Table of Contents

Lesson 01

INTRODUCTION TO PSYCHOLOGICAL TESTING

Learning objectives

After completing this lesson, you would be able to do the following:

- Define the term tests, psychological tests and psychological testing.
- List and describe the three characteristics defining the tests.
- Differentiate between objective and subjective scoring rules
- Describe how tests can be used for rating, placement, selection, competency and proficiency, diagnosis and evaluation.
- Describe the types of tests.
- Differentiate between psychological test and psychological assessment.
- Understand why the control in the use of psychological test is required.
- Identify source books for locating information about tests.
- Describe the ethical issues involved in testing.
- Describe the standards of testing

Test

A test is a measurement device or technique used to quantify behavior or aid in understanding and prediction of behavior. (Kaplan and Saccuzo)

Psychological test

A psychological test is a set of items designed to measure the characteristics of human beings that pertain to behavior. (Kaplan and Saccuzo)

Psychological tests are written, visual or verbal evaluations administered to assess the cognitive and emotional functioning of children and adults.

Testing

Testing refers to the process of administrating, scoring and interpreting psychological tests. (Cohen and Philips)

Three characteristics of psychological tests

1. A psychological test is a sample of behavior.
2. The sample of behavior is obtained under standardized settings.
3. There are established scoring rules for obtaining quantitative information from the behavior sample.

Characteristics of Psychological Tests

1. **A psychological test is a sample of behavior**

All Psychological Tests require the respondent to do something. The subject behavior is used to measure some specific attribute or to predict some specific outcome. Therefore, a variety of measure that do not require the respondent to engage in any overt behavior or that require behavior on the part of the subject that is clearly incidental to what ever is being measured(e.g., a stress EKG) fall outside the domain of psychological tests.

The use of behavior samples in psychological tests has several implications. First, all Psychological Tests are not exhaustive measurements of all behaviors that could be used in measuring and defining a particular attribute. Suppose you wished to develop a test to measure a persons writing ability. One strategy would be to collect and evaluate everything that person had ever written. Such procedure would be highly accurate but impractical. A psychological tests attempts to approximate this exhaustive procedure by collecting a systematic sample of behavior. In this case, a writing test might include a series of short tests, samples essays, memos and like that.

The second implication to using behavior samples to measure psychological variables is that the quality of the test is largely determined by the representativeness of the sample. A good Psychological test is a representative sample of the measured behavior. The behavior elicited by the test must somehow be representative of behaviors that could be outside the testing situations. For example if a scholastic aptitude test is administered in a burning building, it is unlikely that the student's responses to that test would tell us about his/her aptitude. There should be a clear connection between the test and the measured behavior in a real world setting.

2. **The behavior sample is obtained under standardized conditions.**

Each individual taking a psychological test should be tested under essentially identical conditions. The conditions under which a test is administered are certain to affect the behavior of the people or persons taking the test. A student is likely to do better on a test that is given in a regular classroom environment than he or she would if the same test were given in a hot noisy auditorium.

For example, SAT administration instructions pertain to:

Seating Arrangements
Lighting Conditions
Noise Levels
Interruptions
Answering common questions

Standardization is vital because many test results are referential in nature: Your performance is measured relative to everybody else's performance. Standardization reduces between subject variability due to extraneous variables.

It is not possible to achieve the same degree of standardization with all psychological tests. A high degree of standardization may be possible with many written tests, although even within this class of tests the condition of testing might be very difficult to control. Standardization is easier to obtain with tests designed to be administered on masses.

The greatest difficulty in standardization, however probably lies in the broad class of tests that are administered verbally on an individually basis. For example, tests such as the Wechsler Adult Intelligence Scale (WAIS), which are administered individually, are less standardized. WAIS represents one of the best individual intelligence tests and it is administered verbally by a psychologist. It is likely that an examinee will respond differently to a friendly, calm examiner than to the one who is threatening and surely. The individual giving the test is an important variable.

Individually administered tests are difficult to standardize because the examinee is an integral part of the test. The same test given to the same subject by two different examiners is certain to elicit a somewhat different set of behaviors.

Often, psychologists take special training to standardize the way they give the test. Strict adherence to standard procedures for administering various psychological tests helps to minimize the effects of extraneous variables such as physical conditions of testing, the characteristics of an examiner or the subject confusions regarding the demands of the tests.

3. **There are established scoring rules for obtaining quantitative information from the behavior sample.**

Objective Scoring Rules:

Most mass produced tests fall into this category. Different qualified examiners will all come to the same score for an identical set of responses. For example two teachers scoring the same multiple choice test will arrive at the same total score.

Subjective Scoring Rules:

Here the judgment of the examiner is an important part of the test; different examiners can legitimately come to different conclusions concerning the same sample of behavior. The procedure a teacher follows in grading an essay exam provides an example of subjective scoring rules.

Tests vary considerably in the precision and detail of their scoring rules. For multiple choice tests, it is possible to state beforehand the exact score that will be assigned to every possible combination of answers. For an unstructured test, such as the Rorschach inkblot test, in which the subject describes his or her interpretations of an ambiguous abstract figure, general principles for scoring may be described, but it may be impossible to arrive at exact, objective scoring rules.

Good standardized psychological tests all have a set of rules or procedures for scoring responses to a test.

Uses of psychological tests

The question that should be addressed in psychological testing is "why psychological testing is important?" There are several possible answers to this question but we believe that the best answer lies in the simple statement that forms the central theme of our lesson: tests are used to make important decisions a about individuals.

Following are stated the uses of tests.

1. **Rating**

We use tests to rate people when test data help determine where they fall relative to their peers or some standard of performance. Rating, therefore, involves using tests scores to represent an individual's level of performance. For example in educational setting, grades are used as a measure of student performance. In workplace, test can be used to rate worker productivity for comparison either to some standard of performance or to the performance of co workers.

2. **Placement**

Placement involves the evaluation of people so that they can be matched with the appropriate services or environments. Military recruits are placed in training programs based on test performance, business and industry use test to decide where to place new employees. Placement emphasis on finding the right kind of job for the individual.

3. **Selection**

Tests are used for selection of a group of people from a larger pool of applicants and candidates. Private schools, colleges and professional schools use performance on standardized tests as one criterion for admission to their academic programs. Business and industry also use a set of scores as one measure of applicant's suitability for a particular job. Selection emphasis on finding a person with right qualifications.

4. **Competency and Proficiency**

Test can be used to indicate whether or not if the individual's performance meets a preselected criterion. Many school systems require students to pass competency exams prior to receiving a high school diploma. These tests certify that the students have met the minimal requirements necessary for graduations.

5. **Diagnosis**

Diagnosis is one of the more familiar areas of test use. In diagnosis, tests are used to determine the nature and typicality of the individual's underlying characteristics. Clinicians use tests to identify areas of pathology or adjustment problems and to plan treatment approaches.

6. **Outcome evaluations**

The preceding categories all involve in making decisions about individuals. Tests can also be used to make decisions by evaluating an outcome, such as the value of the program, a product or a course of action. Standardized tests can be used to compare the effectiveness of alternative teaching techniques, to determine the effects of drugs or to access the efficacy of different types of therapies. Tests can be used for outcome evaluation both in course of basic science research and in the process of deciding which course of action to choose in applied settings. For example, sets of different tests can be compared to determine which is the best predictor of job performance and thus the best to use for screening job applicants.

It's easier to get information from tests than by interview. Most people won't talk about this, but, believe it or not, many psychologists are rather inept at dealing with people, and so it's a great relief to them to be able to administer a test rather than conduct a competent interview. The information from tests is more scientifically consistent than the information from a interview. If a psychologist is simply trying to arrive at a diagnosis to help determine the course of psychotherapy, an interview is just fine. But when decisions have to be made about legal matters, disability issues, and so on, then the standardized information from tests allows one person to be directly compared with others, and it makes things more fair. It's harder to get away with lying on a test than in an interview. Many tests have multiple "alarms" that go off when a test taker tries to lie. And some tests, such as the Rorschach (the "inkblot test") don't even give a clue as to what preferred, or healthy, responses might be, so it's pretty much impossible to make yourself "look good" by fabricating deceptive answers to a test like this.

Categories of Psychological Tests

1. **Specific Task Performance Tests:**

Tests in which subject performs some specific tasks, such as writing an essay, answering multiple choice questions, or mentally rotating images presented on the computer screen.

Writing an essay, answering multiple-choice item, such as SAT, GRE, ACT are examples of this category of tests. Cronbach (1970) refers to this type of test as a "Test of maximal performance". The defining feature of a performance test relates to the intent or state of mind of the person taking the test. It is assumed that the examinee knows what he or she should do in response to the questions or tasks that make up the performance test and the person being tested exerts maximum effort to succeed.

Performance tests are designed to uncover what an individual can do, given the specific test conditions.

2. **Behavior observations:**

These are tests which involve observations of the subject's behavior within a particular context. Examiner might observe subject having a conversation or some other social interaction. For example Companies recruit observers to pose as salespeople to observe employee's behaviors. Subject's may be unaware they are being tested.

3. **Self-Report Measures:**

The final category of test includes a variety of measures that ask subject to report or describes his or her feelings, attitudes, beliefs, or interests. Many personality inventories such as the MMPI and the 16PF measures are based on self-report. Clinicians include self-report measures as part of their initial examinations of presenting clients.

Self-Report measures are frequently subject to self-censorship. People know their responses are being measured and wish to be seen in a favorable light. This is known as self-serving bias. Items are frequently included to measure the extent to which people provide socially desirable responses. A self-serving bias occurs when people are more likely to claim responsibility for successes than failures. It may also manifest itself as a tendency for people to evaluate ambiguous information in a way beneficial to their interests.

Types of Tests

Tests can be broadly grouped up into two camps: group tests versus individual tests.Group test are largely paper pencil measure suitable to the testing of large group of persons at the same time. Individual tests are instruments that by their design and purpose must be administered one on one.

For convenience we will sort test into eight categories:

Intelligence tests attempt to measure your intelligence—that is, your basic ability to understand the world around you, assimilate its functioning, and apply this knowledge to enhance the quality of your life. Or, as Alfred Whitehead said about intelligence, "it enables the individual to profit by error without being slaughtered by it." Intelligence, therefore, is a measure of a potential, not a measure of what you've learned (as in an achievement test), and so it is supposed to be independent of culture. The challenge is to design a test that can actually be culture-free; most intelligence tests fail in this area to some extent for one reason or another. Tests such as the Stanford-Binet Intelligence Scale and Wechsler Adult Intelligence Scale are examples of intelligence tests.

Aptitude test measure the capability for a relatively specific task or type of skill. It is a test used to predict future performance in a given activity, "intended to predict success in some occupation or training course" (Cronbach, 1984). Scholastic aptitude test is an example of this type of test.

Achievement test measures a person's degree of learning, success or accomplishment in a subject or task. IQ (or cognitive) tests and achievement tests are the most common norm-referenced tests. In either of these types of tests, a series of tasks is presented to the person being evaluated, and the person's responses are graded according to carefully prescribed guidelines. After the test is completed, the results can be compiled and compared to the responses of a norm group, usually comprised of people at the same age or grade level as the person being evaluated.

Academic achievement tests (e.g., WIAT, WRAT) are designed to be administered to either an individual (by a trained evaluator) or to a group of people (paper and pencil tests). The individually-administered tests tend to be more comprehensive, more reliable, more valid and generally to have better psychometric characteristics than group-administered tests. However, individually-administered tests are more expensive to administer because of the need for a trained administrator (psychologist, school psychologist, or psychometrician) and because of the limitation of working with just one client at a time.

Achievement and aptitude tests are usually seen in educational or employment settings, and they attempt to measure either how much you know about a certain topic (i.e., your achieved knowledge), such as mathematics or spelling, or how much of a capacity you have (i.e., your aptitude) to master material in a particular area, such as mechanical relationships.

Creativity tests assess a subject's ability to produce new ideas, insight or artistic creations that are accepted as being of social, aesthetic or scientific value. Thus measures of creativity emphasize novelty and originality in the solutions of fuzzy problems or the productions of artistic works.

Personality tests attempt to measure your basic personality style and are most used in research or forensic settings to help with clinical diagnoses

Two of the most well-known personality tests are

1. The Minnesota Multiphasic Personality Inventory (MMPI), or the revised MMPI-2, composed of several hundred "yes or no" questions, and
2. The Rorschach (the "inkblot test"), composed of several cards of inkblots—you simply give a description of the images and feelings you experience in looking at the blots.

Psychological measures of personality are often described as either objective tests or projective tests.

Objective tests (Rating scale)

Objective tests have a restricted response format, such as allowing for true or false answers or rating using an ordinal scale. Prominent examples of objective personality tests include the Minnesota Multiphasic Personality Inventory, Millon Clinical Multiaxial Inventory-III (Millon, 1994), Child Behavior Checklist (Achenbach & Rescorla, 2001), and the Beck Depression Inventory (Beck & Steer, 1996). Objective personality tests can be designed for use in business for potential employees, such as the NEO-PI, the 16PF, and the Occupational Personality questionnaire, all of which are based on the Big Five taxonomy. The Big Five, or Five Factor Model of normal personality has gained acceptance since the early 1990s when some influential meta-analyses (e.g., Barrick & Mount 1991) found consistent relationships between the Big Five factors.

Table 1.1 Example of personality test items

(a) Adjective Check List
Chose those words which describe you best
( ) relaxed ( ) thoughtful ( ) cheerful ( ) assertive ( ) curious ( ) even-tempered
(b) A True-False Inventory
Circle true or false as each statement applies to you
T F I like sport magazines. T F Most people would lie to get a job. T F I like big parties.

Projective tests (Free response measures)

Projective tests allow for a much freer type of response. An example of this would be the Rorschach test, in which a person states what each of ten ink blots might be. The terms "objective test" and "projective test" have recently come under criticism in the Journal of Personality Assessment. The more descriptive "rating scale or self-report measures" and "free response measures" are suggested, rather than the terms "objective tests" and "projective tests," respectively. There remains some controversy regarding the utility and validity of projective testing which is based on Freud's concept of projecting one's own personality attributes onto a neutral stimulus. However, many practitioners continue to rely on projective testing, and some testing experts (e.g., Cohen, Anastasi) suggest that these measures can be useful in developing therapeutic rapport.

Other projective tests include the House-Tree-Person Test, Robert's Apperception Test, and the Attachment Projective.

Interest inventories measures an individual preference for certain activities or topics and there by help determine occupational preference.

Behavioral procedures objectively describe and count the frequency of behavior, identifying the antecedents and consequences of behavior. Behavioral procedures tend to be highly pragmatic in that they are usually interwoven with treatment approaches.

The Parent-Child Interaction Assessment-II (PCIA; Holigrocki, Kaminski & Frieswyk, 1999) is an example of a direct observation procedure that is used with school-age children and parents. The parents and children are video recorded playing at a make-believe zoo. The Parent-Child Early Relational Assessment (Clark, 1999) is used to study parents and young children and involves a feeding and a puzzle task. The MacArthur Story Stem Battery(MSSB; Bretherton et al., 1990) is used to elicit narratives from children. The Dyadic Parent-Child Interaction Coding System-II(Eyberg, 1981) tracks the extent to which children follow the commands of parents and vice versa and is well suited to the study of children with Oppositional Defiant Disorders and their parents.

Neuropsychological tests measure cognitive, sensory, perceptual and motor performance to determine the extent, locus and behavioral consequences of brain damage. Neuropsychological tests attempt to measure deficits in cognitive functioning (i.e., your ability to think, speak, reason, etc.) that may result from some sort of brain damage, such as a stroke or a brain injury. They usually involve the systematic administration of clearly defined procedures in a formal environment. Neuropsychological tests are typically administered to a single person working with an examiner in a quiet office environment, free from distractions. As such, it can be argued that neuropsychological tests at times offer an estimate of a person's peak level of cognitive performance. Neuropsychological tests are a core component of the process of conducting neuropsychological assessment. One popular test battery is the Halstead-Reitan Test Battery.

Psychological Testing versus Psychological Assessment

Assessment is more than testing. Psychological testing (e.g., an intelligence test, personality test, or mental health test) occurs as part of the process of psychological assessment. Professional psychological assessment usually also includes:

- Interview
- Demographic information
- Medical information
- Personal history
- Observations by others

Thus, the results of a psychological test are rarely used on their own.

The following definitions should help to clarify the difference between assessment and testing in psychology.

Definition of Psychological Testing

"An objective and standardized measure of a sample of behavior"

(Anastasi, 1990)

Definition of Psychological Assessment

"An extremely complex process of solving problems (answering questions) in which psychological tests are often used as one of the methods of collecting relevant data"

(Maloney & Ward, 1976).

Control in the use of psychological tests

Test developers, publishers and psychological examiners generally release psychological tests only to qualified persons who have a legitimate need to study and use these materials. There are three reasons why access to psychological tests is restricted:

1. In the hands of an unqualified person, psychological tests can cause harm.
2. The selection process is rendered invalid for persons who preview test questions.
3. Leakage of item content to the general public completely destroys the efficacy of a test.

