, etc., and ask the test taker to rank these in chronological order. The difficulty with this type of item is the scoring, but Cureton (1960) has provided a table that can be used in a relatively easy scoring procedure that reflects the difference between the person's answers and the scoring key.

Objective-subjective continuum. Different kinds of test items can be thought of as occupying a continuum along a dimension of objective-subjective:

objective ——————————— subjective

From a psychometric point of view objective items, such as multiple-choice items are the best. They are easily scored, contain only one correct answer, and can be handled statistically with relative ease. The shortcoming of such items is that they only yield the information of whether the subject answered correctly or incorrectly, or whether the subject chose "true" rather than "false" or "option A" rather than "option B." They do not tell us whether the choice reflects lucky guessing, test "wiseness," or actual knowledge.

Subjective items, such as essay questions, on the other hand, allow the respondent to respond in what can be a unique and revealing way. Guessing is somewhat more difficult, and the information produced is often more personal and revealing. From a clinical point of view, open-ended items such as, "Tell me more about it?" "What brings you here?" or "How can I be of help?" are much more meaningful in assessing a client. Psychometrically, such responses are difficult to quantify and treat statistically.

Which item format to use? The choice of a particular item format is usually determined by the test constructor's preferences and biases, as well as by the test content. For example, in the area of personality assessment, many inventories have used a "true-false" format rather than a multiple-choice format. There is relatively little data that can serve as guidance to the prospective test author – only some general principles and some unresolved controversies.

One general principle is that statistical analyses require variation in the raw scores. The item, "are you alive at this moment" is not a good item because, presumably, most people would answer yes. We can build in variation by using item formats with several choices, such as multiple-choice items or items that require answering "strongly agree, agree, undecided, disagree, or strongly disagree," rather than simply true-false; we can also increase variation by using more items – a 10-item test can yield scores that range from 0 to 10, while a 20-item test can yield scores that range from 0 to 20. If the items use the "strongly agree . . . strongly disagree" response format, we can score each item from 1 to 5, and the 10-item test now can yield raw scores from 10 to 50.

One unresolved controversy is whether item response formats such as "strongly agree . . . strongly disagree" should have an "undecided" option or should force respondents to choose sides; also should the responses be an odd

number so a person can select the middle "neutral" option, or should the responses be an even number, so the subject is forced to choose?

An example of the data available comes from a study by Bendig (1959) who administered a personality inventory to two samples, one receiving the standard form with a trichotomous response (true, ?, false), the other a form that omitted the ? response. The results were pretty equivalent, and Bendig (1959) concluded that using a dichotomous response was more economical in terms of scoring cost (now, it probably does not make any difference). For another example, see Tzeng, Ware, and Bharadwaj (1991).

Sequencing of items. Items in a test are usually listed according to some plan or rationale rather than just randomly. In tests of achievement or intelligence, a common strategy is to have easy items at the beginning and progressively difficult items toward the end. Another plan is to use a *spiral omnibus* format, which involves a series of items from easy to difficult, followed by another series of items from easy to difficult, and so on. In tests of personality where the test is composed of many scales, items from the same scale should not be grouped together, otherwise the intent of each scale becomes obvious and can alter the responses given. Similarly, some scales contain *filler* items that are not scored but are designed to "hide" the real intent of the scale. The general rule to be followed is that we want test performance to reflect whatever it is that the test is measuring, rather than some other aspect such as fatigue, boredom, speed of response, second-guessing, and so on; so where possible, items need to be placed in a sequence that will offset any such potential *confounding* variables.

Direct assessment. Over the years, great dissatisfaction has been expressed about these various types of items, especially multiple-choice items. Beginning about 1990, a number of investigators have begun to call for "authentic" measurement (Wiggins, 1990). Thus, more emphasis is being given to what might be called direct or performance assessment, that is, assessment providing for direct measurement of the product or performance generated. Thus, if we wanted to test the competence of a football player we would not administer a multiple-choice exam, but would observe that person's ability to throw a ball, run 50 yards, pass, and so on. If we wanted to assess Johnny's arithmetic knowledge we would give him arithmetic problems to solve. Note that in the latter case, we could easily test Johnny's performance by traditional test items, although a purist might argue that we need to take Johnny to the grocery store and see if he can compute how much six oranges and three apples cost, and how much change he will receive from a $5 bill. This is of course, not a new idea. Automobile driving tests, Red Cross swimming certification, and cardiopulmonary resuscitation are all examples of such performance testing. Advocates of direct assessment argue that such assessment should more closely resemble the actual learning tasks and should allow the candidate to show higher-order cognitive skills such as logical reasoning, innovative problem solving, and critical thinking. Thus, the multiple-choice format is being de-emphasized and more focus is being placed on portfolios, writing samples, observations, oral reports, projects, and other "authentic" procedures [see the special issue of *Applied Psychological Measurement*, 2000 (Vol. 24, No. 4)].

The concepts of reliability and validity apply equally well to standard assessment as to authentic measurement, and the difficulties associated with authentic testing are rather challenging (Hambleton & Murphy, 1992; M. D. Miller & Linn, 2000). In addition to individual scholars, researchers affiliated with Educational Testing Service and other companies are researching these issues, although it is too early to tell whether their efforts will have a major future impact.

PHILOSOPHICAL ISSUES

In addition to practical questions, such as what type of item format to use, there are a number of philosophical issues that guide test construction. One such question is, "How do we know when an item is working the way it is supposed to?" Three basic answers can be given: by fiat, by criterion keying, and by internal consistency.

By fiat. Suppose you put together a set of items to measure depression. How would you know that they measure depression? One way, is to simply state that they do, that because you are an expert on depression, that because the items

reflect our best thinking about depression, and that because the content of all the items is clearly related to depression, therefore your set of items must be measuring depression. Most psychologists would not accept this as a final answer, but this method of *fiat* (a decree on the basis of authority), can be acceptable as a first step. The Beck Depression Inventory, which is probably one of the most commonly used measures of depression, was initially developed this way (A. T. Beck, 1967), although subsequent research has supported its utility. The same can be said of the Stanford-Binet test of intelligence.

Criterion-keyed tests. Many of the best known tests such as the MMPI, CPI, and Strong Vocational Interest Blank, were constructed using this method. Basically, a pool of items is administered to a sample of subjects, for whom we also obtain some information on a relevant *criterion*, for example, scores on another test, GPA, ratings by supervisors, etc. For each test item we perform a statistical analysis (often using correlation) that shows whether the item is empirically related to the criterion. If it does, the item is retained for our final test. This procedure may be done several times with different samples, perhaps using different operational definitions for the criterion. The decision to retain or reject a test item is based solely on its statistical power, on its relationship to the criterion we have selected.

The major problem with this approach is the choice of criterion. Let's assume I have developed a pool of items that presumably assess intelligence. I will administer this pool of items to a sample of subjects and also obtain some data for these subjects on some criterion of intelligence. What criterion will I use? Grade point average? Yearly income? Self-rated intelligence? Teacher ratings? Number of inventions? Listing in a "Who's Who?" Each of these has some serious limitations, and I am sure you appreciate the fact that in the real world criteria are complex and far from perfect. Each of these criteria might also relate to a different set of items, so the items that are retained reflect the criterion chosen.

Some psychologists have difficulties with the criterion-keyed methodology in that the retained set of items may work quite well, but the theoretical reason may not be obvious. A scale may identify those who have leadership capabilities to different degrees, but it may not necessarily measure leadership in a theoretical sense because the items were chosen for their statistical relationship rather than their theoretical cogency.

Criterion-keyed scales are typically *heterogeneous* or *multivariate.* That is, a single scale designed to measure a single variable is typically composed of items that, theoretically and/or in content, can be quite different from each other, and thus, it can be argued, represent different variables. In fact, a content analysis or a factor analysis of the scale items might indicate that the items fall in separate clusters. This is because the criterion used is typically complex; GPA does not just reflect academic achievement, but also interest, motivation, grading policies of different teachers, and so on. Retained items may then be retained because they reflect one or more of these aspects.

A related criticism sometimes made about such scales is that the results are a function of the particular criterion used. If in a different situation a different criterion is used, then presumably the scale may not work. For example, if in selecting items for a depression scale the criterion is "psychiatric diagnosis," then the scale may not work in a college setting where we may be more concerned about dropping out or suicide ideation. This of course, is a matter of empirical validity and cannot be answered by speculation. In fact, scales from tests such as the CPI have worked remarkably well in a wide variety of situations.

A good example of empirical scale construction is the study by Rehfisch (1958), who set about to develop a scale for "personal rigidity." He first reviewed the literature to define the rigidity-flexibility dimension and concluded that the dimension was composed of six aspects: (1) constriction and inhibition, (2) conservatism, (3) intolerance of disorder and ambiguity, (4) obsessional and perseverative tendencies, (5) social introversion, and (6) anxiety and guilt. At this point, he could have chosen to write a pool of items to reflect these six dimensions and publish his scale on the basis of its theoretical underpinnings and his status as an "expert" – this would have been the fiat method we discussed above. Or he could have chosen to administer the pool of items to a large group of subjects and through factor analysis determine whether

the results indicated one main factor, presumably rigidity, or six factors, presumably the above dimensions. We discuss this method next.

Instead he chose to use data that was already collected by researchers at the Institute of Personality Assessment and Research of the University of California at Berkeley. At this institute, a number of different samples, ranging from graduate students to Air Force captains, had been administered – batteries of tests, including the CPI and the MMPI, had been rated by IPAR staff on a number of dimensions, including "rigidity." Rehfisch simply analyzed statistically the responses to the combined CPI-MMPI item pool (some 957 true-false statements) of the subjects rated highest and lowest 25% on rigidity. He *cross-validated*, that is replicated the analysis, on additional samples. The result was a 39-item scale that correlated significantly with a variety of ratings, and which was substantially congruent with the theoretical framework. High scorers on this scale tend to be seen as anxious, overcontrolled, inflexible in their social roles, orderly, and uncomfortable with uncertainty. Low scorers tend to be seen as fluent in their thinking and in their speech, outgoing in social situations, impulsive, and original. Interestingly enough, scores on the scale correlated only .19 with ratings of rigidity in a sample of medical school applicants. It is clear that the resulting scale is a "complex" rather than a "pure" measure of rigidity. In fact, a content analysis of the 39 items suggested that they can be sorted into eight categories ranging from "anxiety and constriction in social situations" to "conservatism and conventionality." A subsequent study by Rehfisch (1959) presented some additional evidence for the validity of this scale.

Factor-analysis as a way of test construction. This approach assumes that scales should be *univariate* and *independent*. That is, scales should measure only one variable and should not correlate with scales that measure a different variable. Thus, all the items retained for a scale should be *homogeneous*, they should all be interrelated.

As in the criterion-keying method, we begin with a pool of items that are administered to a sample of subjects. The sample may be one of convenience (e.g., college sophomores) or one of theoretical interest (patients with the diagnosis of anxiety) related to our pool of items. The responses are translated numerically (e.g., true = 1, false = 2), and the numbers are subjected to factor analysis. There are a number of techniques and a number of complex issues involved in factor analysis, but for our purposes we can think of factor analysis as a correlational analysis with items being correlated with a mythical dimension called a *factor*. Each item then has a *factor loading*, which is like a correlation coefficient between responses on that item and the theoretical dimension of the factor. Items that load significantly on a particular factor are assumed to measure the same variable and are retained for the final scale. Factor analysis does not tell us what the psychological meaning of the factor is, and it is up to the test constructor to study the individual items that load on the factor, and name the factor accordingly. A pool of items may yield several factors that appear to be statistically "robust" and psychologically meaningful, or our interest may lie only in the first, main factor and in the one scale.

As with criterion-keying, there have been a number of criticisms made of the factor-analytic approach to test construction. One is that factor analysis consists of a variety of procedures, each with a variety of assumptions and arbitrary decisions; there is argument in the literature about which of the assumptions and decisions are reasonable and which are not (e.g., Gorsuch, 1983; Guilford, 1967b; Harman, 1960; Heim, 1975).

Another criticism is that the results of a factor analysis reflect only what was included in the pool of items. To the extent that the pool of items is restricted in content, then the results of the factor analysis will be restricted. Perhaps I should indicate here that this criticism is true of any pool of items, regardless of what is done to the items, but that usually those of the criterion-keying persuasion begin with pool items that are much more heterogeneous. In fact, they will often include items that on the surface have no relationship to the criterion, but the constructor has a "hunch" that the item might work.

Still another criticism is that the factor analytic dimensions are theoretical dimensions, useful for understanding psychological phenomena, but less useful as predictive devices. Real-life behavior is typically complex; grades in college reflect not just mastery of specific topic areas, but

general intelligence, motivation, aspiration level, the pressures of an outside job, personal relationships such as being "in love," parental support, sleep habits, and so on. A factor analytic scale of intelligence will only measure "pure intelligence" (whatever that may be) and thus not correlate highly with GPA, which is a complex and heterogeneous variable. (To see how a factor analytic proponent answers these criticisms, see P. Kline, 1986.)

ADMINISTERING A TEST

If we consider a test as either an interview or an experiment, then how the test is administered becomes very important. If there is a manual available for the particular test, then the manual may (or may not) have explicit directions on how to administer the test, what specific instructions to read, how to answer subjects' questions, what time limits if any to keep, and so on.

Rapport. One of the major aspects of test administration involves *rapport*, the "bond" that is created between examiner and examinee, so that the subject is cooperative, interested in the task, and motivated to pay attention and put forth a best effort. Sometimes such motivation is strongly affected by outside factors. A premedical student eager to be accepted into medical school will typically be quite cooperative and engaged in the task of taking a medical college admissions test; a juvenile delinquent being assessed at the request of a judge, may not be so motivated.

In the American culture, tests and questionnaires are fairly common, and a typical high school or college student will find little difficulty in following test directions and doing what is being asked in the time limit allotted. Individuals such as young children, prisoners, emotionally disturbed persons, or individuals whose educational background has not given them substantial exposure to testing, may react quite differently.

Rapport then is very much like establishing a special bond with another person, such as occurs in friendships, in marriage, and in other human relationships. There are no easy steps to do so, and no pat answers. Certainly, if the examiner appears to be a warm and caring person, sensitive to the needs of the subject, rapport might be easier to establish. On the other hand, we expect a professional to be friendly but businesslike, so if the warmth becomes "gushiness," rapport might decrease. Rapport is typically enhanced if the subject understands why she or he is being tested, what the tests will consist of, and how the resulting scores will be used. Thus, part of establishing rapport might involve allaying any fears or suspicions the subject may have. Rapport is also enhanced if the subject perceives that the test is an important tool to be used by a competent professional for the welfare of the client.

INTERPRETING TEST SCORES

A test usually yields a raw score, perhaps the number of items answered correctly. Raw scores in themselves are usually meaningless, and they need to be changed in some way to give them meaning. One way is to compare the raw score to a group average – that is what the word "norm" means, normal or average. Thus, you obtained a raw score of 72 on a vocabulary test, and upon finding that the average raw score of a sample of college students is 48, you might be quite pleased with your performance. Knowing the average is, of course, quite limited information. When we have a raw score we need to locate that raw score in more precise terms than simply above or below average. Normative data then typically consist not just of one score or average, but the actual scores of a representative and sizable sample that allow you to take any raw score and translate it into a precise location in that normative group. To do this, raw scores need to be changed into *derived* scores.

Percentiles. Let's suppose that our normative group contained 100 individuals and, by sheer luck, each person obtained a different score on our vocabulary test. These scores could be ranked, giving a 1 to the lowest score and a 100 to the highest score. If John now comes along and takes the vocabulary test, his raw score can be changed into the equivalent rank – his score of 76 might be equivalent to the 85th rank. In effect, that is what percentile scores are. When we have a distribution of raw scores, even if they are not all different, and regardless of how many scores we have, we can change raw scores into percentiles. Percentiles are a rank, but they represent the upper limit of the rank. For example,

a score at the 86th percentile is a score that is higher than 86 out of 100, and conversely lower than 14 out of 100; a score at the 57th percentile is a score that is higher than 57 out of 100, and lower than 43 out of 100. Note that the highest possible percentile is 99 (no score can be above all 100), and the lowest possible percentile is 1 (no one can obtain a score that has no rank).

Percentiles are intrinsically meaningful in that it doesn't matter what the original scale of measurement was, the percentile conveys a concrete position (see any introductory statistical text for the procedure to calculate percentiles). Percentiles have one serious limitation; they are an ordinal scale rather than an interval scale. Although ranks would seem to differ by only one "point," in fact different ranks may differ by different points depending on the underlying raw score distribution. In addition, if you have a small sample, not all percentile ranks will be represented, so a raw score of 72 might equal the 80th percentile, and a raw score of 73, the 87th percentile.

Standard scores. We said that just knowing the average is not sufficient information to precisely locate a raw score. An average will allow us to determine whether the raw score is above or below the average, but we need to be more precise. If the average is 50 and the raw score is 60, we could obviously say that the raw score is "10 points above the mean." That would be a useful procedure, except that each test has its own measurement scale – on one test the highest score might be 6 points above the mean, while on another test it might be 27 points above the mean, and how far away a score is from the mean is in part a function of how variable the scores are. For example, height measured in inches is typically less variable than body weight measured in ounces. To equalize for these sources of variation we need to use a scale of measurement that transcends the numbers used, and that is precisely what the *standard deviation* gives us. If we equate a standard deviation to one, regardless of the scale of measurement, we can express a raw score as being x number of standard deviations above or below the mean. To do so we change our raw scores into what are called *standard* or z scores, which represent a scale of measurement with mean equal to zero and SD equal to 1.

Consider a test where the mean is 62 and the SD is 10. John obtained a raw score of 60, Barbara, a raw score of 72, and Consuelo, a raw score of 78. We can change these raw scores into z scores through the following formula:

$$z = \frac{\mathrm{X} - \mathrm{M}}{\mathrm{SD}}$$

where X is the raw score

M is the mean and
SD is the standard deviation

For John, his raw score of 60 equals:

$$z = \frac{60 - 62}{10} = -0.2$$

For Barbara, her raw score of 72 equals:

$$z = \frac{72 - 62}{10} = +1.0$$

and for Consuelo, her raw score of 78 equals:

$$z = \frac{78 - 62}{10} = +1.60$$

We can plot these 3 z scores on a normal curve graph and obtain a nice visual representation of their relative positions (see Figure 2.1).

Note that changing raw scores into z scores does not alter the relative position of the three individuals. John is still the lowest scoring person, Consuelo the highest, and Barbara is in the middle. Why then change raw scores into z scores? Aside from the fact that z scores represent a scale of measurement that has immediate meaning (a z score of $+3$ is a very high score no matter what the test, whereas a raw score of 72 may or may not be a high score), z scores also allow us to compare across tests. For example, on the test above with mean of 62 and SD of 10, Consuelo obtained a raw score of 78. On a second test, with mean of 106 and SD of 9, she obtained a raw score of 117. On which test did she do better? By changing the raw scores to z scores the answer becomes clear. On test A, Consuelo's raw score of 78 equals:

$$z = \frac{78 - 62}{10} = +1.60$$

On test B, Consuelo's raw score of 117 equals:

$$z = \frac{177 - 106}{9} = +1.22$$

Plotting these on a normal curve graph, as in Figure 2.2, we see that Consuelo did better on test A.

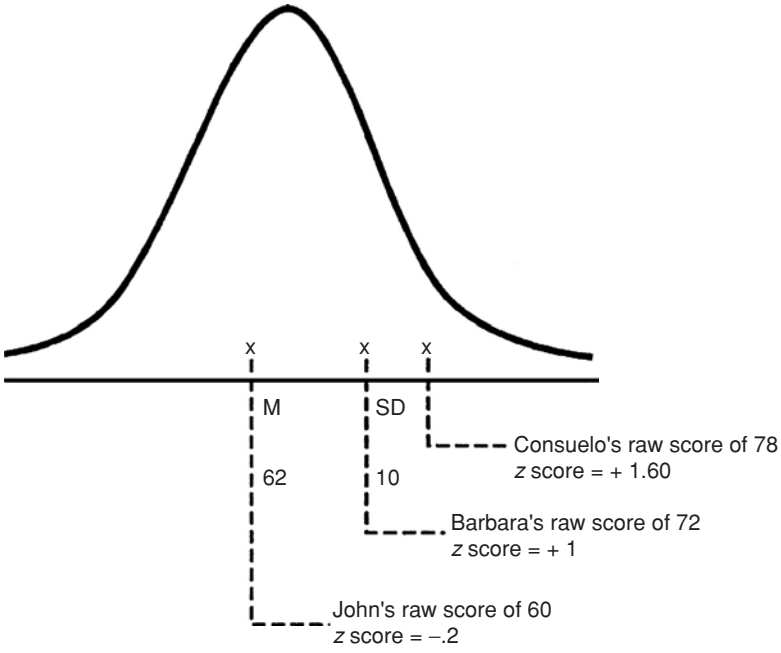


FIGURE 2–1. Relative positions of three *z* scores.

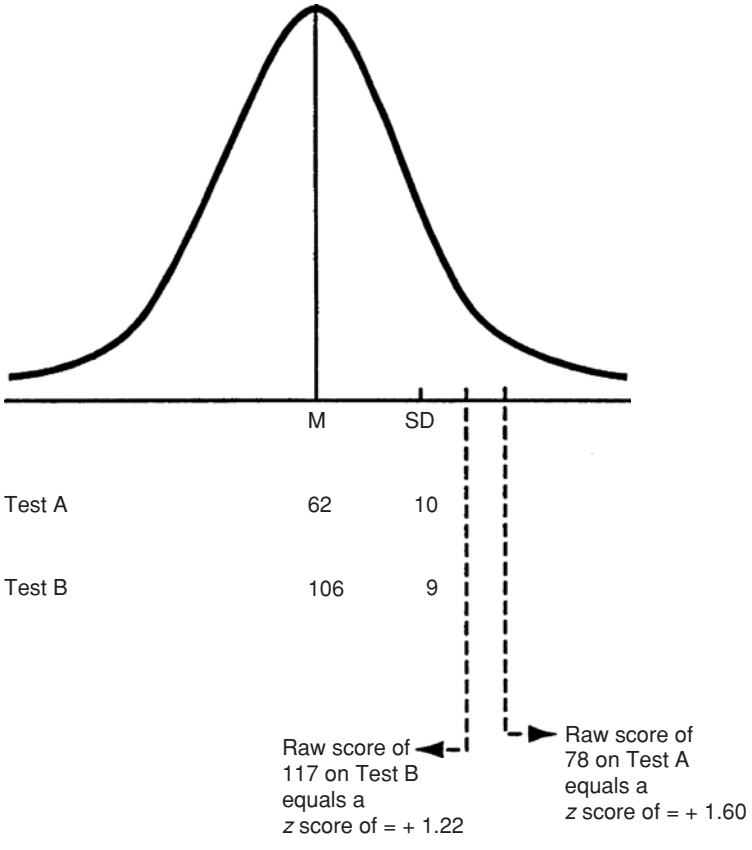


FIGURE 2–2. Equivalency of raw scores to *z* scores.

***T* scores.** The problem with *z* scores is that they can involve both positive and negative numbers as well as decimal numbers, and so are somewhat difficult to work with. This is a problem that can be easily resolved by changing the mean and SD of *z* scores to numbers that we might prefer. Suppose we wanted a scale of measurement with a mean of 50 and a SD of 10. All we need to do is multiply the *z* score we wish to change by the desired SD and add the desired mean. For example, to change a *z* score of + 1.50 we would use this formula:

$$\begin{aligned}\text{new score} &= z(\text{desired SD}) + \text{desired mean}\\ &= +1.50(10) + 50\\ &= 65\end{aligned}$$

This new scale, with a mean of 50 and SD of 10 is used so often in testing, especially for personality tests, that it is given a name: *T scores*; when you see *T* scores reported, you automatically know that the mean is 50 and the SD is 10, and that therefore a score of 70 is two standard deviations above the mean.

Educational Testing Service (ETS) uses a scale of measurement with mean of 500 and SD of 100 for its professional tests such as the SAT and the GRE. These are really *T* scores with an added zero. Note that an individual would not obtain a score of 586 – only 580, 590, and so on.

Stanines. Another type of transformation of raw scores is to a scale called *stanine* (a contraction of standard nine) that has been used widely in both the armed forces and educational testing. Stanines involve changing raw scores into a normally shaped distribution using nine scores that range from 1 (low) to 9 (high), with a mean of 5 and SD of 2. The scores are assigned on the basis of the following percentages:

stanine:	1	2	3	4	5	6	7	8	9
percentage:	4	7	12	17	20	17	12	7	4

Thus, in a distribution of raw scores, we would take the lowest 4% of the scores and call all of them ones, then the next 7% we would call two's, and so on (all identical raw scores would however be assigned the same stanine).

Stanines can also be classified into a fivefold classification as follows:

stanine:	1	2&3	4,5,6	7&8	9
defined as:	poor	below average	average	above average	superior
percentage:	4	19	54	19	4

or a tripartite classification:

stanine:	1,2,3	4,5,6	7,8,9
defined as:	low	average	high
percentage:	23	54	23

Sometimes stanines actually have 11 steps, where the stanine of 1 is divided into 0 and 1 (with 1% and 3% of the cases), and the stanine of 9 is divided into 9 and 10 (with 3% and 1% of the cases). Other variations of stanines have been prepared, but none have become popular (Canfield, 1951; Guilford & Fruchter, 1978). Note that unlike *z* scores and *T* scores, stanines force the raw score distribution into a normal distribution, whereas changing raw scores into *z* scores or *T* scores using the above procedures does not change the shape of the distribution. Don't lose sight of the fact that all of these different scales of measurement are really equivalent to each other. Figure 2.3 gives a graphic representation of these scales.

ITEM CHARACTERISTICS

We now need to take a closer look at two aspects of test items: *item difficulty* and *item discrimination.*

Item Difficulty

The difficulty of an item is simply the percentage of persons who answer the item correctly. Note that the higher the percentage the easier the item; an item that is answered correctly by 60% of the respondents has a *p* (for percentage) value of .60. A difficult item that is answered correctly by only 10% has a $p = .10$ and an easy item answered correctly by 90% has a $p = .90$. Not all test items have correct answers. For example, tests of attitudes, of personality, of political opinions, etc., may present the subject with items that require agreement-disagreement, but for which there is no correct answer. Most items however, have a keyed response, a response that if endorsed is given points. On a scale of anxiety, a "yes" response to the item, "are you nervous most of the time?" might be counted as reflecting anxiety and would be the keyed response.