Sources of information about tests

Information of psychological testing is available from five sources:

- Reference books
- Publisher's catalogue
- Journal
- Databases
- Test manuals

The best single reference source for information on mainstream test is the Mental Measurement Yearbook (MMY) published by the Buros Institute for Mental Measurement at the University of Nebraska. In sixteen volumes to date, the yearbooks provide descriptive information on tests published in English-speaking countries. References to the tests and reviews are included as well as information on availability, scoring and validity. The MMY's also review major books on tests and testing and provide a bibliography of periodicals on testing. The yearbooks are indexed by test title, personal name and subject area. Volumes 9 and beyond also contain a score index.

Since publication of the 9th MMY, supplements to the Mental Measurement Yearbooks are published in an attempt to provide up-to-date information on tests that are new or which have been significantly revised.

Another way to learn about the test is to request test catalogues from the major test publishers. Many psychological journals publish articles on the reliability and validity of better known tests. The best way to locate studies on the specific tests is through PsychINFO, a computerized database of abstracts from dozens of psychology relevant journals, some going back to 1887 in some cases.

Finally, an important and often overlooked source of information about any specific tests is its manual. A good manual contains essential information about norms, standardization, administration, reliability and validity.

Ethical Issues in Testing

Psychological tests have evolved substantially over time, and thus, guidelines in the use of tests need to be reviewed. As well, many legislative changes have occurred and are continuing to emerge that also affect psychologists. Issues regarding psychological tests have changed over the past few years. The given guidelines are designed to assist practicing clinicians negotiate the issues involving use of such tests, and to have a resource to guide them in effective and ethical practice. However, these are guidelines, and are not a substitute for thorough training in testing, and following proper administration principles.

The range of professionals using instruments has also changed and widened, and we realize that we no longer have exclusive domain over many of these tools. Therefore, part of the responsibilities of Psychologists now must increasingly emphasize education of clients and colleagues regarding such instruments, and the need for security.

Additionally, the use of such tests by unskilled clinicians is a significant concern. Educating the public and colleagues regarding core competencies necessary to effectively use tests is also more important now than it ever was. Nonetheless,Psychologists recognize that test instruments can be a useful resource, and that the results of an assessment can have substantial impact on clients. Thus, psychologists aspire to uphold the highest standards of accuracy and fairness when administering psychological test instruments.

The ethics code on competence states that psychologist "provide only those services and use only those techniques for which they are qualified by education, training or experience" (APA,1992, p.1599).

Ethical issues focus on the following points

Informed Consent: Psychologists will obtain the informed consent of test takers before administering tests. Informed Consent requires the test taker to receive full information concerning the purpose of the testing, the persons who may receive the test scores and the use to which the test scores or resulting report may be implemented. Testing of minors requires informed consent of parents/guardians.

Protection of Test Takers: Psychologists are knowledgeable about the legal requirements and protections for test takers that are relevant to the type of test being administered, the setting in which the test is administered and the specific purpose of the test result.

Competency: Psychological assessments are conducted by psychologists with appropriate qualifications, or by properly trained assistants under appropriate supervision. The educational qualifications and standards established by the College of Alberta Psychologists must be met, and as well, the qualifications specified in the test's manual must also have been met. Psychologists will conduct assessments only within their areas of training and experience, or work under the close supervision of professionals with appropriate training and experience. Psychologists are fully responsible for the contents of any reports they sign, including reports written by others under their supervision.

Test Selection: Psychologists should select tests that are appropriate for the intended purpose and intended test takers. Standardized tests have acceptable statistical properties and are supported by research for the intended use.

Alternate Use: Psychologists should avoid using tests for purposes other than those recommended by the test developer unless there is good evidence to support the alternative intended use or interpretation.

Administration: Psychologists should administer and score tests correctly and fairly following established procedures for administering and scoring tests in a standardized manner. Modifications to standardized test content or procedures should be made only on the basis of carefully considered professional judgment. The rationale and potential impact of the modifications on the validity of the scores should be noted in the results.

Interpretation of Scores: Psychologists should avoid using a single test score as the sole determinant of decisions about test takers and interpret test scores in conjunction with other information about individuals. Psychologists will use tests developed for screening purposes only for identifying test takers who may need further evaluation. Results of screening tests alone should not be used to make any decision about a person, unless adequate reliability and validity for those other uses can be demonstrated.

Psychologists must not solely rely on computer-interpreted test results unless they have information on, the principles, on which the computer interpretations were derived, the validity of the interpretations for the intended applications, and the samples on which they were based. The psychologist also has the responsibility to evaluate the computer based interpretation of test performance in light of other evidence. Simple submission of generic computer generated results as the assessment report is not acceptable.

Derived scores such as standard scores, percentiles or age-equivalents should only be disclosed in the context of an interpretive report containing appropriate cautions about the limitations of the reliability and validity of the scores.

Confidentiality and the Duty to Warn: Assessment results are confidential and shared only with those with a legitimate, professional interest. Test results identified by name of individual test takers should not be released to any person or institution without the informed consent of the test taker unless otherwise required by law. Test results used for research purposes should not individually identify the clients who participated in the research.

Psychologists also have the **duty to warn** that stems from the 1976 decision in the Tarasoff case (Wrightsman, Neitzel, Fortune & Greene, 2002). Tanya Tarasoff was a young college student in California who was murdered by Prosenjit Poddar, a student from India. What makes the case relevant to the practice of psychology is that Poddar had made death threats regarding Tarasoff to his campus therapist. Although the therapist had warned the police that Poddar had made death threats, he did not warn Tarasoff. Two months later, Poddar stabbed Tarasoff to death at her home.

Reporting of Results: Psychologists should communicate test results in a timely fashion and in a manner that is easily understood and avoids misunderstanding.

Use of Results: Psychologists have an obligation to make all reasonable efforts to ensure that results of testing are used appropriately by those to whom they report.

Protection of privacy: The right to privacy is defined as the right to decide for oneself how much one will share with others one's thoughts, feelings and facts about one's personal life; this right is further characterized as " essential to ensure freedom and self-determination"

Test Security: Psychologists will do everything that is within their power to protect the security of standardized tests, including respecting copyrights and eliminating opportunities for test takers to obtain information, protocols, or scores by inappropriate means.

Responsibility of test publishers: Test publishers recognize the broad responsibility that only qualified users should be able to purchase their products.

Outdated testing materials will be disposed of in a secure manner. Client files should be securely stored. Supervisors of provisional psychologists are also responsible to ensure the storage and secure maintenance of the files of their supervisee.

Psychologists will use the most current edition of the test and norms, unless there is compelling rationale to use a previous edition.

List of Standards For Educational and Psychological Testing

Part I: Test Construction, Evaluation, and Documentation

1. Validity
2. Reliability, Errors of Measurement and test score information function
3. Test Development and Revision
4. Scales, Norms, standards and Score Comparability
5. Test Administration, Scoring, and Reporting
6. Supporting Documentation for Tests

Part II: Fairness in Testing

7. Fairness in Testing and Test Use
8. The Rights and Responsibilities of Test Takers
9. Testing Individuals of Diverse Linguistic Backgrounds
10. Testing Individuals with Disabilities

Part III: Testing Applications

11. The Responsibilities of Test Users
12. Psychological Testing and Assessment
13. Educational Testing and Assessment
14. Testing in Employment, licensure and certification
15. Testing in Program Evaluation and Public Policy

Reference

Anastassi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York: Prentice-Hall Inc.

Berg, F.L. (1995). *Psychological testing; Design, analysis and use* (ed.) Singapore: A Simon and Schuster company.

Gregory, R.J. (2006). *Psychological testing; History, Principles and applications* (4th ed) India; Dorling Kindersely Pvt, Ltd

Kaplan,R.M and Saccuzzo, D.P. (1982).*Psychological testing; Principles, Application and Issues* (ed). USA; Brooks/Cole Publishing Company.

http://en.wikipedia.org/wiki/Self-serving_bias

http://en.wikipedia.org/wiki/Psychological_testing

http://wilderdom.com/personality/L3-1TestingVsAssessment.html

Lesson 02

HISTORY OF PSYCHOLOGICAL TESTING

Learning objectives

After completing this lesson, you would be able to do the following:

- Identify the major development in the history of psychological testing.

It is common to think of testing as both a recent and American development. Indeed, most of the major developments in testing have occurred in this century and a good number of them have taken place in United States. The origins of testing are neither recent nor American. Historians had obtained evidence that the ancient Chinese had a relatively sophisticated service testing program more than 4000 years ago (DuBois, 1966, 1970). **Chinese** used **competitive examinations** to select mandarins for civil service. Over the centuries they developed an elaborate system for checks and controls to eliminate possible bias in their testing—procedures that in many ways resembled the best of modern practice. For example examinees were isolated to prevent possible cheating, compositions were copied by the trained scribes to eliminate the chance that differences in penmanship may effects scores, and each examinations was evaluated by a pair of graders, differences being resolved by the third judge. In a number of ways, Chinese practice served as a model for developing for civil service examinations in Western Europe and America during the 1800s.

In 1832, east India copies their system. In 1859, **Charles Darwin** published the "origin of species". He also gave the concept of individual differences and survival of the fittest. Perhaps the most basic concept underlying psychological testing that pertains to the concept of individual differences. No two snow flakes are identical and no two fingers are same. Similarly no two people are alike in ability and typical behavior. In his publication of origin of species, he discussed that higher form of life evolved in this planet partially because of differences among individual forms of life within a species or type of animals. Briefly, given that the individual members of a species differ, some will posses characteristics that are more adaptive or more successful than those possessed by others. Darwin also believed that those with the best or most adaptive characteristics will survive at the expense of those who are less fit, and the survivors than pass on their characteristics on to the next generation. Through this process, he claimed, life has evolved to its presently complex and incredibility intelligent levels.

Sir Francis Galton set out to show that some human possessed characteristics that made them more fit than other humans, a theory he articulated in his book "heredity Genius" published in 1869. Galton (1883) subsequently began a series of experimental studies to document the validity of his position. He concentrated on demonstrating that individual differences exist in human sensory and motor functioning such as reaction time, visual acuity and physical strength.

James McKeen Cattell's doctoral dissertation was based on Galton's work concerning individual differences in reaction time. In continuing Galton's explorations of human individual differences and introducing the term mental test, Cattell perpetrated and stimulated the forces that ultimately led to the development of modern psychological tests.

A lot of people worked on the psychophysical measurement, some important names are **Fechner, Wundt and Weber.**

Formal measurement procedures began to appear in Western educational practice during the 19th century. For several centuries, secondary school and universities had been using essay and oral examinations to evaluate student achievement, but in 1897, **Joseph M. Rice** used some of the first union written examinations to test spelling achievement of students in the public schools of Boston. Rice wanted the schools to make room in the curriculum for teaching science and argues that some of the time spent on the spelling drill could be used for that purpose. He demonstrated that amount of time devoted to spellings drills was not related to achievement in spelling and concluded that this time could be reduced, thus making time to teach science. His study represents one of the first times tests were used to evaluate curriculum and make a curriculum decision.

Through out the later half of the 19th century, pioneering work in the infant science of psychology involved developing new ways to measure human behavior and experience. Many of the measurement advances of the time came from laboratory studies such as those of **Hermann Ebbinghaus**, who in 1896 introduced the completion test (fill in the blanks) as a way to measure mental fatigue in students. Other important advances such as the development of correlation coefficient by **Sir Francis Galton and Karl Pearson** were made in the service of research on the distribution and causes of human differences. The late 1800s have been characterized by the DuBois (1970) as the **laboratory period** in the history of psychological measurement.

Increasing interest in human differences in the second half of the 19th century can be traced back to the need to make decisions in three contexts.

- First, there was a growing demand for objectivity and accountability in assessing student's performance in the public school, which resulted from the enactment of mandatory school attendance laws. These laws brought into school for the first time a large number of students who were of middle or lower socioeconomic background and unfamiliar with formal education. Many of these children performed poorly and were considered by many educators of the time to be "**feebleminded"** and unable to learn. The development of accurate measurement methods and instruments was seen as a way to differentiate children with true mental handicaps from those who suffered from disadvantaged background.
- Second, the medical community was in the process of refining its idea about abnormal behavior. Behavioral measurements were seen as a way to classify and diagnose patients.
- Third, all sort of government agencies begin to replace patronage systems for granting employment with examination of the prospective employees' abilities. Tests began to be used as a basis of employee selection.

Not until the first years of the 20th century did well developed prototypes of modern educational and psychological measurements begin to appear. Although it is difficult to identify a single event that started it all, 1905 publication of the Binet-Simon scales of mental ability is often taken as the beginning of the modern era in behavioral measurement. These scales originally published in French but soon translated into English, represented the first successful attempt to measure complex mental processes with a standard set of tasks with graded complexity. The Binet Simon scales were designed to aid educators in identifying students whose mental ability was insufficient for them to benefit from standard public education. On the basis of mental measurement, a decision was then made whether to place these students in special classes. Subsequent editions of scales, published in 1908 and 1911, contained tasks that spanned the full range of abilities for school age children and could be used to identify students at either extremes of the ability continuum.

At the same time Binet and Simon were developing the first measures of intelligence, **E.L Thorndike and his students at Teachers college of Columbia University** were talking problems related to measuring school abilities. Their work ranged from theoretical developments on the nature of the measurement process to the creation of the scales to assess class room learning of reading and arithmetic and level of skill development in task such as handwriting.

It is convenient to divide history of mental testing in the 20th century into five periods:

- The early period
- The boom period
- The first period of criticism
- The battery period
- The period of criticism

The Early Period

The early period, which comprises the years before American entry into World War I, was a period of tentative exploration and theory development. The Binet Simon scales were revised twice by Binet and was brought to the United States by several pioneers in measurement. The most influential of these were the Lewis Terman of Stanford University.

In 1916 **Terman** published the first version of the test that in its fourth edition is still one of the standards by which measures of intelligence are judged: **the Standford Binet Simon scale**. Working with Terman at this time, Arthur Otis began to explore the possibility of testing of mental ability of children and adults in groups. In Australia, **S.D. Porteus** prepared the maze test of intelligence for use with people with hearing or language handicaps.

During the time Binet was developing the first modern test of intelligence, **Charles Spearman** published two important theories relating to the measurement of human abilities.

- The first was a statistical theory that proposed to describe and account for the inconsistency in measurements of human behavior.
- The second theory claimed to account for the fact that different measures of cognitive ability showed substantial consistency in the ways that they ranked people.

Spearman second theory states that there is a single dimension of ability underlying most human performance, played a major role in determining the direction that measures of ability took for many years. Spearman proposed that the consistency of people's performance on different ability measures was the result of the level of intelligence that they possessed.

The Boom Period

The American involvement in the World War I brought a need to expand the army very quickly. For the first time, the new science of psychology was called on to play a part in a military situation. This event started a 15-year boom period during which there were many advances and innovations in the field of testing and measurement. As part of the war effort, a group of psychologist expanded Otis' work to develop and implement the first large scale group testing of ability with the **Army Alpha** (a verbal test) and the **Army Beta** (a test using mazes and puzzles similar to Porteus' that required no spoken or written language). The Army Alpha was the first widely distributed test to use the multiple-choice item form.

The first objective measure of personality, the **Woodworth Personal Data Sheet**, was also developed for the army to help identify those emotionally unfit for the military service. The Army Alpha and Beta were used to select officer's trainees and to remove the intellectually handicapped from the military service.

In the 12 years following the war, the variety of behaviors that were subjected to measurement continued to expand rapidly. **E.K Strong and his students** began to measure vocational interests to help college students choose majors and careers consistent with their interest. Measurements of personality and ability were developed and refined, and the use of standardized tests for educational decisions became more widespread.

In 1929, **L.L. Thurstone** proposed ways to scale and measure attitudes and values. Many people considered it only a matter of time before accurate measurement and prediction of all types of human behavior would be achieved.

The First Period of Criticism

The 1930s saw a crash not only in the stock market but also in the expectations for mental measurements. This time cover the period of criticism and consolidation. To be sure, new tests were published, most notably the original Kuder scales of vocational interest, the Minnesota Multiphasic Personality Inventory, and the first serious competitor for the Standford Binet, Wechsler-Bellevue Intelligence Scale.

Major advances were also made in the mathematical theory underlying tests, particularly L.L. Thurstone's refinements of a statistical procedure known as factor analysis. However, it was becoming clear that the problems of measuring human behavior had not been solved and were much more difficult than they had appeared to be in the heady years of the 1920s.

The Battery Period

In the 1940s, psychological measurement was once again called on for the use in the military service. As part of the war effort, batteries of the test were developed that measured several different abilities. Based on the theory developed by Thurstone and others that there were several distinct types or dimensions of abilities, these test batteries were used to place military recruits in the positions for which they were best suited. The success of this approach in reducing failure rates in various military programs led the measurement field into the period of emphasis on test batteries and factor analysis. For 25 years, until about 1965, efforts were directed towards analyzing the dimensions of human behavior by developing an increasing variety of tests of ability and personality.

During the 1950s, educational and psychological testing grew to become big business. The use of nationally normed, commercially prepared tests to assess student progress became a common feature of school life. Business, industry and the civil service system made increasing use of measurements of attitude and personality, as well as ability, in hiring and promotion decisions. Patients in mental institutions were routinely assessed through a variety of measures of personality and adjustment. The widespread use and misuse of tests brought about a wave of protests.

The Second Period of Criticism

The beginning of the second period of criticism was signaled in 1965 by a series of congressional hearing on testing on invasion of privacy. Since that time there has been a continuing debate over the use of ability and personality testing in public education and employment. Ethical implications gained concern here. A major concern has been the possible use of test to discriminate, intentionally or otherwise, against women and/or members of minority groups in education and employment. As a result of this concern the tests themselves have been very carefully scrutinized for biased content, certain types of testing practices have been eliminated or changed, and much more attention is given to the right of individual. Computer based testing became popular.

The testing industry has responded vigorously to making tests to fair to all who take them, but this has not been sufficient to forestall both legislation and administrative and court decisions restricting the use of tests. This situation is unfortunate because it deprives decisions makers of some of the test information on which to base their actions.

Early Testing In the United States

Early Uses and Abuses of Tests in the United States

In the 1906, **Henry H. Goddard** was hired by the Vineland training School in New Jersey to do research on classification and education of feebleminded children. He soon realized that a diagnostic instrument would be required and was therefore pleased to read of the 1908 Binet-Simon scale. He quickly set about translating the scale, making minor changes so that it would be applicable to American children (Goddard, 1910a).

Goddard (1910) tested 378 residents of the Vineland faculty and categorized them by diagnosis and mental age. He classified 73 residents as *idiots* because their mental age was 2 years or lower; 205 residents were termed *imbeciles* with mental age of 3 to 7; and 100 residents were deemed *feebleminded* with mental age of 8 to 12. It is instructive to note that originally neutral and descriptive terms for portraying levels of mental retardation—*idiot, imbecile and feebleminded*—have made their way into the everyday lexicon of pejorative labels. In fact, Goddard made his own contribution by coining the diagnostic term moron (from the Greek *moromia*, meaning foolish).

Goddard (1911) also tested 1,547 normal children with his translation of the Binet-Simon Scales. He considered children whose mental age was four or more years behind their chronological age to be feebleminded—these constituted 3 percent of his sample. Considering that all of these children were found outside of institutions for the retarded, 3 percent is rather an alarming rate of mental deficiency. Goddard (1911) was of the opinion that these children should be segregated so that they would be prevented from "contaminating society". These early studies piqued Goddard curiosity about "feebleminded" citizenry and the societal burdens they imposed. He also gained a reputation as one of the leading experts on the used of intelligence tests to identify persons with impaired intellect. His talents were soon in heavy demands.

The Binet-Simon and Immigration

In 1910, Goddard was invited to Ellis Island by the commissioner of immigration to help make the examination of immigrants more accurate. A dark and foreboding folklore had grown up around mental deficiency and immigration in the early 1900s:

- It was believed that the feebleminded were degenerate being responsible for many if not most social problems; that they reproduced at an alarming rate and menaced the nation's overall biological fitness; and that their numbers were being incremented by the undesirable "new" immigrants from southern and eastern European countries who had largely supplemented the "old" immigrants from northern and western Europe. (Gelb, 1986)

Initially, Goddard was unconcerned about the supposed threat of feeblemindedness posed by the immigrants. He wrote that adequate statistics did not exist and that the prevalent options about the undue percentages of mentally defective immigrant were grossly overestimated (Goddard, 1912). However with repeated visits, Goddard became convinced that the rate of feeblemindedness were much higher than estimated by the physician who staffed the immigration service. Within a year, he reversed his opinion entirely and called for congressional funding so that Ellis Island could be staffed with experts trained in the use of intelligence test. In the following decade, Goddard became an apostle for the use of intelligence test to identify feeble minded immigrants. Although he wrote that the rates of mentally deficient immigrants were "alarming" he did not join the popular call for immigration restriction (Gelb, 1986).

The fact is that Goddard was one of the most influential psychologists of the early 1900s. Any thoughtful person must therefore wonder why so many contemporary authors have ignored or slighted the person who first translated and applied Binet's test in the United States. We will attempt an answer here based in part of Goddard's original writing but also relying on Gould's (1981) critique of Goddard's voluminous writing on mental deficiency and intelligence testing.

Perhaps Goddard has been ignored in the text books because he was a strict hereditarian who conceived of intelligence in simple minded Mendelian terms. No doubt his call for colonization of morons so as to restrict their breeding has won him contemporary disfavor as well. However, the most likely reason that modern authors have ignored Goddard is that he exemplified a large number of early, prominent psychologists who engaged in blatant misuse of intelligence testing. In his efforts to demonstrate that high rates of immigrants with mental retardation were entering the United states each day, Goddard sent his assistants to Ellis Island to administer his English translation of Binet-Simon test to newly arrived immigrants. The tests were administered through a translator, not long after the immigrant walked ashore. We can guess that many of the immigrants were confused, frightened and disoriented. Thus, a test devised in French, then retranslated into English was, in turn, retranslated back to Yiddish, Hungarian, Italian or Russian; administered to bewildered farmers and laborers who had just endured an Atlantic crossing; and interpreted according to the original French norms.

What did Goddard find and what did he make of his results? In small samples of immigrants, his assistants found 83 percent of the Jews , 80 percent of the Hungarians, 79 percent of the Italians and 87 percent of the Russians to be feeble minded, this is below age 12 on Binet-Simon Scales. His interpretations of this findings, by turns, skeptically cautious and then provocatively alarmist. In one place he claims that his study "makes no determination of actual percentage, even of these groups, who are feebleminded." Yet, later in his reports he states that his figures would only need to be revised by "a relatively small amount" in order to find actual percentage of feeblemindedness among immigrants groups. Further, he concludes that the intelligence of the average immigrant is low; "perhaps a moron grade" but then goes on to cite environmental deprivations as the primary culprit. Simultaneously, Goddard appears to favor deportation for low IQ immigrants but also provides the humanitarian perspective that we might be able to use "moron laborers" if only we are wise enough to train them properly.

Goddard was a complex scholar who refined and contradicted his own professional opinions on number occasions. One ironic example: after the damage was done and his writing help restrict immigrations, Goddard recanted, concluding that feeblemindedness was not incurable and that the feebleminded did not need to be segregated in institution.

The Inventions of Nonverbal Tests in the Early 1900s

Because of the heavy emphasis of the Binet-Simon scale upon verbal skills, many psychologists realized that this new measuring devise was not entirely appropriate for non-English speaking subject, illiterates and those with speech and hearing impairments.

The earliest of the performance measures was the **Seguin form board**, an upright stand with depression into which ten blocks of varying shapes could be fitted. This had been used by Seguin as a training advice for individuals with mental retardations, but was subsequently developed as a test by Goddard and then standardized by R.H Sylvester (1913). This identical board is still used, with the subject blindfolded, in the Halstead-Reitan neuropsychological test battery.

Knox (1914) devised several performance tests for use with Ellis Island immigrants. His tests absolutely required no verbal responses from subjects. The examiner demonstrated each task non-verbally to ensure that subjects understood the instructions. Included in his tests were a simple wooden puzzle and the same digit-symbol substation test which is now found on most of the Wechsler scales of intelligence.

Pinter and Paterson (1917) invented a 15 part scale of performance tests that used several form boards, puzzles and object assembly tests. The object assembly test—reassembling cut up card board versions of common objects such as the horse –is a mainstay of several contemporary intelligence test. The Kohs block test, or Kohs block design test, is a cognitive test for children or adults with a mental age between 3 and

19. It is mainly used to test persons with language or hearing handicaps but also given to disadvantaged and non-English-speaking children. The child is shown 17 cards with a variety of colored designs and asked to reproduce them using a set of colored blocks. Performance is based not just on the accuracy of the drawings but also on the examiner's observation of the child's behavior during the test, including such factors as attention level, self-criticism, and adaptive behavior (such as self-help, communication, and social skills).

The Kohs Block Design Tests which required the subject to assemble painted objects to resemble a pattern is well known to any modern tester who uses the Wechsler scales.

The Porteus maze test is graded series of mazes for which the subject must avoid dead ends while tracing a path from beginning to the end.

This is a fine instrument that is till available today, but underused.

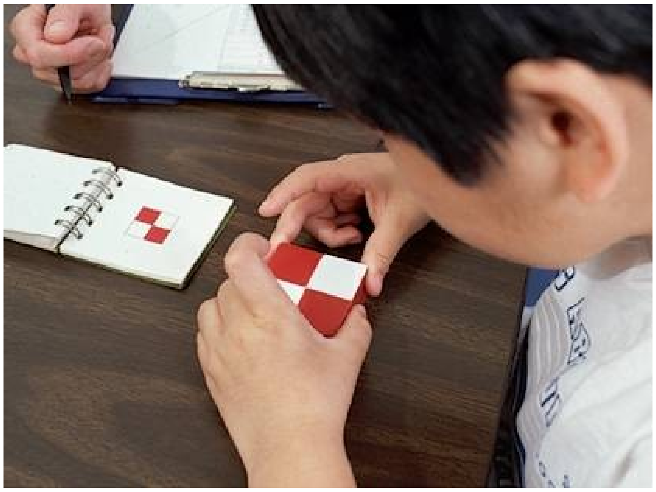


Figure 2.1: Child is shown 17 cards with a variety of colored designs and asked to reproduce them using a set of colored blocks

The Stanford-Binet: The Early Mainstay of IQ

The first useful intelligence test was prepared in 1905 by the French psychologists Alfred Binet and Theodore Simon. The two developed a 30-item scale to ensure that no child could be denied instruction in the Paris school system without formal examination. In 1916 the American psychologist Lewis Terman produced the first Stanford Revision of the Binet-Simon scale to provide comparison standards for Americans from age three to adulthood. The test was further revised in 1937 and 1960, and today the Stanford-Binet remains one of the most widely used intelligence tests. While it was Goddard who first translated the Binet scales in the United States, it was Standford professor Lewis M. Terman who popularized IQ testing with his revision of Binet Scales in the 1916. The number of items was increased to 90 and the new scale was suitable for those with mental retardation, children, and both normal and superior adults. In addition, the Standford Binet had clear and well organized instructions for administrations and scoring. Great care has been taken in securing a representative sample of subjects for use in the standardization of test use.

It is worth mentioning here that Wechsler Scales became a quite popular alternative to the Standford Binet mainly because they just provided more than just an IQ score. In addition to the Full scale IQ, the Wechsler scales provided ten to twelve subsets scores and a verbal and performance IQ. By contrast, the earlier version of Standford Binet intelligence test supplied only a single overall summary score, the global IQ.

The group tests and the classification of the WWI army recruits

The Alpha and Beta Tests of World War I

The first mental tests designed to be used for mass, group testing were developed by psychologists for the U.S. Army in 1917-1918. The group tests were modeled after intelligence tests designed for individual use in one-on-one assessment. In developing the mental tests, the psychologists subscribed to the position that one could be quite intelligent, but illiterate or not proficient in the English language. Based on this reasoning, two major tests were developed, the Army Alpha for literate groups, and the Army Beta for illiterates, low literates or non-English speaking (Yerkes, 1921). Both tests were based on the theoretical position that intelligence was an inherited trait, and the assumption was made that native intelligence was being assessed. Each test was made- up of a number of subtests, the contents of which differed depending on whether the test was for literates or illiterates, low literates or non- English speakers.

The Alpha Test

The Alpha test battery included a wide range of tests of knowledge and various cognitive skills. The Alpha test can be reinterpreted not as a test of native intelligence but as a sampling of a wide variety of cognitive abilities by addressing the person's knowledge base by both oral language and written language.

Test 1: **Following Oral Directions**, involves comprehending simple or complex oral language directions and looking at and marking in the appropriate places on the answer sheet. To a large extent, this is a test of the ability to hold information in working memory and to combine earlier instructions with later ones to determine the correct marking responses.

Test item sample

FOLLOWING ORAL DIRECTIONS

Make a cross in the first and also in the third circle

O O O O O

Test 2: **Arithmetical Problems**, requires both the ability to read and comprehend the stated problem and the knowledge of arithmetic to perform the computations called for. Again, working memory is stressed by having to hold more than one phrase in it that is information bearing, then combining the phrases and performing the required computations.

Test item samples

1. How many are 30 men and 7 men?
2. If you save $7 a month for 4 months, how much will you save?
3. If 24 men are divided into squads of 8, how many squads will there be?
4. How many hours will it take a truck to go 66 miles at the rate of 6 miles an hour?
5. If it takes 6 men 3 days to dig a 180-foot drain, how many men are needed to dig it in half a day?
6. A dealer bought some mules for $800. He sold them for $1,000, making $40 on each mule. How many mules were there?

Test3: **Practical Judgment** clearly requires reading and comprehending language. Additionally, however, it requires knowledge of culturally, normative expectations to make the "correct" choice. In terms of the developmental model of literacy, this means that the person's mind would have had to develop in an external context or environment in which the information needed to make the normatively "correct" response would be presented in some form.

Test item samples

Instructions: This is a test of common sense. Three answers are given to each question. Select the best answer to each question.

Question	**Answers**
Cats are useful animals, because	a. they catch mice b. they are gentle c. they are afraid of dogs
Why are pencils more commonly carried than fountain pens? Because	a. they are brightly colored b. they are cheaper c. they are not so heavy
Why is leather used for shoes? Because	a. it is produced in all countries b. it wears well c. it is an animal product
Why judge a man by what he does rather than by what he says? Because	a. what a man does shows what he really is b. it is wrong to tell a lie c. a deaf man cannot hear what is said
If you were asked what you thought a person whom you didn't know,	a. I will go and get acquainted b. I think he is all right c. I don't know

what should you say? him and can't say

Test 4: **Synonyms-Antonyms**, requires specific vocabulary knowledge, in addition to the knowledge of "same" and "opposite."

Test item samples
Instructions: If the two words of a pair mean the same or nearly the same, select same. If they mean the opposite or nearly the opposite, select opposite.
1. wet-dry
2. in-out
3. hill-valley
4. class-group
5. confess-admit

Test 5: **Disarranged Sentences**, requires semantic knowledge about flies as well as grammatical knowledge to rearrange the sentences, and information has to be held in working memory while rearranging the sentences.

Test item samples
Instructions: The words A EATS COW GRASS in that order are mixed up and don't make a sentence; but they would make a sentence if put in the right order: A COW EATS GRASS, and this statement is true. Below are ten mixed-up sentences. Some of them are true and some are false. Think what each would say if the words were straightened out. Then, if what it would say is true, select 'true'; if what is would say is false, select 'false.'
1. lions strong are
2. days there in are week eight a
3. leg flies one have only
4. honey bees flower gather the from
5. and eat good gold silver to are
6. every times makes mistakes person at

Test 6: **Number Series Completion**, emphasizes reasoning with number knowledge in working memory.

Test item samples

Instructions: Look at each row of numbers below and determine the two numbers that should come next.
a. 3 4 5 6 7 8
b. 3 6 9 12 15 18
c. 8 1 6 1 4 1
d. 27 27 23 23 19 19
e. 1 2 4 8 16 32

Test 7: **Analogies**, clearly emphasizes culturally determined, semantic knowledge retrieval from the long term memory knowledge base, and also information processing in working memory to detect similarities among the different knowledge domains addressed by the analogies.

Test item samples
Instructions: In each of the lines below, the first two words are related to each other in some way. What you are to do in each line is to see what the relation is between the first two words, and then select the word that is related in the same way to the third word.

Word Relationship
1. gun-shoots :: knife- a. run b. cuts c. hat d. bird
2. handle-hammer :: knob- a. key b. room c. shut d. door
3. water-drink :: bread- a. cake b. coffee c. eat d. pie
4. hour-minute :: minute- a. man b. week c. second d. short
5. tiger-carnivorous :: horse- a. cow b. pony c. buggy d. herbivorous

Test 8: **Information** is heavily loaded with cultural knowledge requirements. It is a probe of the person's knowledge base to discover the extent to which it includes both very familiar and less familiar declarative knowledge available in the United States' culture.

Test item samples

1. America was discovered by

a. Drake b. Hudson c. Columbus d. Cabot

2. Pinochle is played with

a. rackets b. cards c. pins d. dice

3. The most prominent industry of Detroit is

a. automobiles b. brewing c. flour d. packing

Rather than regarding the Alpha scores as reflecting the results of literacy practices and years of schooling, the test developers considered that the years of schooling completed reflected the results of the native intelligence measured by the Alpha tests (Yerkes, 1921, p.783).

The Beta Test

In determining who should take the Beta test, decisions were made frequently in terms of the number of years of education reported. Generally, those with fewer than four, five, or six years of education were sent to Beta testing. Additionally, men who were non-English speakers, or very poor in speaking English were sent for Beta testing. In some cases, men who tried the Alpha tests but were subsequently judged to be poor readers were readministered the Beta tests. The procedures were not uniform across the testing locations.