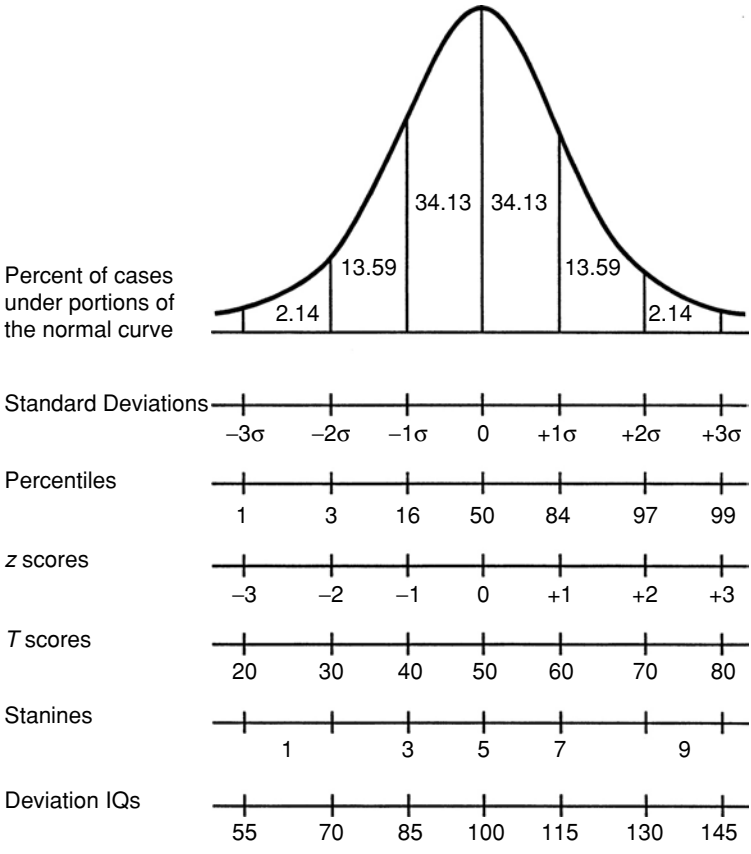


FIGURE 2–3. Relationships of different types of scores, based on the normal distribution.

If the test were measuring "calmness," then a "no" response to that item might be the keyed response. Thus item difficulty can simply represent the percentage who endorsed the keyed response.

What level difficulty? One reason we may wish to know the difficulty level of items is so we can create tests of different difficulty levels, by judicious selection of items. In general, from a psychometric point of view, tests should be of average difficulty, average being defined as $p = .50$. Note that this results in a mean score near 50%, which may seem quite a demanding standard. The reason for this is that a $p = .50$ yields the most discriminating items, items that reflect individual differences. Consider items that are either very difficult ($p = .00$) or very easy ($p = 1.00$). Psychometrically, such items are not useful because they do not reflect any differences between individuals. To the degree that different individuals give different answers, and the answers are related to some behavior, to that degree are the items useful, and thus generally the most useful items are those with p near .50.

The issue is, however, somewhat more complicated. Assume we have a test of arithmetic, with all items of $p = .50$. Children taking the test would presumably not answer randomly, so if Johnny gets item 1 correct, he is likely to get item 2 correct, and so on. If Mark misses item 1, he is likely to miss item 2, and so on. This means, at least theoretically, that one half of the children would get all the items correct and one half would get all of them incorrect, so that there would be only two raw scores, either zero or 100 – a very unsatisfactory state of affairs. One way to get around this is to choose items whose *average* value of difficulty is .50, but may in fact range widely, perhaps from .30 to .70, or similar values.

Another complicating factor concerns the target "audience" for which the test will be used.

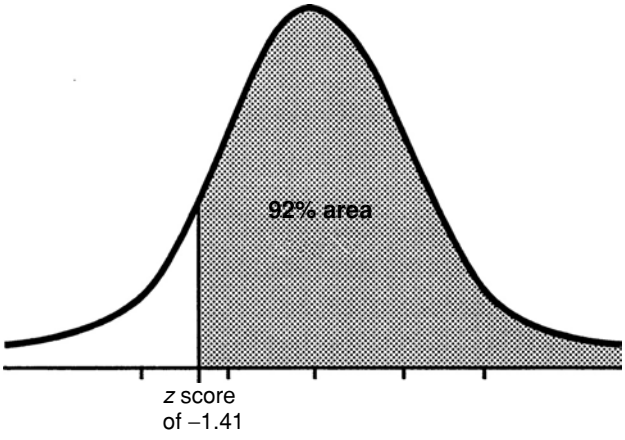


FIGURE 2–4. Example of an easy test item passed by 92% of the sample.

Let's say I develop a test to identify the brightest 10% of entering college freshmen for possible placement in an honors program. In that case, the test items should have an average $p = .10$, that is, the test should be quite difficult with the average p value reflecting the percentage of scores to be selected – in this example, 10%. Tests such as the SAT or GRE are quite demanding because their difficulty level is quite high.

Measurement of item difficulty. Item difficulty then represents a scale of measurement identical with percentage, where the average is 50% and the range goes from zero to 100%. This is of course an ordinal scale and is of limited value because statistically not much can be done with ordinal measurement. There is a way however, to change this scale to an interval scale, by changing the percent to z scores. All we need to do is have a table of normal curve frequencies (see appendix) and we can read the z scores directly from the corresponding percentage. Consider for example, a very easy item with $p = .92$, represented by Figure 2.4. Note that by convention, higher scores are placed on the right, and we assume that the 92% who got this item correct were higher scoring individuals (at least on this item). We need then to translate the percentage of the area of the curve that lies to the right (92%) into the appropriate z score, which our table tells us is equal to -1.41.

A very difficult item of $p = .21$ would yield a z score of $+0.81$ as indicated in Figure 2.5. Note that items that are easy have negative z scores, and items that are difficult have positive z scores. Again, we can change z scores to a more manageable scale of measurement that eliminates negative values and decimals. For example, ETS uses a *delta* scale with a mean of 13 and a SD = 4. Thus delta scores $= z\,(4) + 13$. An item with $p = .58$ would yield a z score of $-.20$ which would equal a delta score of:

$$(-.20)(4) + 13 = 12.2 \text{ (rounding off} = 12)$$

The bandwidth-fidelity dilemma. In developing a test, the test constructor chooses a set of items from a larger pool, with the choice based on rational and/or statistical reasons. Classical test theory suggests that the best test items are those with a .50 difficulty level – for example, a multiple choice item where half select the correct answer, and half the distractors. If we select all or most of the items at that one level of difficulty, we will have a very good instrument for measuring those individuals who indeed fall at that level on the trait being measured. However, for individuals who are apart from the difficulty level, the test will not be very good. For example, a person who is low on the trait will receive a low score based on the few correctly answered items; a person who is high will score high, but the test will be "easy" and again won't provide much information. In this approach, using a "peaked" conventional test (peaked because the items peak at a particular difficulty level), we will be able to measure some of the people very well and some very poorly.

We can try to get around this by using a rectangular distribution of items, that is, selecting a few items at a .10 level of difficulty, a few at .20, a few at .30 and so on to cover the whole range of difficulty, even though the average range of difficulty will still be .50. There will be items here that are appropriate for any individual no matter

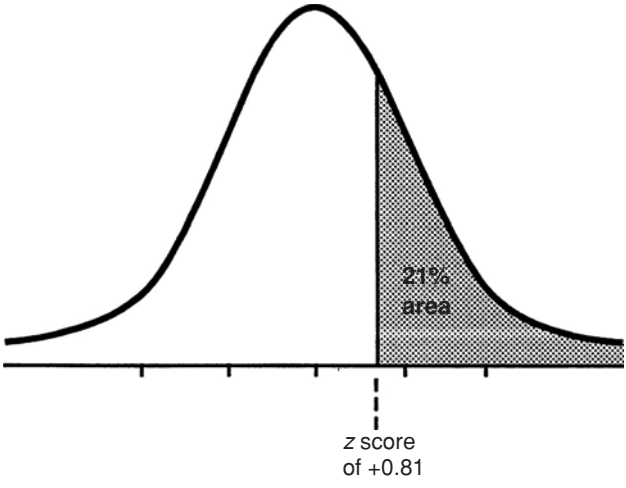


FIGURE 2–5. Example of a difficult test item passed by 21% of the sample.

where they are on the trait, but because a test cannot be too long, the appropriate items for any one person will be few. This means that the test will be able to differentiate between individuals at various levels of a trait, but the precision of these differentiations will not be very great.

A peaked conventional test can provide high fidelity (i.e., precision) where it is peaked, but little bandwidth (i.e., it does not differentiate very well individuals at other positions on the scale). Conversely, a rectangular conventional test has good bandwidth but low overall fidelity (Weiss, 1985).

Guessing. Still another complicating factor in item difficulty is that of guessing. Although individuals taking a test do not usually answer randomly, just as typically there is a fair amount of guessing going on, especially with multiple-choice items where there is a correct answer. This inflates the *p* value because a *p* value of .60 really means that among the 60% who answered the item correctly, a certain percentage answered it correctly by lucky guessing, although some will have answered it incorrectly by bad guessing (see Lord, 1952).

A number of item forms, such as multiple-choice items, can be affected by guessing. On a multiple-choice examination, with each item composed of five choices, anyone guessing blindly would, by chance alone, answer about one fifth of the items correctly. If all subjects guessed to the same degree, guessing would not be much of a problem. But subjects don't do that, so guessing can be problematic. A number of formulas or corrections of the total score have been developed to take guessing into account, such as:

$$\text{score} = \text{right} - \frac{\text{wrong}}{k-1}$$

where $k =$ the number of alternatives per item.

The rationale here is that the probability of a correct guess is $1/k$ and the probability of an incorrect guess is $k - 1/k$. So we expect, on the average, for a person to be correct once for every $k - 1$ times that they are incorrect. The problem is that correction formulas such as the above assume that item choices are equally plausible, and that items are of two types – those that the subject knows and answers correctly and those that the subject doesn't know, and guesses blindly.

Note that the more choices there are for each item, the less significant guessing becomes. In true-false items, guessing can result in 50% correct responses. In five-choice multiple-choice items, guessing can result in 20% correct answers, but if each item had 20 choices (an awkward state of affairs), guessing would only result in 5% correct responses.

A simpler, but not perfect, solution, is to include instructions on a test telling all candidates to do the same thing – that is, guess when unsure, leave doubtful items blank, etc. (Diamond & Evans, 1973).

Item Discrimination

If we have a test of arithmetic, each item on that test should ideally differentiate between those

who know the subject matter and those who don't know. If we have a test of depression, each item should ideally differentiate between those who are depressed and those who are not. *Item discrimination* refers to the ability of an item to correctly "discriminate" between those who are higher on the variable in question and those who are lower. Note that for most variables we don't ordinarily assume a dichotomy but rather a continuous variable – that is, we don't believe that the world is populated by two types of people, depressed and nondepressed, but rather that different people can show different degrees of depression.

There are a number of ways of computing item-discrimination indices, but most are quite similar (Oosterhof, 1976) and basically involve comparing the performance of high scorers with that of low scorers, for each item. Suppose for example, we have an arithmetic test that we have administered to 100 children. For each child, we have a total raw score on the test, and a record of their performance on each item. To compute item discrimination indices for each item, we first need to decide how we will define "high scorer" vs. "low scorer."

Obviously, we could take all 100 children, compute the median of their total test scores, and label those who scored above the median as high scorers, and those below the median as low scorers. The advantage of this procedure is that we use all the data we have, all 100 protocols. The disadvantage is that at the center of the distribution there is a fair amount of "noise." Consider Sarah, who scored slightly above the median and is thus identified as a high scorer. If she were to retake the test, she might well score below the median and now be identified as a low scorer.

At the other extreme, we could take the five children who really scored high and label them high scorers and the five children who scored lowest and label them low scorers. The advantage here is that these extreme scores are not likely to change substantially on a retest; they most likely are not the result of guessing and probably represent "real-life" correspondence. The disadvantage is that now we have rather small samples, and we can't be sure that any calculations we perform are really stable. Is there a happy medium that on the one hand keeps the "noise" to a minimum and on the other maximizes the size of the sample? Years ago, Kelley (1939) showed that the best strategy is to select the upper 27% and the lower 27%, although slight deviations from this, such as 25% or 30%, don't matter much. (Note that in the example of the rigidity scale developed by Rehfisch, he analyzed the top and bottom 25% of those rated on rigidity.)

Table 2–1

Test item	Upper 27	Lower 27	Index of discrimination
1	23 (85%)	6 (22%)	63%
2	24 (89%)	22 (81%)	8%
3	6 (22%)	4 (15%)	7%
4	9 (33%)	19 (70%)	−37%

For our sample of 100 children we would then select the top 27 scorers and call them "high scorers" and the bottom 27 and call these "low scorers." We would look at their answers for each test item and compute the difficulty level of each item, separately for each group, using percentages. The difference between difficulty levels for a particular item is the *index of discrimination* (abbreviated as D) for that item. Table 2.1 gives an example of such calculations.

Note that the index of discrimination is expressed as a percentage and is computed from two percentages. We could do the same calculations on the raw scores, in this cases the number of correct responses out of 27, but the results might differ from test to test, if the size of the sample changes.