Though the attempt was to use the Beta test as an intelligence test comparable to the Alpha but freed of influences of literacy and the English language, examination of the subtests reveals major differences between the Alpha and Beta tests both in terms of the knowledge called for and the information processing skills involved in processing graphically presented information.

In the subtests of the Beta test, it is clear that literacy as the use of graphics technology for problem solving and reasoning is included in every subtest.

Test 1: **Maze**, requires looking at the graphically represented maze while reasoning about the path to be taken.

Test 2: **Cube Analysis** requires counting cubes in the graphic representation and this combines the use of graphics information with knowledge of the language of arithmetic for counting

Test 3: X-O **Series** requires reading graphic displays in left to right sequences while reasoning in working memory.

Test 4: **Digit Symbol** requires scanning the upper number and graphic symbols, holding them in working memory while scanning the lower numbers and then producing the appropriate mark to match the graphic symbol to the number.

Test 5: **Number Checking**, is similar to Test 4 in requiring scanning and matching of graphic symbols, this time in numeric forms.

Test 6: **Picture Completion**, clearly involves the scanning of graphic displays and the knowledge of the depicted objects to complete the picture.

Test 7: **Geometrical Construction** involves studying in working memory the graphics information on the left and mentally rearranging it to construct the figure on the right.

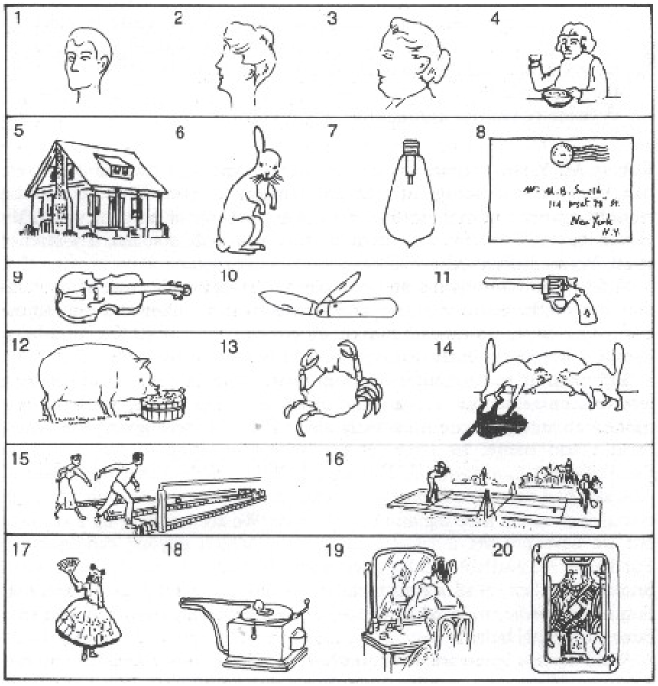


Figure 2.2: Picture completion task

When the Alpha and Beta test total scores (excluding non-English speakers) were correlated with mental age scores on the Stanford-Binet individually administered intelligence test, the resulting coefficients were .81 and .73, respectively. Since the Stanford-Binet is essentially an auding test, in which the administrator speaks of questions and the given information, it is perhaps to be expected that the correlation between the heavily language-laden Alpha and Stanford-Binet tests would be greater than the very low language-based Beta test with the Stanford- Binet. (Yerkes, 1921, p. 782).

Early educational testing

Early testing pioneers such as C.C Brigham used result of individual and group intelligence tests to substantiate ethnic differences in intelligence and thereby justify immigration restriction. Later some of these testing pioneers disavowed their prior view.

Educational testing fell under the purview of the college entrance examination board (CEEB), founded at the turn of the twentieth century. In 1947, the CEEB was replaced by the Educational Testing Service (ETS) which supervised the release of such well known tests as the Scholastic Aptitude Tests and the Graduate Record Exam.

The Development of Aptitude Test

The advent of the multiple aptitude test batteries was made possible with the development of factor analysis by L.L. Thurstone and others. Later, improvement of these test batteries was spurred on the practical need for selecting WWI recruits for highly specialized positions.

Personality Testing and Vocational Testing After WWI

Personality began with Woolworth's Personal Data sheet, a simple yes no checklist of symptoms used to screen WWI recruits for psychoneurosis. Many later inventories, including the popular Minnesota Multiphasic Personality Inventory, borrowed content from the Personal Data Sheet.

The personal data sheet consists of 116 questions that the subject was to answer by underlying yes or no. the questions were exclusively of the face obvious variety and for most part, involved fairly serious symptomatology. Representative items included:

- Do ideas run through your head so that you cannot sleep?
- Were you considered a bad boy?

- Are you bothered by feelings that things are not real?
- Do you have a strong desire to commit suicide?

The next major development was an inventory of neurosis, the Thurstone Personality Schedule (Thurstone and Thurstone, 1930). From the Thurstone sprang the Bernreuter Personality Inventory (Bernreuter, 1931). It was a little more refined than its Thurstone predecessor, measuring four personality dimensions:

- Neurotic tendency
- Self-sufficiency
- Introversion-extroversion
- Dominance-submission

A major innovation in the test construction was that a single test item could contribute to more than one scale.

The Origins of Projective Testing

The projective test began with the **word association technique** pioneered by **Francis Galton** and brought to fruition by **C.G. Jung** in 1910.

As interest in the newly emerging field of psychoanalysis grew in the 1930s, two important projective techniques introduced systematic ways to study unconscious motivation: the Rorschach or inkblot test—developed by the Swiss psychiatrist Hermann Rorschach—using a series of inkblots on cards, and a story-telling procedure called the Thematic Apperception Test—developed by the American psychologists Henry A. Murray and C. D. Morgan. Both of these tests are frequently included in contemporary personality assessment.

The Thematic Apperception Test, or TAT, is a projective measure intended to evaluate a person's patterns of thought, attitudes, observational capacity, and emotional responses to ambiguous test materials. In the case of the TAT, the ambiguous materials consist of a set of cards that portray human figures in a variety of settings and situations. The subject is asked to tell the examiner a story about each card that includes the following elements: the event shown in the picture; what has led up to it; what the characters in the picture are feeling and thinking; and the outcome of the event.

Because the TAT is an example of a projective instrument— that is, it asks the subject to project his or her habitual patterns of thought and emotional responses onto the pictures on the cards— many psychologists prefer not to call it a "test," because it implies that there are "right" and "wrong" answers to the questions. They consider the term "technique" to be a more accurate description of the TAT and other projective assessments.

Figure 2.3: In the TAT, the test subject (the boy shown here) examines a set of cards that portray human figures in a variety of settings and situations, and is asked to tell a story about each card. The story includes the event shown in the picture, preceding events, emotions and thoughts of those portrayed, and the outcome of the event shown. The story content and structure are thought to reveal the subject's attitudes, inner conflicts, and views.

Swiss psychiatrist Hermann Rorschach was the developer of the widely-used personality evaluation method known as the Rorschach test. The Rorschach test involves the assessment by a psychiatrist or psychologist of a subject's responses when asked what he or she sees in a series of inkblots. Rorschach believed that this method could determine the amount of introversion and extroversion a person possessed, as well as clues about such characteristics as intelligence, emotional stability, and problem-solving abilities. In addition to general use in psychiatry and psychology, the test has come to be used by a wide-range of groups such as child development specialists, the military, prisons, and employers. Although the test was Rorschach's only contribution to the field of psychiatry, the popularity of the tool has made his name one that is recognized both inside and outside the profession.

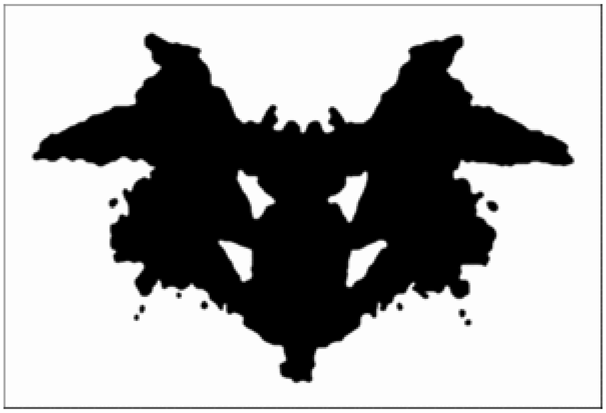


Figure 2.4: Rorschach Ink blot card

During World War II the need for improved methods of personnel selection led to the expansion of large-scale programs involving multiple methods of personality assessment. Following the war, training programs in clinical psychology were systematically supported by U.S. government funding, to ensure availability of mental-health services to returning war veterans. As part of these services, psychological testing flourished, reaching an estimated several million Americans each year. Since the late 1960s increased awareness and criticism from both the public and professional sectors have led to greater efforts to establish legal controls and more explicit safeguards against misuse of testing materials.

The Development of Interest Inventories

The assessment of vocational interest began with Yoakum’s Carnegie Interest Inventory developed in 1919-1920. After several revisions and extensions, this instrument emerged as E.K. Strong vocational Interest Blank. Strong Vocational Interest Blank became one of the most widely used tests of all time (Strong, 1927). Today, everyone from psychologist, counselors, teachers and human resource managers use psychological and educational evaluations. There is scarcely a person over the age of ten who has not taken at least one such test in their lifetime, whether it was an achievement test, an IQ test, a personality evaluation, or a measure of aptitude in a particular field. The key reason for the increase in test use over the last 75 years is ethically correct tests are more reliable and accurate than subjective judgments, which often function as filters when we assess and observe others.

But testing should never be used in a vacuum. As Robert Guion says,

“Testing should not be used as an instrument of decision. It should be used as a flag that either agrees with or contradicts your impression about the person.”

At mind data we agree tests can never replace professional judgment entirely. Rather, they should serve as one source of information to assist in making accurate and fair decisions when hiring and prompting.

References

Gregory, R.J. (2006). *Psychological testing; History, Principles and applications* (4th ed) India; Dorling Kindersely Pvt, Ltd

Kaplan,R.M and Saccuzzo, D.P. (1982).Psychological testing; Principles, Application and Issues (ed). USA; Brooks/Cole Publishing Company.

Robert M. Yerkes (1921). Psychological Examining in the United States Army. Memoirs of the National Academy of Sciences, Vol. XV. Washington, DC: U.S. Government Printing Office.

http://www.minddisorders.com/Py-Z/Thematic-Apperception-Test.htm

http://psychology.jrank.org/pages/363/Kohs-Block-Test.html

Lesson 03

TEST CONSTRUCTION

CHARACTERISTICS OF A GOOD TEST

Although we have discussed the defining features of a test and-, the ways tests are used, we must acknowledge an obvious point: Not all tests—even among those that are published—are good tests. A good test is designed carefully and evaluated empirically to ensure that it generates accurate, useful information. The design phase consists of decisions about test purpose, content, administration, and scoring; the evaluation phase consists of collecting and analyzing data from pilot administrations of the test; these data are then used to identify the psychometric properties of the test.

Design Properties

In terms of its design, we can identify four basic properties of a good test: (1) a clearly defined purpose, (2) a specific and standard content, (3) a standardized administration procedure, and (4) a set of scoring rules. Let's consider each of these.

Property 1. A good test has a clearly defined purpose. To define the purpose of a test, the test developer must answer three questions:

What is the test supposed to measure?

Who will lake the test?

How will the test scores be used?

The first question is one of domain. The domain of a test is the knowledge, skills, or characteristics assessed by the test items. Tests can be designed to measure elements of academic knowledge or skill, personality characteristics, personal attitudes—almost anything you can name. A domain could be a single attribute, such as piano-playing skill, or a set of attributes, such as the central dimensions of personality. Specifying the domain of a test is like outlining the information to be covered in a term paper. When writing a term paper, you begin by selecting a topic and then proceed to decide on a set of specific points or issues to cover. When designing a test, you begin by selecting a domain and then proceed to specify the kinds of knowledge, skills, behaviors, and attitudes that comprise the domain.

The second question is one of audience. A test for adults must necessarily be different from a test for grade-school children, regardless of the test domain. The audience may dictate practical concerns such as whether the questions should be presented orally or in writing or the answers should be ill pictures or in words.

The third question deals with the appropriateness of different types of test items and test scores. Some tests are designed to compare the performance of test takers to each other. Other tests are designed to determine each person's level of performance independently. Some tests measure current level of ability or skill, whereas others are used to predict how well the test taker is likely to perform in the future. Different types of items and scores are used for these different types of comparisons.

Property 2. A good test has a specific and standard content. The content is specific to the domain the test is designed to cover. Questions are selected to cover that domain as comprehensively as possible, taking into account the nature of the test takers and the constraints implied by that audience. The content also is standard, meaning that all test takers are tested on the same attributes or knowledge. This may seem obvious—but there are situations in which examinees may answer different but comparable questions. For example, many tests are available in more than one form. These alternate forms are useful when people must be tested repeatedly or when test takers will be in close quarters and cheating is a concern. Sometimes these alternate forms contain the same questions in different orders; other times, each form contains different questions. In the latter case, the questions on the various forms must require equivalent levels of knowledge or skill. If they do not, your score might differ according to which form you lake.

Property 3. A good test has a set of standard administration procedures. Some tests are self-administered; others require the use of a proctor, who administers the test, or an examiner, who both administers the test and records your responses. In each case, it is critical that all test takers receive the same instructions and materials and have the same amount of time to complete the test. You no doubt have experience with the effort devoted to creating standard conditions during achievement tests or entrance exams such as the SAT or ACT: large black letters in boldface type on the bottom of a page saying, "Stop. Do not go on." Proctors looking at their watches while you eagerly hold your pencil, waiting to hear the words, "You may turn over your test booklet and begin."

Property 4. A good test has a standard scoring procedure. This procedure must be applied the same way to all individuals' who take the test. Objective tests, such as multiple-choice tests, are relatively easy to score in a standard and consistent way. Answers can be coded as right or wrong for knowledge or skill tests or as representing certain attitudes or characteristics for personality tests. It is more difficult to ensure consistency in the scoring of essay tests or projective personality

Psychometric Properties

A good test is one that measures what it is designed to measure in as accurate a way as possible. The measurement characteristics of a test are called psychometric properties. They are determined by analyzing responses to test items during pilot administrations of the test. There are three important psychometric properties of a good test.

Property 1. A good test is reliable. A synonym for reliability is consistency." Just as a reliable person will be consistent in his or her actions and reactions, a reliable test will provide a consistent measure of current knowledge, skills, or characteristics. Without changes in knowledge, skills, or characteristics, an individual taking a reliable test can expect to obtain about the same score on another administration of the test or on another form of the lost. Why is this important? When test scores change, we would like to be able to conclude that the test taker has learned more or has changed somehow. Without a reliable test, we cannot determine what a change in scores really means.

Property 2. A good test is valid. Reliability analysis indicates whether the test provides a consistent measure, but does not tell us what the test measures. The second property, validity, indicates whether the test measures what it was designed to measure.

Property 3. A good test contains items with good item statistics. In addition to determining the reliability and validity of a test, test developers also analyze the pattern of responses to individual test items. This item analysis is important for the identification of items in need of revision. It is a rare test that cannot be improved by the addition, deletion, or rewriting of test items. In fact, identifying and revising poor items can improve the overall reliability and validity of a test.

Preparing a Test Blueprint

A test blueprint (also called a table of specifications for the test) is an explicit plan that guides test construction. The basic components of a test blueprint are the specifications of cognitive processes and the description of content to be covered by the test. These two dimensions need to be matched to show which process relates to each segment of content and to provide a framework for the development of the test. It is useful for the test constructor, in planning the evaluation of a unit, to make a test blueprint that includes not only the cognitive processes and the content but also the method or methods to be used in evaluating student progress toward achieving each objective. The illustrations we use in this section are couched in terms of guidelines that an individual teacher or test constructor would use to produce a good achievement test for local use. Standardized achievement test construction applies the same procedures, but to more broadly specified curricula. When evaluating a test for content validity you would use the same steps of domain definition, but the validity question would relate to whether the test matches your domain rather than to serve as a guide for item writing. In Chapter 15 we cover the principles of item writing and analysis that experts use to produce content-valid measures of educational achievement.

A blueprint for an examination in health for an eighth-grade class is provided in Table 3.1. The test will use a short-answer, or objective, format and contain 6() items. This test is the type for which a formal blueprint is most useful, but the kind of thinking that goes into formulating a blueprint is useful even in constructing an essay test with five or six' items. The issues that are involved in the decision about the type of test item to use are considered later.

The cognitive processes to be assessed by the test are listed in the left hand column of the table. The titles of each of three content units have been entered as column headings. Each box, or cell, under the nit headings contains content entries that relate to the cognitive process on the same line with the cell. The complete blueprint specifies the content deemed important and how it will be measured. Most standardized achievement tests would cover a broader array of content than is shown here, but content definition and test construction would proceed in the same way.

An examination of the blueprint should make it clear to you that tests are just samples of student behavior —for four reasons.

1. Only those objectives suitable for appraisal with a paper-and-pencil test are included in the blueprint.
2. The entries in the cells under each area of content are examples that illustrate, but do not exhaust, the total content.
3. There are an unlimited number of items that could be written for the material that is Included in the blueprint.
4. The time available for testing is limited, and, therefore, the test can include only a small sample from the domain of all possible items.

If a test is to reflect local goals, you must carefully choose the items to include on your tests or select carefully a test that measures those goals. The following four issues should guide your construction or evaluation of the test.

1. What emphasis should each of the content areas and cognitive processes receive on the test? In other words, what proportion of all the items on the test should be written for each content area and for each cognitive process within each content area?
2. What type or types of items should be included on the test?
3. How long should the test be? How many questions or items should the total test contain? How many items should be written for each cell of the blueprint?
4. How difficult should the items be?

Relative Emphasis of Content Areas and Process Objectives. The proportion of test items allocated to each content area and to each cognitive process should correspond to the instructional emphasis and importance of the topic. The decision making process involved is subjective, but the test user should ensure that the test has maintained an appropriate balance in emphasis for both content and mental processes. Allocating a different number of items to each topic and cognitive process is the most obvious way of weighting topics and processes on the test.

The initial weighting of the content areas and cognitive processes requires the assignment of percentages to each content area and cognitive process such that the total for both is 100%. In the blueprint shown in Table 3.1, the test maker decided that Topic A. nutrition, should receive a weight of 40%; Topic B, communicable diseases,

Table 3.1 Blueprint for Final Examination in Health in Eighth Grade

Process Objectives	**Content Areas**
	A. Nutrition, 40%
1. Recognizes terms and vocabulary 20%	Nutrients Incomplete protein Vitamins Complete protein Enzymes Amino acids Metabolism Glycogen Oxidation Carbohydrate A or 5 items
2. Identifies specific facts 30%	Nutrients essential to health Good sources of food nutrients Parts of digestive system Process of digestion Of each nutrient Sources of information about foods and nutrition 7 or 8 items
3. Identifies principles, concepts, and generalizations 30%	Bases of well-balanced diet Enzyme reactions Transfer of materials between cells Cell metabolism Functions of nutrients in body 7 or 8 items
4. Evaluates health information and	Analyzes food and diet advertisements Interprets labels on foods

advertisements 10%	Identifies good sources of information about foods and diets 2 or 3 items
5. Applies principles and generalizations to novel situations 10%	Identifies well-balanced diet Computes calories needed for weight-gaining or weight-losing diet Predicts consequences of changes in enzymes on digestive system Identifies services and protection provided by the Federal Food and Drug Act 2 or 3 items
No. of items	24
Total time for test—90 minutes	

	Content Areas	
B. Communicable . Diseases, 40%	C. Noncommunicable Diseases, 20%	
Immunity Epidemic Virus Pathogenic Carrier Endemic Antibodies Protozoa Incubation period 4 or 5 items	Goiter Deficiency diseases Diabetes Cardiovascular diseases Caries 2 or 3 items	
Common communicable diseases Incidence of various diseases Methods of spreading disease Types of immunization Symptoms of common communicable diseases 7 or 8 items	Specific diseases caused by lack of vitamins Specific disorders resulting from imbalance in hormones Incidence of noncommunicable diseases Common noncommunicable diseases of adolescents and young adults 3 or 4 items	
Basic principles underlying control of disease Actions of antibiotics Body delenses against disease Immune reactions in body 7 or 8 items	Pressure within cardiovascular system Control of diabetes Inheritance of abnormal conditions Abnormal growth of cells 3 or 4 items r	
Distinguishes between adequate and inadequate evidence for medicines Identifies misleading advertisements for medications 2 or 3 items	Identifies errors or misleading information in health material Identifies appropriate source of information for health problems 1 or 2 items	6
Recognizes conditions likely to result in increase of communicable disease Identifies appropriate methods for sterilizing objects Gives appropriate reasons for regulations, processes, or treatments 2 or 3 items	Predicts consequences of changes in secretion of certain hormones Predicts probability of inheriting abnormal conditions 1 or 2 items	60
24	12	Total number of items—60

Types of Items To Be Used The types of items that can be used on a test can be classified into two categories: (1) those for which examinees produce their own answers, which are sometimes labeled supply-response, or constructed-response items, and (2) those for which students select their own answers from several choices, which are labeled select-response items. Examples of supply response items are the essay item requiring an extended answer from the student, the short-answer item requiring no more than one or two sentences for an answer, and the completion item requiring only a word or a phrase for an answer. Examples of selection type items are true-false, multiple-choice, and matching. The decision about which type of item to use will depend on the cognitive process to be measured, the strengths and weaknesses of each item type for the process and content to be measured, and the way the test will be used and scored.

Total Number of Items for the Test. If the decision is made to use an essay type of test, there will be time for only a few questions. The more elaborate the answers required, the fewer the number of questions that it is possible to include. For example, a 40-minutc test in high school might have three or four questions requiring extended answers of a page each. Select-response and short-answer tests can involve a much larger number of items.

The number of test items that can be asked in a given amount of time depends on the following factors:

1. ***The type of item used on the test.*** A short-answer item for which a student has to write his or her answer is likely to require more time than a true-false or multiple-choice item for which a student is only required to choose an answer from among several choices. Of course, "terns that call for more extended written responses will take even more time.
2. ***The age and educational level of the student.*** Students in the primary grades whose reading and writing skills are just beginning to develop require more time per test item than older students do. Young children cannot attend to the same task for a long period of time. Testing time for them must be shorter, further reducing the number of items. With younger children, achievement testing is often distributed in short blocks over several days.
3. ***The ability level of students.*** Compared to lower ability students, high-ability students have better-developed reading and writing skills. They also have a better command of the subject matter and better problem-solving skills. As a rule, high-ability students can answer more questions per unit of testing time than low-ability students of the same age and grade can. Thus, a test for an advanced class could be longer, and a test for a slower-learning class should be shorter than a test for students of average ability.
4. ***The length and complexity of the items.*** If test items are based on reading passages, tabular materials, maps, or graphs, time must be provided for reading and examining the stimulus material. The more stimulus material of this type that is used on a test, the fewer the number of items that can be included on it.
5. ***The type of process objective being tested.*** Items that require only the recall of knowledge can be answered more quickly than those that require the application of knowledge to a new situation. Thus, tests intended to assess higher cognitive processes should include fewer items for a given amount of testing time.
6. ***The amount of computation or quantitative thinking required by the item.*** Most individuals work more slowly when dealing with quantitative materials than when dealing with verbal materials; therefore, if the items require mathematical computations, the time allotted per item must be longer than that for purely verbal items.

Conclusions on Total Number of Test Items. It is impossible to give hard-and-fast rules about the number of items to be included in a test for a given amount of testing time. As a rule, the typical student will require from 30 to 45 seconds to read and answer a simple, factual-type multiple-choice or true-false item and from 75 to 100 seconds to read and answer a fairly complex, multiple-choice item requiring problem solving.

Keep in mind that there is a great deal of variation among students regarding the number of items that each student can complete in a given amount of time and that this variation is not always related to reading ability or knowledge of the content being assessed. This characteristic is also related to individual learning styles. The total amount of time required for a number of items sufficient to provide adequate coverage of the blueprint may, in some cases, be more than is available in a single class period. The most satisfactory solution to this problem is to divide the test into two or more separate subtests that can be given on successive days.

I. ITEM WRITING

When a professor announces that there will be a test, one of the first questions is "What kind of test?" Will it be a true-false test, a multiple-choice test, an essay test, or a test in which one must fill in the blanks? Not all tests are in the same format. As you will learn in Part Two of this book, personality and intelligence tests require all sorts of different responses. After defining the objectives and purpose of the test, the next question faced by the test constructor is the type of response he or she wants to require. In part this choice will be determined by the purpose of the test. For example, if it is a test that requires right or wrong answers, the task will usually be true-false, multiple choices, matching, or essay.

A. Item formats

The type of test you probably have experienced most is one in which you are given credit for a specific response. In classroom situations credit is often given for selection of the "correct" alternative for each test item and only one alternative is scored as correct. True-false and multiple-choice examinations use this system. Similar formats are used for many other purposes, such as evaluating attitudes, determining knowledge about traffic laws, or deciding whether someone has characteristics associated with a particular health condition. The simplest test of this type uses a dichotomous format.

The dichotomous format. .The dichotomous format offers two alternatives for each item. Usually a point is given for selection of only one of the alternatives. The most common example of this format is the true false exam. This test presents students with a series of statements. The student's task is to determine which statements are true and which are false. There are many virtues of the true-false test including ease of construction and ease of scoring, but the method has also become popular because a teacher can easily construct a test by copying lines out of a textbook. The lines that are copied verbatim are designated as "true." Other statements are doctored, so they are no longer true.

The advantages of true-false items include their obvious simplicity, ease of administration, and quick scoring. However, there also are disadvantages. For example, true-false items encourage students to memorize material, and it is often possible for students to perform well on a test covering material they do not really understand. Another problem is that the probability of getting any item correct by chance alone is 50%. Thus, in order for a true-false test to be reliable, it requires many items.

The dichotomous format is not unique to true-false or educational tests. Many personality tests require responses in a dichotomous format. This may be true-false, or it may be in some other two-choice format such as yes-no. Personality test constructors often prefer this type of format because it requires absolute judgment. For example, in response to an item such as "I often worry about my sexual performance," persons are not allowed to be ambivalent— they must respond true or false. Dichotomous items have many advantages for personality tests with many subscales. One is that it makes the scoring of the subscales easy. All that is necessary is to count the number of .items a person endorses from each subscale.

Although the true-false format is popular in educational tests, it is not used as frequently as the multiple-choice test, which represents the polychotomous format.

The polychotomous format. The polychotomous format is similar to the dichotomous format except that each item has more than two alternatives. Typically a point is given for the selection of one of the alternatives, and no point is given for selecting any other choice. The multiple-choice examination is the polychotomous format you have encountered most often because it is a popular method of measuring academic performance in large classes. Multiple-choice tests are easy to score, and the probability of obtaining a correct response by chance is lower than it is for true-false items. A major advantage of this format is that it takes very little time for test takers to respond to a particular item because they do not have to write. Thus the test can cover a large amount of information in a relatively short period of time.

When taking a multiple-choice examination your task is to determine which of several alternatives is "correct." All choices that are not correct are called distractors. The choice of distractor is very important.

Because most students are familiar with multiple choice tests and related formats such as matching, there is no need to elaborate on their description.

One question is "How many distractors should a test have?"

Psychometric theory suggests that the items will be more reliable if there are a large number of distractors. In other words, adding more distractors should increase the reliability of the items. This problem has been studied by a number of psychometricians, and there is general agreement about the results of the calculations. However, in practice adding distractors may not actually increase the reliability because it is difficult to find good ones. The reliability of an item is not enhanced by distractors that are never selected. Studies have shown that it actually is rare to find items for which more than three or four distractors operate efficiently. Ineffective distractors actually may hurt the reliability of the test because they are lime consuming to read and can limit the number of items that can be included in a test. After a careful review of the distractor problem, Wesman (1971) concluded that item writers should try to find three or four, good distractors for each item. They are an essential ingredient of good items.

Another issue is the scoring of multiple-choice examinations. Suppose you bring your roommate to your sociology test and he or she fills out an answer sheet without reading the items. Will your roommate get any items correct? The answer is yes—-by chance alone. If each item has four choices, the test taker would be expected to gel 25% correct. If the items had three choices, a 33.33% correct rate would be expected. Because some answers are "correct" simply by the luck of guessing, a correction for guessing is sometimes used. The formula to correct for guessing is

$$\text{Correction} = R - \frac{w}{n-1}$$

Where R = the number of right responses

W = the number of wrong responses

n = the number of choices for each item

Omitted responses are not included—they provide neither credit nor penalty. For example, suppose I hat your roommate randomly filled out the answer sheet to your sociology test. The test had 100 items, each with four choices. By chance his or her expected score would be 25 correct. Let's assume that he or she got exactly that (if the test was filled out randomly, we might not get exactly 25, which is the average random score). The expected score corrected for guessing would be

$$R - \frac{W}{n-1} = 25 - 75/(4-1) = 25 - 75/3 = 25 – 25 = 0$$

In other words, when the correction for guessing is applied, the expected score is 0.

A question frequently asked by students is "Should I guess on multiple-choice items when I don't know the answer?" The answer depends on how the test will be scored. If a correction for guessing is not used, the best advice is "guess away." By guessing you have a chance of getting the item correct. You do not have this chance if you do not attempt it. However, if a correction for guessing is used, random guessing will do you no good. Some speeded tests are scored so that the correction for the guessing formula includes only the items that were attempted. That is, those which were not attempted are not counted either right or wrong. In this case random guessing and leaving the items blank •each has the same expected effect.

How about cases when you don't know the right answer but are able to eliminate one or two of the alternatives? How many times have you had it down to two alternatives, but couldn't figure out which of the two was correct? In this case it is advisable to guess. The correction formula assumes that you are equally likely to respond to each of the four categories. For a four-choice item it would estimate your chance of getting the item correct by chance alone to be one in four. However, if you can eliminate two alternatives, the chances are actually one in two. This gives you a slight advantage over the correction formula. Other personality and altitude measures do not deem any response "right." Rather, they attempt to quantify characteristics of the response. These formats include the Likert format, the category scale, and the Q-sort.

The Likert format. One popular format for attitude and personality scales requires that a respondent indicate the degree of agreement with a particular attitudinal question. This technique is called the Likert format because it was used as part of Likert's (1932) method of attitude scale construction. A scale using the Likert format would consist of a series of items, such as "I am afraid of heights." Instead of giving a yes or no reply, five alternatives are offered: strongly disagree, disagree, neutral, agree, and strongly agree. This format is very popular in attitude measurement. For example, it is possible to determine the extent to which people endorse statements such as "The government should not regulate private business."

Responses in Likert format can be subjected to factor analysis. This makes it possible to find groups of items that go together. A similar technique that uses a greater number of choices is the category scale.

The category format. Ten-point rating scales have become so commonplace that a Hollywood movie was named "10." When asked for an evaluation of many things, people are requested to rate them on a 10-point scale. The scale need not have exactly 10 points it can have either more or fewer categories. Rating scales of this sort are called category scales.

Although the 10-point scale is very-common to psychological research and to everyday conversation, there is still controversy about when and how it should be used. However, experiments have shown that responses to items on 10-point scales are affected by the groupings of the items being rated. For example, if coaches are asked to rate the abilities of a group of 2p very talented players, they may tend to make fine distinctions among them and use many of the categories on the 10-point scale. A particular player who is rated as a 6 when he is on a team with many outstanding players might be rated as a 9 if he were judged along with a group of poorly coordinated players (Parducci, 1968, 1982). When given a group of objects to rate, subjects have a tendency to spread their responses evenly across the 10 categories (Stevens, 1966).

Recent experiments have shown that this problem can be avoided if the endpoints of the scale are very clearly defined and the Subjects are frequently reminded of the definitions of the endpoints. For example, instead of asking coaches to rate the ability of basketball players on a 10-point scale, they would be shown films of what was meant by 10 and other films of what was meant by 1. Under these circumstances, the subjects are less likely to offer a response that is affected by other stimuli in the group (Kaplan & Ernst, 1980).

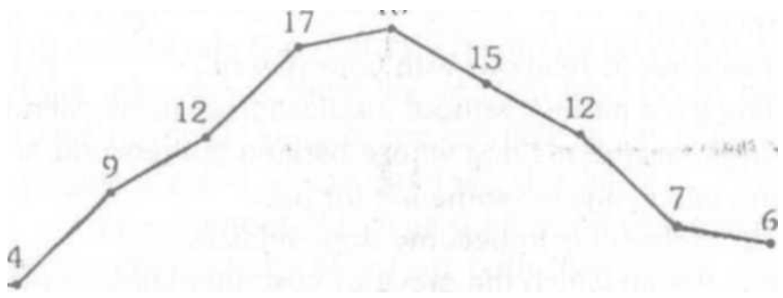
People often ask "Why use a 10-point scale instead of a 13-point scale or a 43-point scale?" As it turns out, this has been a matter of considerable study. Some have argued that the optimal number of points is around 7 (Symonds, 1924), whereas others have suggested that the optimal number of categories should be three times this number (Champney & Marshall, 1939). As is often the case, the number of categories required depends on the fineness of the discrimination subjects are willing to make. If the subjects are unconcerned about something, they will not make fine discriminations, and a scale with just a few categories will do about as well as a scale that has many. However, when people are very involved with some issue, they will tend to use a greater number of categories. For most rating tasks, however, a 10-point scale seems to provide enough discrimination. Anderson (1976) has found that a 10-point scale provides substantial discrimination between objects for a wide variety of stimuli.