The information obtained from such an analysis can be used to make changes in the items and improve the test. Note, for example, that item 1 seems to discriminate quite well. Most of the high scorers (85%) answered the item correctly, while far fewer of the low scorers (22%) answered the item correctly. Theoretically, a perfectly discriminating item would have a D value of 100%. Items 2 and 3 don't discriminate very well, item 2 is too easy and item 3 is too difficult. Item 4 works but in reverse! Fewer of the higher scorers got the item correctly. If this is an item where there is a correct answer, a negative D would alert us that there is something wrong with the item, that it needs to be rewritten. If this were an item from a personality test where there is no correct answer, the negative D would in fact tell us that we need to reverse the scoring.

We have chosen to define high scorer and low scorer on the basis of the total test score itself.

This may seem a bit circular, but it is in fact quite legitimate. If the test measures arithmetic knowledge, then a high scorer on arithmetic knowledge is indeed someone who scores high on the test. There is a second way, however, to define high and low scorers, or more technically to identify *extreme groups*, and that is to use a criterion that is not part of the test we are calibrating. For example, we could use teacher evaluations of the 100 children as to which ones are good in math and which ones are not. For a test of depression, we could use psychiatric diagnosis. For a personality scale of leadership, we could use peer evaluation, self-ratings, or data obtained from observations.

Does it matter whether we compute item discrimination indices based on total test scores or based on an external criterion? If we realize that such computations are not simply an exercise to fill time, but are done so we can retain those items with the highest D values, those items that work best, then which procedure we use becomes very important because different procedures result in the retention of different items. If we use the total test score as our criterion, an approach called *internal consistency*, then we will be retaining items that tend to be homogeneous, that is items that tend to correlate highly with each other. If we use an external criterion, that criterion will most likely be more complex psychologically than the total test score. For example, teachers' evaluations of being "good at math" may reflect not only math knowledge, but how likeable the child is, how physically attractive, outgoing, all-around intelligent, and so on. If we now retain those items that discriminate against such a complex criterion, we will most likely retain heterogeneous items, items that cover a wide variety of the components of our criterion. If we are committed to measuring arithmetic knowledge in as pure a fashion as possible, then we will use the total test score as our criterion. If we are interested in developing a test that will predict to the maximum degree some real-world behavior, such as teachers' recognition of a child's ability, then we will use the external criterion. Both are desirable practices and sometimes they are combined, but we should recognize that the two practices represent different philosophies of testing. Allen and Yen (1979) argue that both practices cannot be used simultaneously, that a test constructor must choose one or the other. Anastasi (1988), on the other hand, argues that both are important.

Philosophies of testing. And so once again we are faced with the notion that we have alternatives, and although the proponents of each alternative argue that theirs is *the* way, the choice comes down to personal preference and to compatible philosophy of testing. With regard to test construction, there seem to be two basic camps. One approach, that of factor analysis, believes that tests should be pure measures of the dimension being assessed. To develop such a pure measure, items are selected that statistically correlate as high as possible with each other and/or with the total test score. The result is a scale that is homogeneous, composed of items all of which presumably assess the same variable. To obtain such homogeneity, factor analysis is often used, so that the test items that are retained all center on the same dimension or factor. Tests developed this way must not correlate with other dimensions. For example, scores on a test of anxiety must not correlate with scores on a test of depression, if the two dimensions are to be measured separately. Tests developed this way are often useful for understanding a particular psychological phenomenon, but scores on the test may in fact not be highly related to behavior in the real world.

A second philosophy, that of *empiricism*, assumes that scales are developed because their primary function is to predict real-life behavior, and items are retained or eliminated depending on how well they correlate with such real-life behavior. The result is a test that is typically composed of heterogeneous items all of which share a correlation with a non test criterion, but which may not be highly correlated with each other. Such scales often correlate significantly with other scales that measure different variables, but the argument here is that, "that's the way the world is." As a group, people who are intellectually bright also tend to be competent, sociable, etc., so scales of competence may most likely correlate with measures of sociability, and so on. Such scales are often good predictors of real-life behaviors, but may sometimes leave us wondering why the items work as they do. For an interesting example of how these two philosophies can lead their proponents to entirely different views, see the reviews of the CPI in the seventh MMY (Goldberg, 1972; Walsh, 1972), and in the ninth MMY (Baucom, 1985; Eysenck, 1985).

Item response theory (IRT). The "classical" theory of testing goes back to the early 1900s when Charles Spearman developed a theoretical framework based on the simple notion that a test score was the sum of a "true" score plus random "error." Thus a person may obtain different IQs on two intelligence tests because of differing amounts of random error; the true score presumably does not vary. Reliability is in fact a way of assessing how accurately obtained scores covary with true scores.

A rather different approach known as item response theory (IRT) began in the 1950s primarily through the work of Frederic Lord and George Rasch. IRT also has a basic assumption and that is that performance on a test is a function of an unobservable proficiency variable. IRT has become an important topic, especially in educational measurement. Although it is a difficult topic that involves some rather sophisticated statistical techniques beyond the scope of this book (see Hambleton & Swaminathan, 1985; Lord, 1980), the basic idea is understandable.

The characteristics of a test item, such as item difficulty, are a function of the particular sample to whom the item was administered. A vocabulary item may, for example, be quite difficult for second graders but quite easy for college students. Thus in classical test theory, item difficulty, item discrimination, normative scores, and other aspects are all a function of the particular samples used in developing the test and generating norms; typically, a raw score is interpreted in terms of relative position within a sample, such as percentile rank or other transformation. IRT, on the other hand, focuses on a theoretical mathematical model that unites the characteristics of an item, such as item difficulty, to an underlying hypothesized dimension. Although the parameters of the theoretical model are estimated from a specific set of data, the computed item characteristics are not restricted to a specific sample. This means, in effect, that item pools can be created and then subsets of items selected to meet specific criteria – for example, a medium level of difficulty. Or subset of items can be selected for specific examinees (for a readable review of IRT see Loyd, 1988).

Basically, then, IRT is concerned with the interplay of four aspects: (1) the ability of the individual on the variable being assessed, (2) the extent to which a test item discriminates between high- and low-scoring groups, (3) the difficulty of the item, and (4) the probability that a person of low ability on that variable makes the correct response.

NORMS

No matter what philosophical preferences we have, ultimately we are faced with a raw score obtained from a test, and we need to make sense of that score. As we have seen, we can change that raw score in a number of ways, but eventually we must be able to compare that score with those obtained for a normative sample, and so we need to take a closer look at norms.

How are norms selected? Commercial companies that publish tests (for a listing of these consult the MMY) may have the financial and technical means to administer a test to large and representative groups of subjects in a variety of geographical settings. Depending on the purpose of the test, a test manual may present the scores of subjects listed separately for such variables as gender (males vs. females), school grade (e.g., fifth graders, sixth graders, etc.), time of testing (e.g., high-school seniors at the beginning of their senior year vs. high-school seniors near the end of the school year), educational level (high-school graduates, college graduates, etc.), geographical region (Northeast, Southwest, etc.) and other relevant variables or combination of variables.

Sometimes the normative groups are formed on the basis of *random sampling*, and sometimes they are formed on the basis of certain criteria, for example U.S. Census data. Thus if the census data indicate that the population is composed of different economic levels, we might wish to test a normative sample that reflects those specific percentages; this is called a *stratified sample.* More typically, especially with tests that are not commercially published, norms are made up of *samples of convenience.* An investigator developing a scale of leadership ability might get a sample of local business leaders to take the test, perhaps in return for a free lecture on "how to improve one's leadership competence," or might have a friend teaching at a graduate college of business agree to administer the test to entering students. Neither of these samples would be random, and

one might argue neither would be representative. As the test finds continued use, a variety of samples would be tested by different investigators and norms would be accumulated, so that we could learn what average scores are to be expected from particular samples, and how different from each other specific samples might be. Often, despite the nonrandomness, we might find that groups do not differ all that much – that the leadership level exhibited by business people in Lincoln, Nebraska, is not all that different from that exhibited by their counterparts in San Francisco, Atlanta, or New York City.

Age norms. Often we wish to compare a person's test score with the scores obtained by a normative group of the same age. This makes sense if the variable being assessed changes with age. When we are testing children, such age norms become very important because we expect, for example, the arithmetic knowledge of a 5-year-old to be different from that of a 9-year-old. With some variables, there may be changes occurring well within a short time span, so we might need age norms based on a difference of a few months or less. With adults, age norms are typically less important because we would not expect, for example, the average 50-year-old person to know more (or less) arithmetic than the average 40-year-old. On the other hand, if we are testing college students on a measure of "social support" we would want to compare their raw scores with norms based on college students rather than on retired senior citizens.

School grade norms. At least in our culture, most children are found in school and schooling is a major activity of their lives. So tests that assess school achievement in various fields, such as reading, social studies, etc., often have norms based on school grades. If we accept the theoretical model that a school year covers 10 months, and if we accept the fiction that learning occurs evenly during those 10 months, we can develop a test where each item is assigned a score based on these assumptions. For example, if our fifth-grade reading test is composed of 20 items, each item answered correctly could be given one-half month-credit, so a child answering all items correctly would be given one school-year credit, a child answering 16 items correctly would be given eight months' credit, and so on.

Unfortunately, this practice leads to some strange interpretations of test results. Consider Maria, a fourth grader, who took a reading comprehension test. She answered correctly all of the items at the fourth grade and below, so she receives a score of 4 years. In addition however, she also answered correctly several items at the fifth-grade level, several items at the sixth-grade level, a few at the seventh-grade level, and a couple at the eighth-grade level. For all of these items, she receives an additional 2 years credit, so her final score is sixth school year. Most likely when her parents and her teacher see this score they will conclude incorrectly that Maria has the reading comprehension of a sixth grader, and that therefore she should be placed in the sixth grade, or at the very least in an accelerated reading group. In fact, Maria's performance is typical. Despite our best efforts at identifying test items that are appropriate for a specific grade level, children will exhibit *scatter*, and rarely will their performance conform to our theoretical preconceptions. The test can still be very useful in identifying Maria's strengths or weaknesses, and in providing an objective benchmark, but we need to be careful of our conclusions.

A related approach to developing grade-equivalent scores is to compute the median score for pupils tested at a particular point in time. Let's say, for example, we assess eight graders in their fourth month of school and find that their median score on the XYZ test of reading is 93. If a child is then administered the test and obtains a score of 93, that child is said to have a grade equivalent of 8.4. There is another problem with grade-equivalent scores and that is that school grades do not form an interval scale, even though the school year is approximately equal in length for any pupil. Simply consider the fact that a second grader who is one year behind his classmates in reading is substantially more "retarded" than an eighth grader who is one year behind.

Expectancy tables. Norms can be presented in a variety of ways. We can simply list the mean and SD for a normative group, or we can place the data in a table showing the raw scores and their equivalent percentiles, *T* scores, etc. For example,

Table 2–2

	Equivalent percentiles	
Raw score	Male	Female
47	99	97
46	98	95
45	98	93
44	97	90
43	96	86
42	94	81
etc.		

Table 2.2 gives some normative information, such as you might find in a test manual.

If we are using test scores to predict a particular outcome, we can incorporate that relationship into our table, and the table then becomes an *expectancy table*, showing what can be expected of a person with a particular score. Suppose, for example, we administer a test of mechanical aptitude to 500 factory workers. After 6 months we obtain for each worker supervisors' ratings indicating the quality of work. This situation is illustrated in Table 2.3. Note that there were 106 individuals who scored between 150 and 159. Of these 106.51 received ratings of excellent and 38 of above average. Assuming these are the type of workers we wish to hire, we would expect a new applicant to the company who scores between 150 and 159 to have a 89/106 or 84% chance to do well in that company. Note, on the other hand, that of the 62 individuals who scored between 60 and 69, only 1 achieved a rating of excellent, so that we would expect any new applicant with a score of 60–69 not to do well. In fact, we could calculate what score a person would need to obtain to be hired; such a score is called the *cutoff* score.

A few additional points follow about expectancy tables. Because we need to change the frequencies into percentages, a more useful expectancy table is one where the author has already done this for us. Second, decisions based on expectancy tables will not be foolproof. After all, one of the lowest scoring persons in our example turned out to be an excellent worker. An expectancy table is based on a sample that may have been representative at the time the data was collected, but may no longer be so. For example, our fictitious company might have gotten a reputation for providing excellent benefits, and so the applicant pool may be larger and more heterogeneous. Or the economy might have changed for the worse, so that candidates who never would have thought of doing manual labor are now applying for positions. To compute an expectancy table, we need to have the scores for both variables for a normative sample, and the two sets of scores must show some degree of correlation. Once the data are obtained for any new candidate, only the test score is needed to predict what the expected performance will be. Expectancy tables need not be restricted to two variables, but may incorporate more than one variable that is related to the predicted outcome.