Checklists and Q-sorts. One format common in personality measurement is the adjective checklist (Gough, 1960). With this method a subject is given a long list of adjectives and asked to indicate whether each one is characteristic of himself or herself. Adjective checklists can be used for describing either oneself or someone else. For example, in one study at the University of California at Berkeley raters checked the traits they thought were characteristic of a group of 40 graduate students. Half these students had been designated by their instructors as exceptional in originality, and the other half had been characterized as low in originality. The results demonstrated that the adjectives chosen to describe members of these two groups differed. The highly original students were characterized most often by the traits "adventurous," "alert," "curious," "quiet," "imaginative," and "fair-minded." In contrast, the low originality students were seen as "confused," "conventional," "defensive," "polished," "prejudiced," and "suggestible."

The adjective checklist requires subjects to endorse or not endorse self-descriptive adjectives, and it only allows these two choices. A similar technique known as the Q-sort increases the number of categories. The Q-sort can be used to describe oneself or to provide ratings of others (Stephenson, 1953). With this technique a subject is given statements and asked to sort them into nine piles. For example, Block (1961) gave observers 100 statements about personal characteristics. The statements were sorted into piles indicating the degree to which they appeared to be accurate about some particular person. If you were using this method you might be asked to rate your roommate. You would be given a set of 100 cards. Each card would have a statement on it, such as the following

- Has a wide range of interests
- Is productive; gets things done
- Is self-dramatizing; is histrionic
- Is overreactive to minor frustrations; is irritable
- Seeks reassurance from others
- Appears to have a high degree of intellectual capacity
- Is basically anxious

If a statement really hit home, you would place it in Pile 9. Those that were not at all descriptive would be placed in Pile 1. Most of the cards are usually placed in is Piles 4, 5, and 6. The frequency of items placed in each of the categories usually looks like a bell-shaped curve (see Figure 3.1). The items that end up in the extreme categories usually say something interesting about the person.



_______I_____1_____I_____1____ I_____1____ 1____ I
I

Figure 3.1: The California Q-sort. The number of items distributed in the nine piles of the California Q-sort approaches a normal distribution.

B. Other possibilities

The formats for items we have discussed are only a few of the many possibilities. If you are interested in learning more about item writing and item formats, you might check some classic references (Guilford, 1954; Edwards, 1957; Torgerson, 1958).

Unfortunately, there is no simple formula for item writing. Several people have studied the issue carefully and have contributed many useful suggestions (Ebel, 1972; Stanley & Hopkins, 1972; Sax. 1980; Wesman, 1971). If you need to write test items, you should consult these sources. However, writing good items remains an art rather than a science. There is no substitute for using precise language, knowing the subject matter, being familiar with the level of examinees, and using your imagination (Wesman, 1971) Once the items are written and they have been administered, there are item analysis techniques that can be used to evaluate them.

Essay Type Items

The essential characteristics of the task set by an essay test are that each student

1. Organizes his own answers, with a minimum of constraint.
2. Uses his own words (usually his own handwriting).
3. Answers a small number of questions.
4. Produces answers having all degrees of completeness and accuracy.

WRITING THE ITEMS FOR AN OBJECTIVE TEST

Writing good test items is an art. It is a little like writing a good sonnet and a little like baking a good cake. The operation is not quite so free and fanciful as writing the sonnet: it is not quite so standardized as baking the cake. It lies somewhere in between. So a discussion of item writing lies somewhere between the exhortation to the poet to go put and express himself and the precise recipes of a good cookbook. The point we wish to make is that we do not have a science of test construction. The guides and maxims that we shall offer are not tested out by controlled scientific experimentation. Rather, they represent a distillation of practical experience and professional judgment. As with the recipe in the cookbook, if carefully followed they yield a good product.

We shall first go over some suggestions that apply to almost any type of objective item. Then we will consider specific item types, indicating some of the general virtues and limitations of the type of item and giving more specific suggestions for writing and editing. A number of the principles that we set forth will seem very obvious. However, experience in reviewing and editing items indicates that these most obvious faults are the ones that are most frequently committed by persons who try to prepare objective tests. Thus, it hardly seems necessary to insist that a multiple-choice item must have one and only one right answer, and yet items with no right answer or several occur again and again in tests that are carelessly prepared.

GENERAL MAXIMS FOR ITEM WRITING

1. Keep the Reading Difficulty of Test Items Low in relation to the group who are to take the test, unless the purpose is to measure verbal and reading abilities. Ordinarily you do not want language difficulties to interfere with a pupil's opportunity to show what he knows.

Example

Poor: The legislative enactment most distasteful to the protagonists of labor has been the
A. Walsh-Healy Act
B. Norris-LaGuardia Act
C. Wagner Act.
D. Taft-Hartley Act

Better: The law to which labor supporters have objected most has been the
A. Walsh-Healy Act
B. Norms-LaGuardia Act
C. Wagner Act
D. Taft-Hartley Act

2. Do Not Lift a Statement Verbatim from the Textbook. This places a premium upon rote memory with a minimum of understanding. A statement can at least be paraphrased. Better still, in many cases it' may be possible to imbed the specific knowledge in an application.

Example

Poor: An injunction is a court order forbidding specified actions, such as striking or picketing by unions. T F

Better: (Paraphrased) If a court issued an order forbidding Union X to strike, this order would be called an injunction. T F

3. If an Item Is Based an. Opinion or Authority. Indicate Whose Opinion or What Authority. Ordinarily statements of a controversial nature do not make good items, but there are instances where knowing what some particular person thinks may be important for its own sake. The student should presumably be acquainted with the viewpoint of his textbook or instructor, but he should not be placed in the position of having to endorse it as indisputable fact.

Example

Poor: The basic cure for jurisdictional disputes is to remove the fear of un-employment. T F

Better: According to Faulkner and Starr, the basic cure for jurisdictional disputes is to remove the fear of unemployment. T F

4. In Planning Set of Items for a Test, Care Must Be Taken that One Item Does Not Provide Cues to the Answer of Another item or Items. The second item below gives cues to the first.

Example

1. A court order restraining a union from striking is called
A. a boycott
B. an injunction
C. a lockout
D. an open shop

2. The Taft-Hartley Act provides that an injunction may be called for 80
days to prevent strikes
A. of government or municipal workers
B. of public utility workers
C. arising out of jurisdictional disputes
D which threaten to endanger the public welfare

5. Avoid the Use of Interlocking or Interdependent Items. The answer to one item should not be required as a condition for solving the next item. This is the other side of the principle stated in 4 above. Every individual should have a fair chance al .each item m it comes. Thus, in the example shown below, the person who does not know the answer to the first question is in a very weak position as far as attacking the second one is concerned.

Example

The new labor technique introduced in the big automobile and steel strikes of 1937 was the (sit down strike) .

Public reaction to this technique was generally (unfavorable)

6. In a set 0f Items, Let the Occurrence of the correct responses Follow Essentially a Random Pattern. Avoid favoring certain responses, i.e., either true or false, or certain locations in a set of responses. Do not have the responses follow any systematic pattern.
7. Avoid Trick and Catch Questions, except in the rare case in which the test has a specific purpose of measuring ability to keep out of traps.

Example

Under the leadership of John L. Lewis, the Congress of Industrial Organizations broke away from the American Federation of Labor in 1935. T F

(This would have to be scored false, because at that date the organization was known as the Committee for Industrial Organization.)

8. Try to Avoid Ambiguity of Statement and Meaning. This is a general admonition somewhat like sin no more," and it may be no more effective. However, it is certainly true that ambiguity of statement and meaning is the most pervasive fault in objective test items. Many of the specific points already covered and to be covered deal with specific aspects of the reduction of ambiguity.

Example

The general trend in union membership since 1880 has paralleled very closely

A. business cycles

H. general economic conditions

C. the labor force

D. fluctuations in the cost of living

The keyed answer to the above question was B, but the examinee trying to answer the item is faced with several problems. First of all what is meant by "general trend"? Does this mean the general increase from 1880 to 1950, or does it also refer to all the ups and clowns in between? .Secondly, what did the writer have in mind when he wrote "union membership"? Does it mean the actual number of people belonging to a union, or does it mean the percentage of all of the people in the labor force who belong to unions? Third, how close does the relationship between union membership and any one of the options have to be before one can say that they parallel each other very closely?

Now look at the options. None of the options make very satisfactory completions for the stem. Option A, "business cycles," and option D, "fluctuations in the cost of living," are included within B, "general economic conditions." Option C is not clear. Does the writer mean the number of people in the labor force as a whole? Does he mean the occupational distribution of the people e labor force? Does he consider unemployed workers or part-workers as part of the labor force? The item needs to be sharpened up in several respects. The example below would appear to test the same knowledge and to provide less occasion for misunderstanding of what the examiner trying to say.

Example

Between 1880 and 1950 the number of workers belonging to unions increased most rapidly

A. when economic conditions were good.

B. during-periods of economic depression.

C. after court decisions that were unfavorable to labor.

D. when factories moved to rural arid undeveloped areas.

9. Beware of Items Dealing with Trivia. An item on a test should appraise some important item of knowledge or some significant understanding. Avoid the type of item that could quite justifiably b;-answered, "Who cares?" Ask yourself in each case whether knowing for not knowing the answer would make a significant difference in the individual's competence in the area being appraised.

Example

Poor: The Taft-Hartley Act was passed in

A. 1945.

B. 1946.

C. 1947.

D. 1948.

Better: Which of the following contract provisions between management and labor would be specifically prohibited under the Taft-Hartley Act?

A. All newly hired employees must join the union within 90 days after
employment.

B. No person can be employed unless he is already a member of the union.

C. . Union members will be given preference in the hiring of new employees.

D. New employees will be hired without regard to union membership or promise to join the union.

TRUE-FALSE ITEMS

The true-false item has had popularity in teacher made objective tests far beyond that warranted by its essential nature. This is probably because bad true-false items can be written quickly and easily. To write good ones is quite a different matter. Even when they are well written, true-false items are relatively restricted in the types of educational objective they can measure. They should be, limited, to statements that are unequivocally true, or demonstrably false. For this reason, they are adapted to measuring relatively specific, isolated and often trivial facts. They can also be used fairly well to test meanings and definitions of terms. But items testing genuine understandings, inferences, and applications are usually very hard to cast in true-false form. The true-false item is particularly open to attack as fostering piecemeal, fractionated, superficial learning and is probably responsible for many of the attacks upon the objective test.

It is also in this form of test that the problem of guessing becomes most acute.

The commonest variety of true-false item presents a simple declarative statement, and requires of the examinee only that he indicate whether it is true or false.

Example

T F from 1950 to 1953 John L. Lewis was the president of the CIO.

Several variations have been introduced in an attempt to improve the item type. One simple variation is to underline a part of the statement, viz., CIO in the above example. The instructions indicate that this is the key-part of the statement and that it determines whether the statement is true or false. That is, the correct-ness or appropriateness of the rest of the statement is guaranteed. The examinee can focus his attention upon the more specific issue of whether the underlined part is compatible with the rest of the statement. This seems to reduce guessing and make for more consistent measurement.

A further variation is to require the examinee to correct the item if it is false. This works well if combined with the underlining described above but is likely to be confusing if no constraints are introduced in the situation. Our example could be corrected by changing the name of the individually changing the dates, or by changing the name of the organization. Requiring that the item be corrected minimizes guessing and provides some further cue to the individual's knowledge.

CAUTIONS IN WRITING TRUE-FALSE ITEMS

1. Beware of " Specific Determiners," words that give cues to the probable answer, such as all never, usually, etc. Statements that contain '"all," "always" "no" "never," and such all-inclusive terms represent such broad generalizations that they are likely to be false. Qualified statements involving: such terms as "usually" or "sometimes" are likely to be true. The test-wise student knows this, and will use these cues, if he is given a chance, to get credit for knowledge he does not possess. "All" or "no" may sometimes be used to advantage in true statements, because in this case guessing will lead the examinee astray.

Example

Poor: All unions in the AF of L have always been craft unions. T F

Better: All closed shop contracts require that the workers belong to a union. T F

2. Beware of Ambiguous and Indefinite Terms of Degrees or Amount. Expressions such as "frequently," "greatly" "to a considerable degree." and "in most cases" are not interpreted in the same way by everyone who reads them. Ask a class or other group what they think of when you say that something happens "frequently." Is it once a week or once an hour? Is it 90 per cent of the time or 50 per cent? The variation will be very great. An item in which the answer depends on the interpretation of such terms as these is an unsatisfactory one.
3. Beware of Negative Statements and Particularly of Double Negatives. The negative is likely to be overlooked in hurried reading of an item, and the double negative is hard to read and confusing.

Example

Poor: A non-union shop is not one in which an employee must refrain from joining a union.T F

Better: Employees in u non-union shop are permitted to belong to a union. T F

4. Beware of Items that Include More than One Idea in the Statement Especially If one is true and the other Is False. This type of item borders on the category of trick items. It places a premium on care and alertness in reading. The reader must not restrict his attention to one idea to the exclusion of the other or he will be misled. The item tends to be a measure of reading, skills rather than knowledge or understanding of subject content.

Examples

According to the Taft-Hartley Act Jurisdictional strikes are forbidden but the closed shop is approved as an acceptable labor practice.T F

The CIO was composed of industrial unions, whereas the AF of L was composed entirely of craft unions.
T F

5. Beware of Giving Cues to the Correct Answer by the Length of the Item. There is a general tendency for true statements to be longer than false ones. This is a result of the necessity of including qualifications and limitations to make the statement true. The item writer must be aware of this trend and make a conscious effort to overcome it.

SHORT ANSWER AND COMPLETION ITEMS

The short-answer and the completion item tend to be very nearly the same thing, differing only in the form in which the problem is presented. If it is presented as a question it is a short-answer item, whereas if it is presented as an incomplete statement it is a completion item.

Example

Short Answer: Who followed John L. Lewis as president of the CIO? Completion: John L. Lewis was followed as president of the CIO by (Philip Murray).

Items of this type are well suited to testing knowledge of vocabulary, names or dates, identification of concepts, and ability to solve algebraic or numerical problems. Numerical problems that yield a .specific numerical solution are "short answers" in their very nature. The measurement of more complex understandings and applications is difficult to accomplish with items of this type.

Furthermore, evaluation of the varied responses that are given is likely to call for some skill and to introduce some subjectivity into the scoring procedure.

MAXIMS CONCERNING COMPLETION ITEMS

1. Beware of Indefinite or "Open" Completion Items. In the first' illustration below, there are many words or phrases that give factually correct and reasonably sensible completions to the statement: "a man," "forceful," "beetle-browed," "elected in 1936." The problem needs to be more fully defined, as is done in the revised statement.

Example

Poor: The first chairman of the CIO was (John L. Lewis)

Better: The name of the man who was the first chairman of the CIO is (John L. Lewis)

2. Don't Leave Too Many Blanks in a Statement; Omit Only Key Words. Overmutilation of a statement reduces the task of the examinee to a guessing game or an intelligence test.

Example

Poor: The (Taft-Hartley Act) makes the (closed) shop (illegal)

Better: The closed shop was outlawed by the (Taft-Hartley), Act.

3. Blanks Are Better Put Near the end of a Statement, Rather Than at the Beginning. This makes the item more like a normal question. The respondent has had the problem defined before he meets the blank.

Example

A (n) (injunction) is a court order that forbids workers to strike. r: A court order that forbids workers to strike is called a(n) (injunction).

If the Problem Requires a Numerical Answer Indicate the Units In Which It Is to Be Expressed. This will simplify the problem of scoring and will remove one possibility of ambiguity in the examinee's response.

MLTIPLE-CHOICE ITEMS

The multiple-choice item is the most flexible and most effective of the objective item types. It is effective for measuring information, vocabulary, understandings, application of principles or ability to interpret data. In fact, it can be used to test practically any educational objective that can be measured by a pencil-and-paper test except the ability to organize and present material. The versatility effectiveness of the multiple-choice item is limited only by the ingenuity and talent of the item writer.

The multiple-choice item consists of two parts: the stem, which it's the problem, and. the list of possible answers or options. Western may be presented in the form of an incomplete statement or a question.

Example

Incomplete statement: Jurisdictional strikes are illegal under the

A. Taft-Hartley Act.
B. Wagner Act.
C. Walsh-Healy Act.
D. Fair Labor Standards Act.

Question: Which one of the following labor acts outlawed jurisdictional strikes?

A. The Taft-Hartley Act.
B. The Wagner Act.
C. The Walsh-Healy Act.
D. The Fair Labor Standards Act.

Inexperienced item writers usually find it easier to use the question form of stem than the incomplete sentence form. The use of the question forced the item writer to state the problem explicitly. It rules out certain types of faults that may creep into the incomplete statement, which we will consider presently. However, the incomplete statement is often more concise and pointed than the question, if it is skillfully used.

The number of options used in the multiple-choice question differs "in different tests, and there is no real reason why it cannot vary for items in the same test. However, to reduce the guessing factor, it is preferable to have four or five options for each item. On the other hand, it is better to have only three good options for an item than to have five, two of which are so obviously wrong that no one ever chooses them.