Relativity of norms. John, a high-school student, takes a test of mechanical aptitude and obtains a score of 107. When we compare his score with available norms, we might find that his score is at the 85th percentile when compared

Table 2–3

	Supervisors' ratings				
Mechanical aptitude scores	Excellent	Above average	Average	Below average	Poor
150–159	51	38	16	0	1
140–149	42	23	8	3	0
130–139	20	14	7	2	1
120–129	16	9	3	0	0
110–119	0	2	4	7	8
100–109	1	0	3	12	16
90–99	1	0	0	14	19
80–89	2	1	2	23	23
70–79	0	1	0	19	26
60–69	1	0	0	30	31
Totals:	134	88	43	110	125

with the high-school sample reported in the test manual, that his score is at the 72nd percentile when compared with students at his own high school, and that his score is at the 29th percentile when compared with those applicants who have been admitted to the prestigious General Dynamics School of Automobile Training. Thus different normative groups give different meaning to a particular score, and we need to ask, "Which norm group is most meaningful?" Of course, that depends. If John is indeed aspiring to be admitted to the General Dynamics school, then that normative group is more meaningful than the more representative but "generic" sample cited in the test manual.

Local norms. There are many situations where local norms, data obtained from a local group of individuals, are more meaningful than any national norms that attempt to be representative. If decisions are to be made about an individual applicant to a particular college or a specific job, it might be better to have local norms; if career counseling is taking place, then national norms might be more useful. Local norms are desirable if we wish to compare a child's relative standing with other children in the same school or school district, and they can be especially useful when a particular district differs in language and culture from the national normative sample. How to develop local norms is described in some detail by Kamphaus and Lozano (1984), who give both general principles and a specific example.

Criterion-referenced testing. You might recall being examined for your driving license, either through a multiple choice test and/or a driving test, and being told, "Congratulations, you've passed." That decision did not involve comparing your score or performance against some norms, but rather comparing your performance against a *criterion*, a decision rule that was either explicit (you must miss less than 6 items to pass) or implicit (the examiner's judgment that you were skillful enough to obtain a driver's license).

Glaser (1963) first introduced the term *criterion-referenced* testing and since then the procedure has been widely applied, particularly in educational testing. The intent is to judge a person's performance on a test not on the basis of what others can do, but on the basis of some criterion. For example, we may define mental retardation not on the basis of a normative IQ, but whether a child of age 5 can show mastery of specific tasks such as buttoning her shirt, or following specific directions. Or we may admit a child to preschool on the basis of whether the child is toilet trained. Or we may administer a test of Spanish vocabulary and require 80% correct to register testees for Advanced Spanish.

Clearly, we must first of all be able to specify the criterion. Toilet training, mastery of elementary arithmetic, and automobile driving can all be defined fairly objectively, and generally agreed upon criteria can be more or less specified. But there are many variables, many areas of competency, where such criteria cannot be clearly specified.

Second, criteria are not usually arbitrary, but are based on real-life observation. Thus, we would not label a 5-year-old as mentally retarded if the child did not master calculus because few if any children of that age show such mastery. We would, however, expect a 5-year-old to be able to button his shirt. But that observation is in fact based upon norms; so criterion-referenced decisions can be normative decisions, often with the norms not clearly specified.

Finally, we should point out that criterion-referenced and norm-referenced refer to how the scores or test results are interpreted, rather than to the tests themselves. So Rebecca's score of 19 can be interpreted through norms or by reference to a criterion.

Criterion-referenced testing has made a substantial impact, particularly in the field of educational testing. To a certain degree, it has forced test constructors to become more sensitive to the domain being assessed, to more clearly and concretely specify the components of that domain, and to focus more on the concept of mastery of a particular domain (Carver, 1974; Shaycoft, 1979).

The term mastery is often closely associated with criterion-referenced testing, although other terms are used. Carver (1974) used the terms *psychometric* to refer to norm referenced and *edumetric* to refer to criterion referenced. He argued that the psychometric approach focuses on individual differences, and that item selection and the assessment of reliability and validity are determined by statistical procedures. The edumetric

approach, on the other hand, focuses on the measurement of gain or growth of individuals, and item selection, reliability and validity, all center on the notion of gain or growth.

COMBINING TEST SCORES

Typically, a score that is obtained on a test is the result of the scoring of a set of items, with items contributing equal weight, for example 1 point each, or different weights (item #6 may be worth one point, but item #18 may be worth 3 points). Sometimes, scores from various subtests are combined into a composite score. For example, a test of intelligence such as the Wechsler Adult Intelligence Scale is composed of eleven subtests. Each of these subtests yields a score, and six of these scores are combined into a Verbal IQ, while the other five scores are combined into a Performance IQ. In addition, the Verbal IQ and the Performance IQ are combined into a Full Scale IQ. Finally, scores from different tests or sources of information may be combined into a single index. A college admissions officer may, for example, combine an applicant's GPA, scores on an entrance examination, and interview information, into a single index to decide whether the applicant should be admitted. There are thus at least three basic ways of combining scores, and the procedures by which this is accomplished are highly similar (F. G. Brown, 1976).

Combining scores using statistics. Suppose we had administered ten different tests of "knowledge of Spanish" to Sharon. One test measured vocabulary, another, knowledge of verbs, still a third, familiarity with Spanish idioms, and so on. We are not only interested in each of these ten components, but we would like to combine Sharon's ten different scores into one index that reflects "knowledge of Spanish." If the ten tests were made up of one item each, we could of course simply sum up how many of the ten items were answered correctly by Sharon. With tests that are made up of differing number of items, we cannot calculate such a sum, since each test may have a different mean and standard deviation, that is represent different scales of measurement. This would be very much like adding a person's weight in pounds to their height in inches and their blood pressure in millimeters to obtain an index of "physical functioning." Statistically, we must equate each separate measurement before we add them up. One easy way to do this, is to change the raw scores into z scores or T scores. This would make all of Sharon's ten scores equivalent psychometrically, with each z score reflecting her performance on that variable (e.g., higher on vocabulary but lower on idioms). The ten z scores could then be added together, and perhaps divided by ten.

Note that we might well wish to argue, either on theoretical or empirical grounds, that each of the ten tests should not be given equal weight, that for example, the vocabulary test is most important and should therefore be weighted twice as much. Or if we were dealing with a scale of depression, we might argue that an item dealing with suicide ideation reflects more depression than an item dealing with feeling sad, and therefore should be counted more heavily in the total score. There are a number of techniques, both statistical and logical, by which *differential weighting* can be used, as opposed to *unit weighting*, where every component is given the same scoring weight (see Wang & Stanley, 1970). Under most conditions, unit weighting seems to be as valid as methods that attempt differential weighting (F. G. Brown, 1976).

Combining scores using clinical intuition. In many applied situations, scores are combined not in a formal, statistical manner, but in an informal, intuitive, judgmental manner. A college admissions officer for example, may consider an applicant's grades, letters of recommendation, test scores, autobiographical sketch, background variables such as high school attended, and so on, and combine all of these into a decision of "admit" or "reject." A personnel manager may review an applicant's file and decide on the basis of a global evaluation, to hire the candidate. This process of "clinical intuition" and whether it is more or less valid than a statistical approach has been studied extensively (e.g., Goldberg, 1968; Holt, 1958; Meehl, 1954; 1956; 1957). Proponents of the intuitive method argue that because each person is unique, only clinical judgment can encompass that uniqueness; that clinical judgment can take into account both complex and atypical patterns (the brilliant student who flunks high school but does extremely well in medical

school). Proponents of the statistical approach argue that in the long run, better predictive accuracy is obtained through statistical procedures, and that "intuition" operates inefficiently, if at all.

Multiple cutoff scores. One way to statistically combine test scores to arrive at a decision, is to use a *multiple cutoff* procedure. Let us assume we are an admissions officer at a particular college, looking at applications from prospective applicants. For each test or source of information we determine, either empirically or theoretically, a cutoff score that separates the range of scores into two categories, for example "accept" and "reject." Thus if we required our applicants to take an IQ test, we might consider an IQ of 120 as the minimum required for acceptance. If we also looked at high school GPA, we might require a minimum 86% overall for acceptance. These cutoff scores may be based on clinical judgment – "It is my opinion that students with an IQ less than 120 and high school GPA less than 86% do not do well here" – or on statistical evidence – a study of 200 incoming freshmen indicated that the flunk rate of those below the cutoff scores was 71% vs. 6% for those above the cutoff scores.

Note that using this system of multiple cutoffs, a candidate with an IQ of 200 but a GPA of 82% would not be admitted. Thus we need to ask whether superior performance on one variable can compensate for poor performance on another variable. The multiple cutoff procedure is a noncompensatory one and should be used only in such situations. For example, if we were selecting candidates for pilot training where both intelligence and visual acuity are necessary, we would not accept a very bright but blind individual.

There are a number of variations to the basic multiple cutoff procedure. For example, the decision need not be a dichotomy. We could classify our applicants as accept, reject, accept on probation, and hold for personal interview. We can also obtain the information sequentially. We might, for example, first require a college entrance admission test. Those that score above the cutoff score on that test may be required to take a second test or other procedure and may then be admitted on the basis of the second cutoff score.

Multiple regression. Another way of combining scores statistically is through the use of a multiple regression, which essentially expresses the relationship between a set of variables and a particular outcome that is being predicted. If we had only one variable, for example IQ, and are predicting GPA, we could express the relationship with a correlation coefficient, or with the equation of a straight line, namely:

$$Y = a + bX$$

where Y is the variable being predicted, in this case GPA
X is the variable we have measured, in this case IQ
b is the slope of the regression line (which tells us as X increases, by how much Y increases)
a is the intercept (that is, it reflects the difference in scores between the two scales of measurement; in this case GPA is measured on a 4-point scale while IQ has a mean of 100)

When we have a number of variables, all related statistically to the outcome, then the equation expands to:

$$Y = a + b_1x_1 + b_2x_2 + bx \ldots \text{etc.}$$

A nice example of a regression equation can be found in the work of Gough (1968) on a widely used personality test called the California Psychological Inventory (CPI). Gough administered the CPI to 60 airline stewardesses who had undergone flight training and had received ratings of in-flight performance (something like a final-exam grade). None of the 18 CPI scales individually correlated highly with such a rating, but a four-variable multiple regression not only correlated +.40 with the ratings of in-flight performance, but also yielded an interesting psychological portrait of the stewardesses. The equation was:

$$\text{In-flight rating} = 64.293 + .227(So) - 1.903(Cm) + 1.226(Ac) - .398(Ai)$$

where 64.293 is a weight that allows the two sides of the equation to be equated numerically,
So is the person's score on the Socialization scale

Cm is the person's score on the Communality scale
Ac is the person's score on the Achievement by Conformance scale
and *Ai* is the person's score on the Achievement by Independence scale

Notice that each of the four variables has a number and a sign (+ or −) associated with it. To predict a person's rating of in-flight performance we would plug in the scores on the four variables, multiply each score by the appropriate weight, and sum to solve the equation. Note that in this equation, Communality is given the greatest weight, and Socialization the least, and that two scales are given positive weights (the higher the scores on the *So* and *Ac* scales, the higher the predicted in-flight ratings), and two scales are given negative weights (the higher the scores the lower the predicted in-flight rating). By its very nature, a regression equation gives differential weighting to each of the variables.

The statistics of multiple regression is a complex topic and will not be discussed here (see J. Cohen & P. Cohen, 1983; Kerlinger & Pedhazur, 1973; Pedhazur, 1982; Schroeder, Sjoquist, & Stephan, 1986), but there are a number of points that need to be mentioned.

First of all, multiple regression is a compensatory model, that is, high scores on one variable can compensate for low scores on another variable. Second, it is a *linear* model, that is, it assumes that as scores increase on one variable (for example IQ), scores will increase on the predicted variable (for example, GPA). Third, the variables that become part of the regression equation are those that have the highest correlations with the criterion *and* low correlations with the other variables in the equation. Note that in the CPI example above, there were 18 potential variables, but only 4 became part of the regression equation. Thus, additional variables will not become part of the equation even if they correlate with the criterion but do not add something unique, that is, have low or zero correlations with the other variables. In most practical cases, regression equations are made up of about two to six variables. The variables that are selected for the equation are selected on the basis of statistical criteria, although their original inclusion in the study might have reflected clinical judgment.

Discriminant analysis. Another technique that is somewhat similar to multiple regression is that of *discriminant analysis.* In multiple regression, we place a person's scores in the equation, do the appropriate calculations, and out pops the person's predicted score on the variable of interest, such as GPA. In discriminant analysis we also use a set of variables, but this time we wish to predict group membership rather than a continuous score. Suppose for example, that there are distinct personality differences between college students whose life centers on academic pursuits (the "geeks") vs. students whose life centers on social and extracurricular activities (the "greeks"). John has applied to our university and we wish to determine whether he is more likely to be a geek or a greek. That is the aim of discriminant analysis. Once we know that two or more groups differ significantly from each other on a set of variables, we can assess an individual to determine which group that person most closely resembles. Despite the frivolous nature of the example, discriminant analysis has the potential to be a powerful tool in psychiatric diagnosis, career counseling, suicide prevention, and other areas (Tatsuoka, 1970).

SUMMARY

In this chapter we have looked at three basic issues: the construction, the administration, and the interpretation of tests. Test construction involves a wide variety of procedures, but for our purposes we can use a nine-step model to understand the process. Test items come in all shapes and forms, though some, like multiple choice, seem to be more common. Test construction is not a mere mechanical procedure, but in part involves some basic philosophical issues. A primary issue in test administration is that of establishing rapport. Once the test is administered and scored, the raw scores need to be changed into derived scores, including percentiles, standard scores, *T* scores, or stanines. Two aspects of test items are of particular interest to test constructors: item difficulty and item discrimination. Finally, we need to interpret a raw score

in terms of available norms or a criterion. Scores can also be combined in a number of ways.

SUGGESTED READINGS

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*, 481–489.

This article discusses the design, development, and evaluation of scales for use in counseling psychology research. Most of the methods discussed in this article will be covered in later chapters, but some of the basic issues are quite relevant to this chapter.

Hase, H. D., & Goldberg, L. R. (1967). Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin, 67*, 231–248.

This is an old but still fascinating report. The authors identify six strategies by which personality inventory scales can be developed. From the same item pool, they constructed sets of 11 scales by each of the 6 strategies. They then compared these 66 scales with 13 criteria. Which set of scales, which type of strategy, was the best? To find the answer, check the report out!

Henderson, M., & Freeman, C. P. L. (1987). A self-rating scale for bulimia. The "BITE." *British Journal of Psychiatry, 150*, 18–24.

There is a lot of interest in eating disorders, and these authors report on the development of a 36-item scale composed of two subscales – the Symptom Subscale and the Severity scale, designed to measure binge eating. Like the study by Zimmerman and Coryell (1987) listed next, this study uses fairly typical procedures, and reflects at least some of the steps mentioned in this chapter.

Nield, A. F. (1986). Multiple-choice questions with an option to comment: Student attitudes and use. *Teaching of Psychology, 13*, 196–199.

The author reports on a study where introductory psychology students were administered multiple-choice questions with an option to explain their answers. Such items were preferred by the students and found to be less frustrating and anxiety producing.

Zimmerman, M., & Coryell, W. (1987). The Inventory to Diagnose Depression (IDD): A self-report scale to diagnose major depressive disorder. *Journal of Consulting and Clinical Psychology, 55*, 55–59.

The authors report on the development of a 22-item self-report scale to diagnose depression. The procedures and methodologies used are fairly typical and most of the article is readable, even if the reader does not have a sophisticated statistical background.

DISCUSSION QUESTIONS

1. Locate a journal article that presents the development of a new scale (e.g., Leichsenring, 1999). How does the procedure compare and contrast with that discussed in the text?
2. Select a psychological variable that is of interest to you (e.g., intelligence, depression, computer anxiety, altruism, etc.). How might you develop a direct assessment of such a variable?
3. When your instructor administers an examination in this class, the results will most likely be reported as raw scores. Would derived scores be better?
4. What are the practical implications of changing item difficulty?
5. What kind of norms would be useful for a classroom test? For a test of intelligence? For a college entrance exam?

3 Reliability and Validity

AIM This chapter introduces the concepts of reliability and of validity as the two basic properties that every measuring instrument must have. These two properties are defined and the various subtypes of each discussed. The major focus is on a logical understanding of the concepts, as well as an applied understanding through the use of various statistical approaches.

INTRODUCTION

Every measuring instrument, whether it is a yardstick or an inventory of depression, must have two properties: the instrument must yield consistent measurement, i.e., must be reliable, and the instrument must in fact measure the variable it is said to measure, i.e., must be valid. These two properties, reliability and validity, are the focus of this chapter.

RELIABILITY

Imagine that you have a rich uncle who has just returned from a cruise to an exotic country, and he has brought you as a souvenir a small ruler – not a pygmy king, but a piece of wood with markings on it. Before you decide that your imaginary uncle is a tightwad, I should tell you that the ruler is made of an extremely rare wood with an interesting property – the wood shrinks and expands randomly – not according to humidity or temperature or day of the week, but randomly. If such a ruler existed it would be an interesting conversation piece, but as a measuring instrument it would be a miserable failure. Any measuring instrument must first of all yield *consistent* measurement; the actual measurement should not change unless what we are measuring changes. Consistency or reliability does not necessarily mean sameness. A radar gun that always indicates 80 miles per hour even when it is pointed at a stationary tree does not have reliability. Similarly, a bathroom scale that works accurately except for Wednesday mornings when the weight recorded is arbitrarily increased by three pounds, does have reliability.

Note that reliability is not a property of a test, even though we speak of the results as if it were (for example, "the test-retest reliability of the Jones Depression Inventory is .83"). Reliability really refers to the consistency of the data or the results obtained. These results can and do vary from situation to situation. Perhaps an analogy might be useful. When you buy a new automobile, you are told that you will get 28 miles per gallon. But the actual mileage will be a function of how you drive, whether you are pulling a trailer or not, how many passengers there are, whether the engine is well tuned, etc. Thus the actual mileage will be a "result" that can change as aspects of the situation change (even though we would ordinarily not expect extreme changes – even the most careful driver will not be able to decrease gas consumption to 100 miles per gallon) (see Thompson & Vacha-Haase, 2000).

True vs. error variation. What then is reliability? Consider 100 individuals of different heights.

When we measure these heights we will find variation, statistically measured by *variance* (the square of the standard deviation). Most of the variation will be "true" variation – that is, people really differ from each other in their heights. Part of the variation however, will be "error" variation, perhaps due to the carelessness of the person doing the measuring, or a momentary slouching of the person being measured, or how long the person has been standing up as opposed to lying down, and so on. Note that some of the error variation can be eliminated, and what is considered error variation in one circumstance may be a legitimate focus of study in another. For example, we may be very interested in the amount of "shrinkage" of the human body that occurs as a function of standing up for hours.

How is reliability determined? There are basically four ways: *test-retest* reliability, *alternate (or equivalent) forms* reliability, *split-half* reliability, and *interitem consistency.*

TYPES OF RELIABILITY

Test-retest reliability. You have probably experienced something like this: you take out your purse or wallet, count your money, and place the wallet back. Then you realize that something is not quite right, take the wallet out again and recount your money to see if you obtain the same result. In fact, you were determining test-retest reliability. Essentially then, *test-retest* reliability involves administering a test to a group of individuals and retesting them after a suitable interval. We now have two sets of scores for the same persons, and we compare the consistency of these two sets typically by computing a correlation coefficient. You will recall that the most common type of correlation coefficient is the Pearson product moment correlation coefficient, typically abbreviated as *r*, used when the two sets of scores are continuous and normally distributed (at least theoretically). There are other correlation coefficients used with different kinds of data, and these are briefly defined and illustrated in most introductory statistics books.

You will also recall that correlation coefficients can vary from zero, meaning that there is no relationship between one set of scores and the second set, to a plus or minus 1.00, meaning that there is a perfect relationship between one set of scores and the second. By convention, a correlation coefficient that reflects reliability should reach the value of .70 or above for the test to be considered reliable.

The determination of test-retest reliability appears quite simple and straightforward, but there are many problems associated with it. The first has to do with the "suitable" interval before retesting. If the interval is too short, for example a couple of hours, we may obtain substantial consistency of scores, but that may be more reflective of the relative consistency of people's memories over a short interval than of the actual measurement device. If the interval is quite long, for example a couple of years, then people may have actually changed from the first testing to the second testing. If everyone in our sample had changed by the same amount, for example had grown 3 inches, that would be no problem since the consistency (John is still taller than Bill) would remain. But of course, people don't change in just about anything by the same amount, so there would be inconsistency between the first and second set of scores, and our instrument would appear to be unreliable whereas in fact it might be keeping track of such changes. Typically, changes over a relatively longer period of time are not considered in the context of reliability, but are seen as "true" changes.

Usually then, test-retest reliability is assessed over a short period of time (a few days to a few weeks or a few months), and the obtained correlation coefficient is accompanied by a description of what the time period was. In effect, test-retest reliability can be considered a measure of the stability of scores over time. Different periods of time may yield different estimates of stability. Note also that some variables, by their very nature, are more stable than others. We would not expect the heights of college students to change over a two-week period, but we would expect changes in mood, even within an hour!

Another problem is related to motivation. Taking a personality inventory might be interesting to most people, but taking it later a second time might not be so exciting. Some people in fact might become so bored or resentful as to perhaps answer randomly or carelessly the second time around. Again, since not everyone would become careless to the same degree, retest scores would change differently for different people,

and therefore the proportion of error variation to true variation would become larger; hence the size of the correlation coefficient would be smaller.

There are a number of other problems with test-retest reliability. If the test measures some skill, the first administration may be perceived as a "practice" run for the second administration, but again not everyone will improve to the same degree on the second administration. If the test involves factual knowledge, such as vocabulary, some individuals might look up some words in the dictionary after the first administration and thus change their scores on the second administration, even if they didn't expect a retesting.

Alternate form reliability. A second way to measure reliability is to develop two forms of the same test, and to administer the two forms either at different times or in succession: Good experimental practice requires that to eliminate any practice or transfer effects, half of the subjects take form A followed by form B, and half take form B followed by form A. The two forms should be equivalent in all aspects – instructions, number of items, etc. – except that the items are different. This approach would do away with some of the problems mentioned above with test-retest reliability, but would not eliminate all of them.

If the two forms of the test are administered in rapid succession, any score differences from the first to the second form for a particular individual would be due to the item content, and thus reliability could be lowered due to *item sampling,* that is the fact that our measurement involves two different samples of items, even though they are supposed to be equivalent. If the two forms are administered with some time interval between them, then our reliability coefficient will reflect the variation due to both item sampling and temporal aspects.

Although it is desirable to have alternate forms of the same test to reduce cheating, to assess the effectiveness of some experimental treatment, or to maintain the security of a test (as in the case of the GRE), the major problem with alternate form reliability is that the development of an alternate form can be extremely time consuming and sometimes simply impossible to do, particularly for tests that are not commercially published. If we are developing a test to measure knowledge of arithmetic in children, there is almost an infinite number of items we can generate for an alternate form, but if we are developing a test to assess depression, the number of available items related to depression is substantially smaller.

Let's assume you have developed a 100-item, multiple-choice vocabulary test composed of items such as:

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

You have worked for five years on the project, tried out many items, and eliminated those that were too easy or too difficult, those that showed gender differences, those that reflected a person's college major, and so on. You now have 100 items that do not show such undue influences and are told that you must show that your vocabulary test is indeed reliable. Test-retest reliability does not seem appropriate for the reasons discussed above. In effect, you must go back and spend another 5 years developing an alternate form. Even if you were willing to do so, you might find that there just are not another 100 items that are equivalent. Is there a way out? Yes, indeed there is; that is the third method of assessing reliability, known as split-half reliability.

Split-half reliability. We can administer the 100-item vocabulary test to a group of subjects, and then for each person obtain two scores, the number correct on even-numbered items and the number correct on odd-numbered items. We can then correlate the two sets of scores. In effect, we have done something that is not very different from alternate-form reliability; we are making believe that the 100-item test is really two, 50-item tests. The reliability estimate we compute will be affected by item sampling – the odd-numbered items are different from the even-numbered items, but will not be affected by temporal stability because only one administration is involved.

There is however, an important yet subtle difference between split-half reliability and test-retest. In test-retest, reliability was really a reflection of temporal stability; if what was being measured did not appreciably change over time, then our measurement was deemed consistent or reliable. In split-half reliability the focus of consistency has changed. We are no longer concerned about temporal stability, but are now concerned with *internal consistency.* Split-half reliability makes sense to the degree that each item in

our vocabulary test measures the same variable, that is to the degree that a test is composed of *homogeneous* items. Consider a test to measure arithmetic where the odd-numbered items are multiplication items and the even-numbered items deal with algebraic functions. There may not be a substantial relationship between these two areas of arithmetic knowledge, and a computed correlation coefficient between scores on the two halves might be low. This case should not necessarily be taken as evidence that our test is unreliable, but rather that the split-half procedure is applicable only to homogeneous tests. A number of psychologists argue that indeed most tests should be homogeneous, but other psychologists prefer to judge tests on the basis of how well they work rather than on whether they are homogeneous or heterogeneous in composition. In psychological measurement, it is often difficult to assess whether the items that make up a scale of depression, or anxiety, or self-esteem are psychometrically consistent with each other or reflect different facets of what are rather complex and multidimensional phenomena.

There are of course many ways to split a test in half to generate two scores per subject. For our 100-item vocabulary test, we could score the first 50 items and the second 50 items. Such a split would ordinarily not be a good procedure because people tend to get more tired toward the end of a test and thus would be likely to make more errors on the second half. Also, items are often presented within a test in order of difficulty, with easy items first and difficult items later; this might result in almost everyone getting higher scores on the first half of the test and differing on the second half – a state of affairs that would result in a rather low correlation coefficient. You can probably think of more complicated ways to split a test in half, but the odd vs. even method usually works well. In fact, split-half reliability is often referred to as odd-even reliability.