The difficulty of a multiple-choice item will depend upon the "closeness of the options and the process called for in the item. Consider the set of three items shown below, all centered around the meaning of "strike" or "jurisdictional strike/' One can predict with a good deal of confidence that I will be passed by more pupils than II, and II by more than III. The difference between I and II is in the process involved—I calls for quite direct memory of the definition of a term, whereas II calls for recognition of the concept embedded in the complexities of a concrete situation. The difference between II and III is one of closeness of options—II calls for rather gross discrimination of major concepts, whereas III calls for differentiation of subvarieties within a single concept.

I. When the members of a union refuse to work it is called

A. a boycott.
B. an injunction.
C. a lockout.
D. a strike.

II. On a building project the bricklayers were setting up some wooden platforms to hold their bricks. Then the carpenters refused to work, claiming that this was work that they should do. This is an example of

A. a boycott.
B. an injunction.
C. a lockout.
D. a strike.

III. On a building project the bricklayers were setting up some wooden platforms to hold their bricks. Then the carpenters refused to work, claiming that this was work that they should do. This is an example of

A. a general strike.
B. a jurisdictional strike.
C. a sit-down strike.

D. a sympathy strike.

MAXIMS FOR MULTIPLE-CHOICE ITEMS

1. The Stem of a Multiple-Choice Item Should Clearly Formulate a Problem. All the options should be possible answers to a problem that is raised by the stem. When the stem is phrased as a question, it is clear that a single problem has been raised, but this should be equally the case when the stem is in the form of an incomplete statement. Avoid items that are really a series of unrelated true-false items dealing with the same general topic.

Example

Poor: The Taft-Hartley Act
A. outlaws the closed shop.
B. prevents unions from participating in politics
C. is considered unfair by management.
D, has been replaced by the Wagner Act.

Better: The Taft-Hartley Act outlaws the
A. closed shop.
B. preferential shop.
C. union shop.
D. open shop.

2. Include as -Muck of the Item as Possible in the Stem. In the interests of economy space, economy of reading time, and clear statement of the .problem, it is usually desirable to try to word and arrange the item so that the stem is relatively long and the several options relatively short. This cannot always be achieved but is an objective to be worked toward. This principle ties in with the one previously stated of formulating the problem fully in the stem.

Example

Poor: Organized labor during the 1920's
A. encountered much unfavorable Federal legislation.
B. was disrupted by internal splits.
C. showed the usual losses associated with a period of prosperity.
D. was weakened by a series of unfavorable court decisions.

Better: During the years from 1920 to 1930 the position of organized labor was weakened by
A. much unfavorable Federal legislation.
B. splits within labor itself.
C. the effects of business prosperity.
D. a series of unfavorable court decisions.

3. Don't Load the Stem Down with Irrelevant material. In certain special cases, the purpose of an item may be to test the examinee's ability to identify and pick out the essential facts. In this case, it is appropriate to hide the crucial aspect of the problem in a set of details that are of no importance. Except for this case, however, the item should be written so as to make the nature of the problem posed as clear as possible. The less irrelevant reading the examinee has to do the better.

Example

Poor: During the early 1900's employers were generally hostile to organized labor and used many devices to try to stop their workers from organizing labor unions. One of these devices was the
A. boycott.."
B. black list.
C. closed shop.
D. checkoff.

Better: A device that has sometimes been used by employers to combat the formation of unions is the
A. boycott.
B. black list.
C. closed shop.
D. checkoff.

4. Be Sure that There Is One and Only One Correct or Clearly Best Answer. It hardly seems necessary to specify that a multiple-choice item must have one and only one right answer, but in

practice this is one of the most pervasive and insidious faults in item writing. Thus, in the following example, though choice A was probably deled to be the correct answer, there is a large clement of correct-also in choices B and D. The item could be improved as shown the revised form.

Example

Poor: The provisions of the Wagner Aft (National Labor Relations Act) vigorously criticized by

A. management
B. AF of L
C. the CIO
D. Congress

Better: The provisions of the Wagner Act (National Labor Relations Act*) most vigorously criticized by

A. L the National Association of Manufacturers
B. the railroad brotherhoods
C. the industrial unions in the CIO
D. the Democrats in Congress

5. Items Designed to Measure Understanding Insights or Ability to apply Principles Should Be Presented in Novel Terms. If the situations used to measure understandings follow very closely the examples used in text or class, the possibility of a correct answer being based on rote memory of what was read or heard is very real, e second and third variations of the example on p. 60 illustrate attempt to move away from the form in which the concept was originally stated.
6. Beware of Clang Associations. If the stem and the keyed answer sound alike the examinee may get the question right just by using these superficial cues. However, superficial associations in the wrong answers represent one of the effective devices for attracting those who do not really know the fact or concept being tested. This last practice must be used with discretion, or one may prepare trick questions.

Example

Poor: In what major labor group have the unions been organized on an industrial basis?

A. Congress of Industrial Organizations.
B. Railway Brotherhoods.
C. American Federation of Labor.
D. Knights of Labor.

Better: In what major federation of labor unions would all the workers in a given company be likely to belong to a single union?

A. Congress of Industrial Organizations.
B. Railway Brotherhoods.
C. American Federation of Labor.
D. Kgights of Labor.

7. Beware of Irrelevant Grammatical Cues. Be sure that each option is a grammatically correct completion of the stem. Cues from form of word ("a" versus "an"), number or tense of verb, etc. must be excluded. Note, for example, that in. the illustration on p. 60 it is necessary to include the article in each of the separate response options.
8. Beware of Cues from the Length of the Options. There is a tendency for the correct answer to be longer than incorrect answers, due to the need to include the specifications and qualifications that make it true. Examine your items, and if necessary lengthen some of the distracters (wrong answers).

THE MATCHING ITEM

The matching item is actually a special form of the multiple-choice item. The characteristic that distinguishes it from the ordinary multiple-choice item is that instead of a single problem or stem with a group of suggested answers, there are several problems whose answers must be drawn from a single list of possible answers.

The matching item has most frequently been used to measure factual information such as the meaning of words, dates of events association of authors with titles of books or titles with plot or characters, names associated with particular events, or association of chemical symbols with names of chemicals. The matching item is a compact and efficient way of measuring this type of achievement.

Effective matching items may often be built by basing the set of items upon a graph, chart, map, diagram, or picture of equipment. Features of the figure may be labeled, and the examinee may be asked to match names, functions, etc., with the labels on the figure. This type of item is particularly useful in tests dealing with science or technology, e.g., identification of organs in an anatomy test.

However, there are many topics to which the matching item is not very well adapted. The items making up a set should bear some relationship to each other; that is, they should be homogeneous. In the case of many of the outcomes one would like to test, it is difficult to get enough homogeneous items to make up a set for a matching item.

Consider the example that appears below:

Instructions: In the blank in front of the number of each statement in Column I, place the letter of the word or phrase in Column II that is most closely related to it.

	Column I		Column II
(C)	1. Organized by crafts.	A.	Taft-Hartley Act.
(D)	2. A refusal on the part of employees to work.	B.	Industrial Revolution.
(E)	3 First president of the CIO.	C.	AF of L.
(B)	4. Resulted in a change of economic relationship between employer and employee.	D.	Strike.
		E.	John L. Lewis
(A)	5. Outlaws the closed shop.		

This example illustrates two of the most common mistakes made in preparing matching items. Look at the statements in Column I. These statements have nothing in common except that all of them refer to labor. The first statement is vague and indefinite in the way it is stated but appears to ask for a union organized by crafts. Column II includes the name of only one labor organization. Successful matching here requires almost no knowledge on the part of the student. Each item in the set can be matched in the same way. Note, too, that there are only five choices in Column IT to match "the five items in Column I. If the instructions indicate that each answer is to be used only once, then the person who knows four of the answers automatically gets the fifth by elimination, and the person who knows three has a fifty-fifty chance on the last two.

MAXIMS ON MATCHING ITEMS

1. When Writing Matching Item's, the Items in a Set Should Be Homogeneous. For example, they should all be names of labor leaders, or all dates of labor legislation, or all provisions of different congressional acts.
2. The Number of Answer Choices Should Be Greater Than the Number of Problems Presented. This holds except when each answer choke is used repeatedly, as in variations that we shall consider presently.
3. The Set of Items Should Be Relatively short It 'is better to make several relatively short matching sets than one long one because (1) it is easier to keep the items in the set homogeneous and (2) it is easier for the student to find and record the answer.
4. Response Options Should Be Arranged in a Logical Order, if One Exists. Arranging names in alphabetical order or dates in chronological order reduces the clerical task for the examinee.
5. The Directions Should Specify the Basis for Matching and Should Indicate Whether an Answer Choice May Be Used More Than Once. These precautions will guarantee a more uniform task for all examinees.

A variation on the matching; type of item which is sometimes effective is the classification type or master list. This pattern presents an efficient means of exploring range of mastery of a concept or related set of concepts.

Example

Below are given some newspaper reports about actions taken by employees. For each of these, you are to mark the action

A if it is an ordinary strike.

B if it is a sit-down strike.

C if it is a jurisdictional strike.
D if it is a sympathetic strike.
E if it is none of the above.

(E) 1. At 12 noon yesterday all the government employees in France stopped work for 24 hours.

(B) 2. At 10:30 this morning all the workers in the Forman plant shut down their machines. They refused to leave the plant.

(C) 3. An election at the electric appliance factory failed to give a majority to either the CIO or AE of None of the workers reported for work today.

(D) 4. Truck drivers quit today when ordered to haul freight from docks-on which the longshoremen were on strike.

The above question can be made to yield further information about the examinees by requiring each student to classify each action further by the code P, permitted by the Taft-Hartley Act: F. forbidden by the Taft-Hartley Act; or U, there is some doubt whether the Taft-Hartley Act permits this action or not. In this instance, the additional task would probably also substantially increase the* difficulty of the item.

Another setting in which the master list variation of the classification type" of item .can often be used to advantage is that of testing knowledge of the general chronology' or sequence of events. An example is shown here.

For each event on the left, pick the choice on the right that tells when, the event took place.

	Event		Time
(A)	1. Beginnings of Industrial Revolution.	A.	Before 1820.
(D)	2. Formation of CIO.	B.	Between 1820 and 1860.
(B)	3. Formation of Knights of Labor.	C.	Between 1860 and 1915.
(E)	4. Taft-Hartley Act.	D.	Between 1915 and 1940.
(C)	5. The great Pullman strike (and possibly others).	E.	Since 1940.

THE ESSAY TEST

The essay test consists of such problems as:

Compare the craft guilds of medieval Europe with the two major labor unions that exist in the United States today.

What were the reasons for the decrease in union membership after World War I?

How did the Wagner Act affect labor's rights to bargain collectively?

Depending on the purpose of the test, different types of test scores can be used. Norm-referenced scoring provides ii relative evaluation of test takers, comparing them to each other or to a specific norm group or standardization sample. When maximal-performance tests use norm-referenced scores, item difficulty is an important concern. When typical-performance tests use norm-referenced scores, item discrimination is an important concern. Criterion-referenced scoring evaluates test-taker performance relative to the test domain, determining where test takers fall relative to some standard of performance or criterion. Criterion-referenced scores are only found in maximal-performance tests.

Lesson 04

ITEM ANALYSIS

The term item analysis refers to a group of statistics that can be calculated for individual test items. There are a variety of item statistics that can be computed and a variety of calculation techniques. The Three most commonly used statistics are item difficulty, item discrimination and distractor power. Although these statistics are usually discussed in regard lo multiple-choice ability tests, two of them, item difficulty and item discrimination, can be adapted to the analysis of short-answer and essay questions and are also used on personality, interest, and altitude tests.

LOGIC OF ITEM ANALYSIS

Consider a 30-item test containing 10 items on which all test takers are incorrect. It is easy to illustrate how the presence of these difficult items threatens both the reliability and validity of the test. If all test takers are incorrect on 10 out of 30 items, the distribution of scores on this test is likely to be similar to the Item Analysis distribution of scores on a 20-item test. The presence of so many difficult items in essence reduces test length by one-third. Reducing the number of items on which test takers can differ reduces the potential variability of test scores.

Test length and the variability of test scores were identified as factors influencing test reliability. In fact, the increase in reliability to be expected by increasing the length of a test can be estimated by, the Spearman-Brown prophesy formula. Conversely, when test length decreases, reliability decreases as well. A test with a large proportion of difficult items generates scores with properties similar to a shorter test including lower reliability.

How does the presence of these 10 very difficult items affect the validity of the test? Chapter 7 presented several important relationships between the reliability and the validity of a test. For example, reliability was discussed as a "necessary but not sufficient condition" for test validity. Furthermore the reliability of a test was identified as a factor placing actual statistical limits on that test's potential validity. If the presence of a large proportion of difficult items is likely to reduce test reliability, it also is likely to-reduce the test's validity.

In short, item statistics such as item difficulty can help explain why a test shows a "certain level of reliability and Validity....Item, analysis .is particularly useful when tests are unreliable or fail to demonstrate predicted relationships with criterion measures. The test may include poorly worded questions that elicit guessing or questions not measuring the appropriate construct or content domain. The reliability and validity of the test can be improved-by-removing or rewriting these items.

Table 4.1 presents fictitious data from a 20-item multiple-choice test administered to a 30person biology class. These data will be used to illustrate item-analysis procedures throughout the chapter.

Table 4.1 Sample Test for Item Analysis

The following charts present data on a 20-question, four-alternative, multiple-choice test taken by 30 biology students. The first chart shows the answers for the 10 students with the highest test grades (top 10), the second chart the answers for the 10 students with the middle 10 grades (middle 10), and the third chart the answers for 10 students with the lowest grades (bottom 10).

In each chart,, the top row lists the total grade for each student in that group (maximum correct = 20). The left-hand column in each chart lists the question numbers (I to 20), followed by the correct answer in parentheses. Questions answered correctly are marked with a. The remaining entries indicate the incorrect alternatives each student selected for each question.

1. Top 10 students in the class

Question Number	Correct Answer	Total Number of Questions Correct

		19	**18**	**18**	**17**	**17**	**16**	**16**	**16**	**16**	**16**
I	(A)	*	*	*	*	*	*	*	*	*	*
2	(C)	*	B	*	*	*	*	*	A	*	*
3	(D)	*	*	*	*	*	*	*	*	B	*
4	(B)	*	*	*	*	*	A	A	*	*	D
5	(C)	*	*	*	*	B	*	*	*	*	B
6	(B)	C	*	*	*	*	*	*	*	*	*
7	(A)	*	*	*	B	B	*	D	*	*	*
8	(D)	*	*	*	*	*	*	*	B	C	*
9	(A)	*	*	*	B	*	*	*	*	*	C
l0	(C)	*	A	B	*	*	*	*	*	*	*
11	(B)	*	*	*	*	*	C	A	C	*	*
12	(A)	*	*	*	*	*	*	*	B	*	*
13	(D)	*	*	*	*	*	*	B	*	*	*
14	(C)	*	*	*	A	*	*	*	*	A	*
I5	(B)	*	*	C	*	*	*	*	*	*	*
16	(A)	*	*	*	*	*	B	*	*	C	*
17	(D)	*	*	*	*	*	*	*	*	*	*
18	(C)	*	*	*	*	A	D	*	*	*	*
19	(D)	*	*	*	*	*	*	*	*	*	A
20	(B)	*	*	*	*	*	*	*	*	*	*

ITEM-DIFFICULTY ANALYSIS

Item-difficulty analysis is appropriate for maximal performance tests— achievement and aptitude tests— because the analysis requires that test items be score as correct or incorrect.' It is not appropriate for typical performance tests, such as personality tests or interest inventories. The most common measure of item difficulty is the percentage of test takers who answer the item correctly. Referred to as the item-difficulty index, it is represented by the symbol p and calculated as follows:

$$p = \frac{\textit{Number} \text{ of persons answering item correctly}}{N}$$

In which

p = item difficulty for a particular test item

N = the total number of people taking the test

2. Middle 10 students in the class

Question Number	**Correct Answer**	**Total Number of Questions Correct**

		15	**15**	**15**	**15**	**14**	**14**	**14**	**13**	**13**	**12**
I	(A)	*	*	*	*	*	*	*	*	*	*
2	(C)	B	B	*	*	D	*	*	A	*	D
3	(D)	*	*	A	*	*	*	*	*	B	*
4	(B)	*	A	*	A	C	*	*	D	D	D
5	(C)	*	*	*	*	*	B	B	*	*	B
6	(B)	C	*	D	*	*	*	*	*	*	*
7	(A)	*	*	*	D	B	B	B	C	*	*
8	(D)	*	*	*	*	*	*	*	B	C	*
9	(A)	*	*	*	*	*	*	*	*	*	C
l0	(C)	*	A	B	*	*	*	*	*	*	*
11	(B)	A	C	*	*	A	C	C	*	C	A
12	(A)	*	*	*	*	*	*	*	B	*	*
13	(D)	*	*	*	B	*	*	*	*	*	*
14	(C)	*	*	*	*	A	*	*	*	A	*
I5	(B)	A	*	C	C	*	*	*	*	*	A
16	(A)	*	*	*	*	B	D	D	C	B	*
17	(D)	C	B	C	A	*	C	C	B	A	*
18	(C)	*	*	*	*	*	A	A	*	*	D
19	(D)	*	*	*	*	*	*	*	*	*	A
20	(B)	*	*	*	*	*	*	*	*	*	*

The calculation of *p* may seem somewhat familiar. It is a component of the Ruder-Richardson 20 formula for calculating an internal consistency reliability coefficient, along with its counterpart *q*, the proportion of people incorrect on the item.