Each half score represents a sample, but the computed reliability is based only on half of the items in the test, because we are in effect comparing 50 items vs. 50 items, rather than 100 items. Yet from the viewpoint of item sampling (not temporal stability), the longer the test the higher will its reliability be (Cureton, 1965; Cureton, et al., 1973). All other things being equal, a 100-item test will be more reliable than a 50-item test – going to a restaurant 10 different times will give you a more "stable" idea of what the chef can do than only two visits. There is a formula that allows us to estimate the reliability of the entire test from a split-half administration, and it is called the *Spearman-Brown* formula:

$$\text{estimated } r = \frac{k\ (\text{obtained } r)}{1 + (k-1)(\text{obtained } r)}$$

In the formula, k is the number of times the test is lengthened or shortened. Thus, in split-half reliability, k becomes 2 because we want to know the reliability of the entire test, a test that is twice as long as one of its halves. But the Spearman-Brown formula can be used to answer other questions as these examples indicate:

EXAMPLE 1 I have a 100-item test whose split-half reliability is .68. What is the reliability of the total test?

$$\text{estimated } r = \frac{2(.68)}{1 + (1)(.68)} = \frac{1.36}{1.68} = \boxed{.81}$$

EXAMPLE 2 I have a 60-item test whose reliability is .61; how long must the test be for its reliability to be .70? (Notice we need to solve for k.)

$$.70 = \frac{k(.61)}{1 + (k-1)(.61)}$$

cross-multiplying we obtain:

$$\begin{aligned} k(.61) &= .70 + .70(k-1)(.61) \\ k(.61) &= .70 + (.427)(k-1) \\ k(.61) &= .70 + .427k - .427 \\ k(.183) &= .273 \\ k &= \boxed{1.49} \end{aligned}$$

the test needs to be about 1.5 times as long or about 90 items (60×1.5).

EXAMPLE 3 Given a 300-item test whose reliability is .96, how short can the test be to have its reliability be at least .70? (Again, we are solving for k.)

$$\begin{aligned} .70 &= \frac{k(.96)}{1 + (k-1)(.96)} \\ k(.96) &= .70 + .70(.96)(k-1) \\ k(.96) &= .70 + .672(k-1) \\ k(.96) &= .70 + .672k - .672 \\ k(.96) &= .028 = .672k \\ k(.288) &= 0.28 \\ k &= \boxed{.097} \end{aligned}$$

The test can be about one tenth of this length, or 30 items long (300 × .097).

The calculations with the Spearman-Brown formula assume that when a test is shortened or lengthened, the items that are eliminated or added are all equal in reliability. In fact such is not the case, and it is quite possible to increase the reliability of a test by eliminating the least reliable items. In this context, note that reliability can be applied to an entire test or to each item.

The Rulon formula. Although the Spearman-Brown formula is probably the most often cited and used method to compute the reliability of the entire test, other equivalent methods have been devised (e.g., Guttman, 1945; Mosier, 1941; Rulon, 1939). The Rulon formula is:

$$\text{estimated } r = 1 - \frac{\text{variance of differences}}{\text{variance of total scores}}$$

For each person who has taken our test, we generate four scores: the score on the odd items; the score on the even items, a difference score (score on the odd items minus score on the even items), and a total score (odd plus even). We then compute the variance of the difference scores and the variance of the total scores to plug into the formula. Note that if the scores on the two halves were perfectly consistent, there would be no variation between the odd item score and the even item score, and so the variance of the difference scores would be zero, and therefore the estimated *r* would equal 1. The ratio of the two variances in fact reflects the proportion of error variance that when subtracted from 1 leaves the proportion of "true" variance, that is, the reliability.

Variability. As discused in Chapter 2, variability of scores among individuals, that is, *individual differences,* makes statistical calculations such as the correlation coefficient possible. The item, "Are you alive as you read this?" is not a good test item because it would yield no variability – everyone presumably would give the same answer. Similarly, gender as defined by "male" or "female" yields relatively little variability, and from a psychometric point of view, gender thus defined is not a very useful measure. All other things being equal, the greater the variability in test scores the better off we are. One way to obtain such variability is to increase the range of responses. For example, instead of just asking do you agree or disagree, we could use a five-point response scale of strongly agree, agree, undecided, disagree, strongly disagree. Another way to increase variability is to increase the number of items – a 10-item true-false scale can theoretically yield scores from 0 to 10, but a 25-item scale can yield scores from 0 to 25, and that of course is precisely the message of the Spearman-Brown formula. Still another way to increase variability is to develop test items that are neither too easy nor too difficult for the intended consumer, as we also discussed in Chapter 2. A test that is too easy would result in too many identical high scores, and a test that is too difficult would result in too many identical low scores. In either case, variability, and therefore reliability, would suffer.

Two halves = four quarters. If you followed the discussion up to now, you probably saw no logical fallacy in taking a 100-item vocabulary test and generating two, scores for each person, as if in fact you had two, 50-item tests. And indeed there is none. Could we not argue however, that in fact we have 4 tests of 25 items each, and thus we could generate four scores for each subject? After all, if we can cut a pie in two, why not in four? Indeed, why not argue that we have 10 tests of 10 items each, or 25 tests of 4 items each, or 100 tests of 1 item each! This leads us to the fourth way of determining reliability, known as *interitem consistency.*

Interitem consistency. This approach assumes that each item in a test is in fact a measure of the same variable, whatever that may be, and that we can assess the reliability of the test by assessing the consistency among items. This approach rests on two assumptions that are often not recognized even by test "experts." The first is that interitem reliability, like split-half reliability, is applicable and meaningful only to the extent that a test is made up of *homogeneous* items, items that all assess the same domain. The key word of course is "same." What constitutes the same domain? You have or will be taking an examination in this course, most likely made up of multiple-choice items. All of the items focus on your knowledge of psychological testing, but some of the items may require rote memory, others, recognition of key words, still others, the ability to reason logically,

and others, perhaps the application of formulas. Do these items represent the same or different domains? We can partially answer this statistically, through *factor analysis.* But if we compute an interitem consistency reliability correlation coefficient, and the resulting *r* is below .70, we should not necessarily conclude that the test is unreliable.

A second assumption that lurks beneath interitem consistency is the notion that if each item were perfectly reliable, we would only obtain two test scores. For example, in our 100-item vocabulary test, you would either know the meaning of a word or you would not. If all the items are perfectly consistent, they would be perfectly related to each other, so that people taking the test would either get a perfect score or a zero. If that is the case, we would then only need 1 item rather than 100 items. In fact, in the real world items are not perfectly reliable or consistent with each other, and the result is individual differences and variability in scores. In the real world also, people do not have perfect vocabulary or no vocabulary, but differing amounts of vocabulary.

Measuring interitem consistency. How is interitem consistency measured? There are two formulas commonly used. The first is the Kuder-Richardson formula 20, sometimes abbreviated as K-R 20 (Kuder & Richardson, 1937), which is applicable to tests whose items can be scored on a dichotomous (e.g., right-wrong; true-false; yes-no) basis. The second formula is the coefficient alpha, also known as Cronbach's alpha (Cronbach, 1951), for tests whose items have responses that may be given different weights – for example, an attitude scale where the response "never" might be given 5 points, "occasionally" 4 points, etc. Both of these formulas require the data from only one administration of the test and both yield a correlation coefficient. It is sometimes recommended that Cronbach's alpha be at least .80 for a measure to be considered reliable (Carmines & Zeller, 1979). However alpha increases as the number of items increases (and also increases as the correlations among items increase), so that .80 may be too harsh of a criterion for shorter scales. (For an in-depth discussion of coefficient alpha, see Cortina, 1993).

Sources of error. The four types of reliability just discussed all stem from the notion that a test score is composed of a "true" score plus an "error" component, and that reliability reflects the relative ratio of true score variance to total or observed score variance; if reliability were perfect, the error component would be zero.

A second approach to reliability is based on *generalizability* theory, which does not assume that a person has a "true" score on intelligence, or that error is basically of one kind, but argues that different conditions may result in different scores, and that error may reflect a variety of sources (Brennan, 1983; Cronbach, Gleser, Rajaratnam, & Nanda, 1972; see Lane, Ankenmann, & Stone, 1996, for an example of generalizability theory as applied to a Mathematics test). The interest here is not only in obtaining information about the sources of error, but in systematically varying those sources and studying error experimentally. Lyman (1978) suggested five major sources of error for test scores:

1. The individual taking the test. Some individuals are more motivated than others, some are less attentive, some are more anxious, etc.
2. The influence of the examiner, especially on tests that are administered to one individual at a time. Some of these aspects might be whether the examiner is of the same race, gender, etc., as the client, whether the examiner is (or is seen as) caring, authoritarian, etc.
3. The test items themselves. Different items elicit different responses.
4. Temporal consistency. For example, intelligence is fairly stable over time, but mood may not be.
5. Situational aspects. For example, noise in the hallway might distract a person taking a test.

We can experimentally study these sources of variation and statistically measure their impact, through such procedures as analysis of variance, to determine which variables and conditions create lessen reliability. For example, whether the retest is 2 weeks later or 2 months later might result in substantial score differences on test X, but whether the administrator is male or female might result in significant variation in test scores for male subjects but not for female subjects. (See Brennan, 1983, or Shavelson, Webb, & Rowley,

1989, for a very readable overview of generalizability theory.)

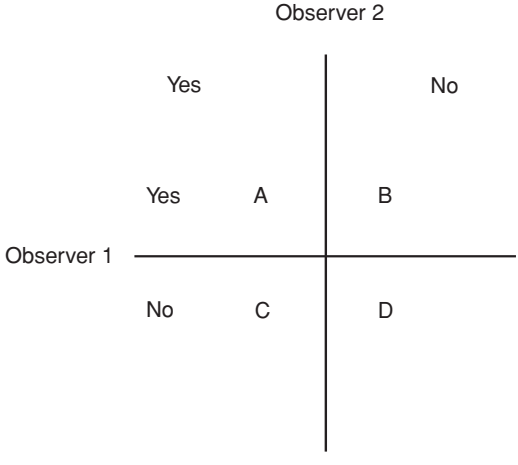
Scorer reliability. Many tests can be scored in a straightforward manner: The answer is either correct or not, or specific weights are associated with specific responses, so that scoring is primarily a clerical matter. Some tests however, are fairly subjective in their scoring and require considerable judgment on the part of the scorer. Consider for example, essay tests that you might have taken in college courses. What constitutes an "A" response vs. a "B" or a "C" can be fairly arbitrary. Such tests require that they be reliable not only from one or more of the standpoints we have considered above, but also from the viewpoint of scorer reliability – would two different scorers arrive at the same score when scoring the same test protocol? The question is answered empirically; a set of test protocols is independently given to two or more scorers and the resulting two or more sets of scores are compared, usually with a correlation coefficient, or sometimes by indicating the percentage of agreement (e.g., Fleiss, 1975).

Quite often, the scorers need to be trained to score the protocols, especially with scoring sophisticated psychological techniques such as the Rorschach inkblot test, and the resulting correlation coefficient can be in part reflective of the effectiveness of the training. Note that, at least theoretically, an objectively scored test could have a very high reliability, but a subjectively scored version of the same test would be limited by the scorer reliability (for example, our 100-item vocabulary test could be changed so that subjects are asked to define each word and their definitions would be judged as correct or not). Thus, one way to improve reliability is to use test items that can be objectively scored, and that is one of several reasons why psychometricians prefer multiple-choice items to formats such as essays.

Rater reliability. Scorer reliability is also referred to as rater reliability, when we are dealing with ratings. For example, suppose that two faculty members independently read 80 applications to their graduate program and rate each application as "accept," "deny," or "get more information." Would the two faculty members agree with each other to any degree?

Chance. One of the considerations associated with scorer or rater reliability is chance. Imagine two raters observing a videotape of a therapy session, and rating the occurrence of every behavior that is reflective of anxiety. By chance alone, the observers could agree 50% of the time, so our reliability coefficient needs to take this into account: What is the actual degree of agreement over and above that due to chance? Several statistical measures have been proposed, but the one that is used most often is the *Kappa coefficient* developed by Cohen (1960; see also Hartmann, 1977). We could of course have more than two raters. For example, each application to a graduate program might be independently rated by three faculty members, but not all applications would be rated by the same three faculty. Procedures to measure rater reliability under these conditions are available (e.g., Fleiss, 1971).

Interobserver reliability. At the simplest level, we have two observers independently observing an event – e.g., did Brian hit Marla? Schematically, we can describe this situation as:



Cells A and D represent agreements, and cells B and C represent disagreements. From this simple schema some 17 different ways of measuring observer reliability have been developed, although most are fairly equivalent (A. E. House, B. J. House, & Campbell, 1981). For example, we can compute percentage agreement as:

$$\text{Percentage agreement} = \frac{A + D}{A + B + C + D} \times 100$$

From the same schema we can also compute coefficient Kappa, which is defined as:

$$\frac{Po - Pe}{1 - Pe}$$

where *Po* is the observer proportion of agreement and *Pe* is the expected or chance agreement.

To calculate Kappa, see Fleiss (1971) or Shrout, Spitzer, and Fleiss (1987).

Correction for attenuation. Reliability that is less than perfect, as it typically is, means that there is "noise in the system," much like static on a telephone line. But just as there are electronic means to remove that static, there are statistical means by which we can estimate what would happen if we had a perfectly reliable test. That procedure is called correction for attenuation and the formula is:

$$r_{\text{estimated}} = \frac{r_{12}}{\sqrt{r_{11} r_{22}}}$$

where $r_{\text{estimated}}$ is the "true" correlation between two measures if both the test and the second measure were perfectly reliable;
r_{12} is the observed correlation between the test and the second measure;
r_{11} is the reliability of the test; and
r_{22} is the reliability of the second measure.

For example, assume there is a correlation between the Smith scholastic aptitude test and grades of .40; the reliability of the Smith is .90 and that of grades is .80. The estimated true correlation between the Smith test and GPA is:

$$= \frac{.40}{(.90)(.80)} = \frac{.40}{.85} = \boxed{.47}$$

You might wonder how the reliability of GPA might be established? Ordinarily of course, we would have to assume that grades are measured without error because we cannot give grades twice or compare grades in the first three courses one takes vs. the last three courses in a semester. In that case, we would assign a 1 to r_{22} and so the formula would simplify to:

$$r_{\text{estimated}} = \frac{r_{12}}{\sqrt{r_{11}}}$$

The standard error of measurement. Knowing the reliability coefficients for a particular test gives us a picture of the stability of that test. Knowing for example, that the test-retest reliability of our 100-item vocabulary test is .92 over a 6-month period tells us that our measure is fairly stable over a medium period of time; knowing that in a sample of adults, the test-retest reliability is .89 over a 6-year period, would also tell us that vocabulary is not easily altered by differing circumstances over a rather long period of time. Notice however, that to a certain degree this approach does not focus on the individual subject. To compute reliability the test constructor simply administers the test to a group of subjects, chosen because of their appropriateness (e.g., depressed patients) or quite often because of their availability (e.g., college sophomores). Although the obtained correlation coefficient does reflect the sample upon which it is based, the psychometrician is more interested in the test than in the subjects who took the test. The professional who uses a test, however, a clinical psychologist, a personnel manager, or a teacher, is very interested in the individual, and needs therefore to assess reliability from the individual point of view. This is done by computing the *standard error of measurement (SEM)*.

Imagine the following situation. I give Susan, a 10-year-old, an intelligence test and I calculate her IQ, which turns out to be 110. I then give her a magic pill that causes amnesia for the testing, and I retest her. Because the test is not perfectly reliable, because Susan's attention might wander a bit more this second time, and because she might make one more lucky guess this time, and so on, her IQ this second time turns out to be 112. I again give her the magic pill and test her a third time, and continue doing this about 5,000 times. The distribution of 5,000 IQs that belong to Susan will differ, not by very much, but perhaps they can go from a low of 106 to a high of 118. I can compute the mean of all of these IQs and it will turn out that the mean will in fact be her "true" IQ because error deviations are assumed to cancel each other out – for every lucky guess there will be an unlucky guess. I can also calculate the variation of these 5,000 IQs by computing the standard deviation. Because this is a very special standard deviation (for one thing, it is a theoretical notion based on an impossible

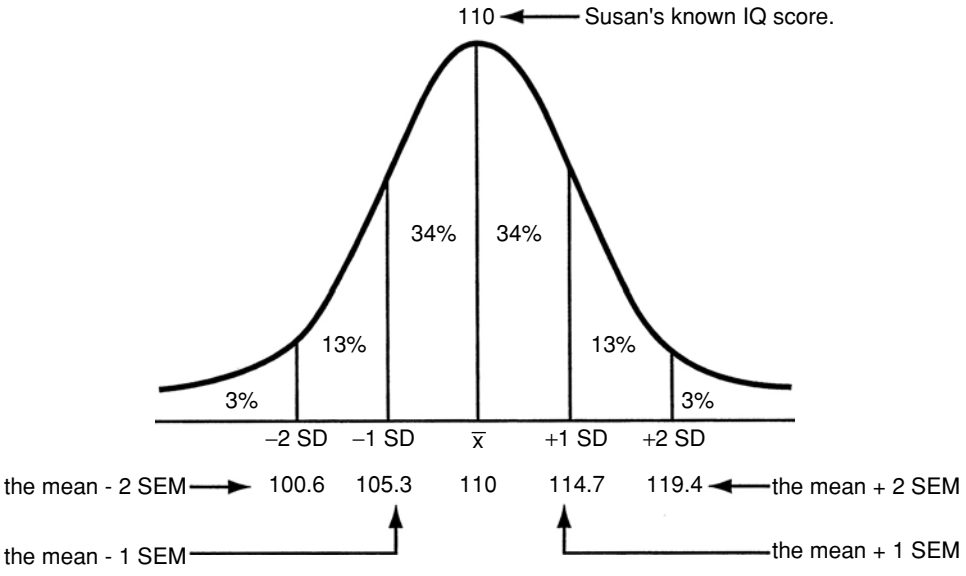


FIGURE 3–1. Hypothetical distribution of Susan's IQ scores.

example), it is given a special name: the standard error of measurement or SEM (remember that the word standard really means average). This SEM is really a standard deviation: it tells us how variable Susan's scores are.

In real life of course, I can only test Susan once or twice, and I don't know whether the obtained IQ is near her "true" IQ or is one of the extreme values. I can however, compute an estimate of the SEM by using the formula:

$$\text{SEM} = \text{SD}\sqrt{1 - r_{11}}$$

where SD is the standard deviation of scores on the test, and
r_{11} is the reliability coefficient.

Let's say that for the test I am using with Susan, the test manual indicates that the SD = 15 and the reliability coefficient is .90. The SEM is therefore equal to:

$$15\sqrt{(1-.90)} \text{ or } 4.7$$

How do we use this information? Remember that a basic assumption of statistics is that scores, at least theoretically, take on a normal curve distribution. We can then imagine Susan's score distribution (the 5,000 IQs if we had them) to look like the graph in Figure 3.1.

We only have one score, her IQ of 110, and we calculated that her scores would on the average deviate by 4.7 (the size of the SEM). Therefore, we can assume that the probability of Susan's "true" IQ being between 105.3 and 114.7 is 68%, and that the probability of her "true" IQ being between 100.6 and 119.4 is 94%. Note that as the SD of scores is smaller and the reliability coefficient is higher, the SEM is smaller. For example, with an SD of 5, the

$$\text{SEM} = 5\sqrt{(1-.90)} = 1.58$$

with an SD of 5 and a reliability coefficient of .96 the

$$\text{SEM} = 5\sqrt{(1-.96)} = 1.$$

Don't let the statistical calculations make you lose sight of the logic. When we administer a test there is "noise in the system" that we call error or lack of perfect reliability. Because of this, an obtained score of 120 could actually be a 119 or a 122, or a 116 or a 125. Ordinarily we don't expect that much noise in the system (to say that Susan's IQ could be anywhere between 10 and 300 is not very useful) but in fact, most of the time, the limits of a particular score are relatively close together and are estimated by the SEM, which reflects the reliability of a test as applied to a particular individual.

The SE of differences. Suppose we gave Alicia a test of arithmetic and a test of spelling. Let's assume that both tests yield scores on the same

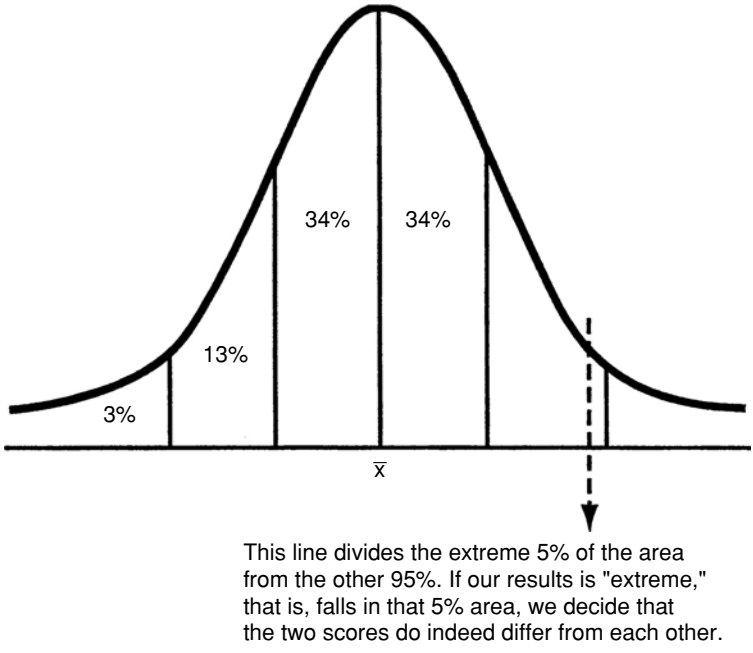


FIGURE 3–2. Normal curve distribution.

numerical scale – for example, an average of 100 and a SD of 10 – and that Alicia obtains a score of 108 on arithmetic and 112 on spelling. Can we conclude that she did better on the spelling test? Because there is "noise" (that is, unreliability) on both tests, that 108 on arithmetic could be 110, and that 112 on spelling could be 109, in which case we would not conclude that she did better on spelling. How can we compare her two scores from a reliability framework? The answer again lies in the standard error, this time called the *standard error of differences, SED*. Don't lose sight of the fact that the SE is really a SD telling us by how much the scores deviate on the average.

The formula for the SED is:

$$\text{SED} = \sqrt{(\text{SEM})_1^2 + (\text{SEM})_2^2}$$

which turns out to be equal to

$$\text{SED} = \text{SD}\sqrt{2 - r_{11} - r_{22}}$$

where the first SEM and the first r refer to the first test
and the second SEM and the second r refer to the second test
and SD = the standard deviation (which is the same for both tests).

Suppose for example, that the two tests Alicia took both have a SD of 10, and the reliability of the arithmetic test is .95 and that of the spelling test is .88. The SED would equal:

$$10\sqrt{2 - .95 - .88} \text{ or } 4.1.$$

We would accept Alicia's two scores as being different, if the probability of getting such a difference by chance alone is 5 or fewer times out of 100, i.e., $p < .05$. You will recall that such a probability can be mapped out on the normal curve to yield a z score of $+1.96$. We would therefore take the SED of 4.1 and multiply it by 1.96 to yield approximately 8, and would conclude that Alicia's two scores are different only if they differ by at least 8 points; in the example above they do not, and therefore we cannot conclude that she did better on one test than the other (see Figure 3.2).

Reliability of difference scores. Note that in the above section we focused on the difference between two scores. Quite often the clinical psychologist, the school or educational psychologist, or even a researcher, might be more interested in the relationship of pairs of scores rather than individual single scores; we might for example be interested in relating discrepancies between verbal and nonverbal intelligence to evidence of

possible brain damage, and so we must inquire into the reliability of difference scores. Such reliability is not the sum of the reliability of the two scores taken separately because the difference score is not only affected by the errors of measurement of each test, but is also distinguished by the fact that whatever is common to both measures is canceled out in the difference score – after all, we are looking at the difference. Thus the formula for the reliability of difference scores is:

$$r_{\text{difference}} = \frac{\frac{1}{2}(r_{11} + r_{22}) - r_{12}}{1 - r_{12}}$$

For example, if the reliability of test A is .75 and that of test B is .90, and the correlation between the two tests is .50 then

$$r_{\text{difference}} = \frac{\frac{1}{2}(.75 + .90) - .50}{1 - .50} = \frac{.325}{.50} = \boxed{.65}$$

In general, when the correlation between two tests begins to approach the average of their separate reliability coefficients, the reliability of the difference score lowers rapidly. For example, if the reliability of test A is .70, that of test B is also .70, and the correlation between the two tests is .65, then

$$r_{\text{difference}} = \frac{\frac{1}{2}(.70 + .70) - .65}{1 - .65} = \frac{.05}{.35} = \boxed{.14}$$

The point here is that we need to be very careful when we make decisions based on difference scores. We should also reiterate that to compare the difference between two scores from two different tests, we need to make sure that the two scores are on the same scale of measurement; if they are not, we can of course change them to z scores, T scores, or some other scale.

Special circumstances. There are at least two categories of tests where the determination of reliability requires somewhat more careful thinking. The first of these are speeded tests where different scores reflect different rates of responding. Consider for example a page of text where the task is to cross out all the letters "e" with a time limit of 40 seconds. A person's score will simply reflect how fast that person responded to the task. Both test-retest and equivalent forms reliability are applicable to speeded tests, but split-half and internal consistency are not, unless the split is based on time rather than number of items.

A second category of tests, requiring special techniques, are criterion-referenced tests, where performance is interpreted not in terms of norms but in terms of a pass-fail type of decision (think of an automobile driving test where you are either awarded a license or not). Special techniques have been developed for such tests (e.g., Berk, 1984).

VALIDITY

Consider the following: Using a tape measure, measure the circumference of your head and multiple the resulting number by 6.93. To this, add three times the number of fingers on your left hand, and six times the number of eyeballs that you have. The resulting number will be your IQ. When I ask students in my class to do this, most stare at me in disbelief, either wondering what the point of this silly exercise is, or whether I have finally reached full senility! The point, of course, is that such a procedure is extremely reliable, assuming your head doesn't shrink or expand, and that you don't lose any body parts between test and retest. But reliability is not sufficient.

Once we have established that a test is reliable, we must show that it is also *valid,* that it measures what it is intended to measure. Does a test of knowledge of arithmetic really measure that knowledge, or does it measure the ability to follow directions, to read, to be a good guesser, or general intelligence? Whether a test is or is not valid depends in part on the specific purpose for which it is used. A test of knowledge of arithmetic may measure such knowledge in fifth graders, but not in college students. Thus validity is not a matter of "is this test valid or not" but is the test valid for this particular purpose, in this particular situation, with these particular subjects. A test of academic aptitude may be predictive of performance at a large state university but not at a community college. From a classical point of view, there are three major categories of validity, and these are called content validity, criterion validity, and construct validity. The division of validity into various parts has been objected to by many (e.g., Cronbach, 1980; Guion, 1980; Messick, 1975; Tenopyr, 1977). As Tenopyr and Oeltjen (1982) stated, it is difficult to imagine a measurement situation that does not involve all aspects of validity. Although these will be presented as separate categories, they really are