Several points are worth nothing before going further first, *p* is a *proportion* that varies between 0.0 and 1.0. It cannot he negative since it is based on the number of people who answer correctly. Second, *p* is based on the number of people *correct* on the item. A high *p* value, such as .9 means that most people answered the item correctly. An item with a high *p* value is actually a rather easy item. On the other hand, a low *p* value, such as .2, means that most people answered the item incorrectly. Difficult items, therefore, have *p* values closer to 0.0.

Third, calculation of *p* requires answers to test items to be categorized as correct or incorrect. On alternate choice items, such as multiple choice or true false, test-taker responses naturally fall into these dichotomous categories. On a free-response item, such as a short-answer or essay question, test-taker responses are likely to fall into several categories representing the number of points earned on the item. It is possible, however, to dichotomize performance on free-response items. The test developer can select a criterion to classify a test faker's response as correct or incorrect. For example, on a 5-poinl short-answer question, the test developer might decide that persons earning at least 4 points will be counted as correct, while those earning less than 4 points will be counted as incorrect. The criterion can then be used to transform the existing scores to the dichotomous categories needed for an item-difficulty analysis.

3. Bottom 10 students in the class

Question Number	Correct Answer	Total Number of Questions Correct

		11	**11**	**11**	**10**	**10**	**10**	**9**	**9**	**9**	**8**
I	(A)	*	B	*	C	*	*	*	C	C	D
2	(C)	*	*	A	*	A	D	D	*	D	A
3	(D)	*	*	*	B	*	C	C	A	*	*
4	(B)	C	*	*	*	*	*	*	D	*	*
5	(C)	A	B	A	A	B	A	A	B	A	A
6	(B)	*	*	*	*	*	*	*	*	*	*
7	(A)	D	D	*	B	D	C	B	D	*	B
8	(D)	*	A	B	C	*	*	A	*	B	C
9	(A)	*	*	*	*	C	B	*	C	D	D
l0	(C)	B	*	*	*	*	*	D	A	*	B
11	(B)	*	C	*	D	*	*	*	*	A	A
12	(A)	C	*	C	*	*	*	C	B	B	C
13	(D)	*	*	*	*	A	C	*	B	C	*
14	(C)	B	D	*	A	D	A	D	A	A	*
I5	(B)	A	*	D	D	C	*	*	*	*	*
16	(A)	C	*	B	C	D	C	B	*	B	C
17	(D)	A	A	A	*	*	*	*	*	*	B
18	(C)	*	*	*	*	B	B	**B**	A	D	A
19	(D)	*	C	C	A	*	B	*	*	*	*
20	(B)	*	C	D	*	A	*	A	*	*	*

Finally, no test item can be said to have a single *p* value. Item difficulty, like test reliability, can be calculated every time a test is administered. An item's difficulty index is specific to the test data under study. It is possible that a test item will vary in level of difficulty across different types of test takers. For example, a test item covering an aspect of third grade math might have a low /; value when used with second graders, but a high /; value when used with fourth graders.

Table 4.2 illustrates the calculation of item difficulty using the 20-item multiple-choice test presented in Table 4.1. Most items have *p* values in the .6 to .8 range. Items 7 and 16 stand out as the most difficult items (*p* = .5); item 6 is the easiest item {*p* = .9). But how are these values to be interpreted'.'

Table 4.2 Calculating Item Difficulty

Item	Number of People Correct	*p*[1]
1	25	.83
2	17	.57
3	23	.77
4	18	.60
5	16	.53
6	27	.90
7	15	.50
8	20	.67
9	21	.70
10	22	.73
11	16	.53
12	22	.73
13	24	.80
14	18	.60
15	21	.70

16	15	.50
17	18	.60
18	19	.63
19	24	.80
20	26	.87

$$p = \frac{\text{Number of people correct}}{\text{Total number of people}}$$

Based on the item-difficulty analysis, are these question any good or should they be revised?

The use of *P* values as a measure of item difficulty has several interesting implications. First, the *p* value is basically a behavioral measure. Rather than defining difficulty in terms of some intrinsic characteristic of the item, with this method difficulty is defined in terms of the relative frequency with which those taking the test choose the correct response. Second, difficulty is a characteristic of both the item and the population taking the test. A math problem that is very difficult when given in a high school course will be very easy when given in a graduate physics course.

Perhaps the most useful implication of the *p* value is that it provides a common measure of the difficulty of test items that measure completely different domains. It is very difficult to determine whether answering a particular question about history involves knowledge that is more obscure, complex, or specialized than that needed to answer a math problem. When *p* values are used to define difficulty, it is very simple to determine whether an item on a history test is more difficult than a specific item on a math test taken by the same group of students.

Effects of Item Difficulty on Test Scores

When everyone in a class chooses the wrong answers to a particular test item, there is usually uproar, and, with luck, the offending, item will be dropped from the test. Although infuriated students may not see the problem in technical terms, there is a sound psychometric basis for dropping such an item.

As stated in Chapter 1. one of the basic assumptions of measurement is that there are systematic individual differences in the construct or the content domain being measured. Tests represent a method of quantifying these differences. When nobody chooses the correct answer (a *p* value of .0) there are no individual differences in the "score" on that item. Note that the same is true when *everyone* chooses the correct response (a *p* value of 1.0). This suggests one of the most important principles of item analysis. An item with a *p* value of .0 or a *p* value of 1.0 does not contribute to measuring individual differences and thus is almost certain to be useless.

When we compare the test scores of two people, we are typically interested in knowing who had the highest score or how far apart the two scores were. A test item with a *p* value of ,0 or 1.0 has no effect on the differences between scores received by two subjects—dropping all of the test items with /; values of .0 or 1.0 from a test will not affect the rank order or the size of the differences between different people's scores. Test items with *p* values of .0 or 1.0 may affect the test mean but have no affect whatsoever on the test's reliability or validity, nor on the decisions based upon the test scores.

Item difficulty has a profound effect on both the variability of test scores and the precision with which test scores discriminate among different groups of examinees. The effects of difficulty on the variance of test scores are fairly obvious when *p* values are extreme. When all the test items are extremely difficult, the great majority of the test scores will be very low. When all items are extremely easy, most test scores will be extremely high. In either case, test scores will show very little variability. Thus, extreme *p* values directly restrict the variability of test scores.

The variability of test scores is maximized when *p* values average around .5. In fact, test scores are more variable when *the entire p* values cluster around .5 than when there is some range of *p* values. Since most tests are designed to measure some construct or attribute over a range of possible scores, some variability in test scores is to be expected. When tests are designed for the general measurement of some continuous variable (e.g., intelligence, need for achievement), there is little doubt that items with *p* values near .5 are preferred over extremely easy or extremely difficult items. In most analyses of item difficulty, items with *p* values near .5 should be considered optimum.

Some tests are not designed to provide equally good measurement at all points on a continuum. For example, a simple reading comprehension test may be used to screen out people whose reading ability is below some minimal level required on a job. If the job demands that the workers be above the twentieth percentile in a general reading ability, it makes little difference whether the worker reads at a high school, college, or post-graduate level. A test will be maximally effective in making this type of screening decision if the item difficulties correspond roughly with the cutting point. In order to decide whether each applicant is above or below the twentieth percentile, a test that consist of items with *p* values near .2 would be best (Lord, 1952). If the objective is to screen the very top group of applicants (e.g.. medical admissions), the test should be composed of very difficult items.

Item Discrimination

Every item can be thought of as a separate test. Some items do a pretty good job of measuring the same thing that is measured by the test as a whole. Some items measure nothing at all (i.e., items with *p* values of .0 or 1.0); others measure the wrong thing altogether. One of the principal aims of item analysis is to discover which items best measure the construct or attribute the test was designed to measure.

If the test and a single item both measure the same thing, one would expect people who do well on the test to answer that item correctly and those who do poorly to answer that item incorrectly. In other words, a good item *discriminates* between those who do well on the test and those who do poorly. In this section, we discuss three statistics that can be used to measure the discriminating power of an item: the discrimination index, the item-total correlation, and inter-item correlations.

Discrimination Index

The method of *extreme groups* can be applied to compute a very simple measure of the discriminating power of a test item. If a test is given to a large group of people, the discriminating power of an item can be measured by comparing the number of people with high test scores (e.g., those in the lop 25 percent of the class) who answered that item correctly with the number of people with low scores (e.g., the bottom 25 percent) who answered the same item correctly. If a particular item is doing a good job of discriminating between those who score high and those who score low more people in the top-scoring group will have answered the item correctly.

The item discrimination index, *D,* is based on the simple rationale described above. The first step in computing this index is to select an upper group and a lower group. Customarily, these extreme groups are made up of either those with scores in the upper and lower 27 percent or those with scores in the upper and lower 33 percent of the group taking the test (Cureton. 1957). For all practical purposes, these extreme groups can include anywhere from 25 percent to 35 percent of the examinees; any breakdown within this range yields similar discrimination measures. Once extreme groups are formed, the next step is to calculate the percentage of examinees passing each item in both the upper group and the lower group. The item discrimination index is simply the difference between these two percentages.

An example will help clarify the item discrimination index. A 40-item test was given to 100 students. The 27 students with the highest test scores formed the upper group, and the 27 with the lowest scores formed the lower group. The percentage of those in the upper group and the percentage of those in the lower group passing each item on the test were then computed. Data from four of these items are presented in Table 4.3.

Recall that the *p* value is a measure of item difficulty. The data in the table suggest that both items 1 and 2 were substantially more difficult for the lower group than for the upper group. The logic of the

Table 4.3 PERCENT PASSING

Item	Upper group	Lower group	*D*
1	71	42	29
2	60	24	36
3	47	42	5

4	38	61	-23

D statistic is simple. The test itself was by definition substantially more difficult for the lower group than for the upper group. If an item and a test both measure the same thing, the item should also he more difficult for the lower group than for the upper. The *D* statistic provides a simple but efficient measure of the discriminating power of each test item (Engelhart. 1965).

Item 3 docs not show much discriminating power. The *D* value for this item is very small, reflecting the fact that the item was about equally difficult in the upper and lower groups. The statistics for item 4 are even more discouraging. This item shows plenty of discriminating power, but in the wrong direction. A negative A) index indicates that an item is easier for people who do poorly on the test than for those who do well on the test. This is the exact opposite of what would be expected if the item and the test were measuring the same thing. The *D* values suggest that both item 3 and item 4 are poor test items. The small *D* value for item 3 suggests that it does not make any discrimination between those who do well on tests and those who do poorly. The negative *D* value for item 4 suggests that the item discriminates, but in the wrong direction. One goal of test construction is to generate items that. Like items 1 and 2 allow for valid discrimination between high and low scorers.

Shortcuts in computing D. The procedure for computing D described in the previous section is fairly simple, but there are even simpler methods available. The general formula for the D statistic is

$$D = \frac{U}{n_u} - \frac{L}{n_l}$$

where
U = number of people in the upper group who passed the item
n_u = number of people in the upper group
L = number of people in the lower group who passed the item
n_l = number of people in the lower group

In many cases, the upper group and the lower group are equal in size. In this case, where $n_u = n_l$

Above mentioned formula reduces to

$$D = \frac{U - L}{n}$$

where

$n = n_l = n_u$

Thus, in many cases D can be obtained by computing the difference between the number of people in the upper group who pass an item and the number in the lower group who pass the same item, and dividing by n.

Item-total correlation. The discrimination index is a simple statistic that provides a good deal of information on the degree to which a particular item and the total test measure the same thing. There is a more familiar statistic, the item-total correlation that provides the same sort of information. As the name implies, this statistic represents the simple correlation between the "score" on an item (a correct response usually receives a score of 1; an incorrect response receives a score of 0) and the total test score. This correlation is often referred to as a point-biserial correlation. Years ago when computers were not available, there was some real utility in developing simple computational formulas for different types of correlations. Since computers are now readily available, there is no good reason to treat an item-total correlation differently than any other correlation. Thus, we will avoid the term "point biserial" and emphasize that the item-total correlation is a simple correlation coefficient.

The item-total correlation is interpreted in much the same way as the item discrimination index, D. A positive item-total correlation indicates that the item successfully discriminates between those who do well on the test and those who do poorly. More important, a positive item-total correlation indicates that the item measures the same thing that is being measured by the test. An item-total correlation near zero indicates that the item does not discriminate between high and low scores. A negative item-total correlation indicates that the scores on the item and scores on the test. Those who do well on an item with a negative item-total r do poorly on the test.

The principal advantage of the item-total correlation is its familiarity. It is the simple correlation between item scores and test scores. Therefore, it is easy to test the statistical significance of an item-total correlation. It also is easy to make judgments about practical significance. If the item-total r is .40, we know that the item can account for 16 percent of the variability in test scores. An item discrimination index of .40 cannot be reexpressed in such concrete terms. Finally, item-total correlations are directly related to the reliability of a test (Nunnally, 1978).

Inter-item correlations. When conducting an item analysis, it makes sense to compute the correlations among all test items. The resulting inter-item correlation matrix is packed with information. First, it is possible to compute the reliability of a test given the average inter-item correlation and the number of items on the test. This is a useful fact, but this property of inter-item correlations is not the most important. The most important use of inter-item correlations is in understanding measures of item discrimination. Up to this point, we have not said why a particular item might show high or low levels of discriminating power. It might be obvious that items that show large positive item-total correlations also will show positive correlations with the most test items. It is not always obvious why other test items show low item-total correlations. Examination of the inter-item correlations can help us understand why some items fail to discriminate between those who do well on the test and those who do poorly.

If the item-total correlation is low, there are two possible explanations. First, it is possible that the item in question is not correlated with any of the other items on the test. In this case, the item should be either substantially rewritten or discarded altogether. The second possibility is that the item shows positive correlations with some test items but correlations near zero, or perhaps even negative correlations, with other items on the test. This would occur, for example, if the test measured two distinct attributes. Thus, a mathematics test that contained several complexly worded problems might measure reading comprehension as well as mathematical ability. Reliability theory suggests that such a test, in which different items measure different things (and thus are uncorrected), would not provide consistent measurement.

DISTRACTOR POWER ANALYSIS

Tests of achievement and aptitude, whether norm or criterion referenced, often use a multiple-choice format. In a multiple-choice question, an incorrect alternative is called a distractor. Distractor power analysis evaluates the percentage of people selecting each incorrect alternative to determine if the distractors are useful. Perhaps you have read multiple-choice questions containing one or two obviously wrong alternatives. Even if you didn't know the material covered by the question, you could eliminate these alternatives and thus increase the probability of correctly guessing the right answer. The presence of these poorly written distractors detracts from the overall quality of the test because people can earn higher scores just by guessing. The test's ability to identify people who know the material and people who do not is reduced. Furthermore, the ability to eliminate one or two alternatives may encourage people lo guess more often. Guessing reduces test reliability and therefore limits the test's potential validity. Distractor power analysis identifies alternatives in need of revision and therefore can facilitate the development of a more reliable and valid test.

Expected and Actual Distractor Power

A well-written multiple-choice question has two characteristics. First, people who possess the knowledge or skills being tested select the correct answer. Second, people who lack the knowledge or skills tested select randomly from the available choices. If people in this latter group approach the item randomly, an equal proportion should select each alternative. Some of these people will guess correctly. The remaining people, those who are incorrect on the item, should be equally distributed across the different distractors.

The number of people expected lo pick each distractor by random choice is the expected distractor power or "expected pull" and is calculated as follows:

$$\text{expected distractor power} = \frac{\text{number of people incorrect on item}}{\text{number of distractors}}$$

The number of people selecting each distractor is the actual distractor power or "actual pull." It is compared lo the expected power to judge the adequacy of the distractors. Distractors that are never selected or are selected less often than expected should be examined closely and probably be rewritten. It is possible that these distractors are obvious wrong answers. Distractors that are selected much more frequently than expected also need rewriting. In fact, these distractors may be so similar to the correct answer that even people who know the information selects them.

Interactions among Item Characteristics

An item analysis yields three types of information: (a) information about distractors, (b) information about item difficulty, and (c) information about item discrimination power. These three types of information are conceptually distinct, but empirically related. Thus, examining distractors reveals something about difficulty; item difficulty directly affects item discriminating power. In this section, we briefly describe ways in which item characteristics interact.

Distractors and Difficulty

A multiple-choice item which asks the year in which B. F. Skinner published *The Behavior of Organisms* could be very difficult or very easy depending on the distractors. Consider these two versions of the same item.

A. In what year did Skinner publish The Behavior of Organisms'!

a. 1457
b. 1722
c. 1938
d. 1993

B. In what year did Skinner publish The Behavior of Organisms'?

a. 1936
b. 1931
c. 1938
d. 1942

Even a person who knows very little about the history of psychology is likely to find version A to be an easy item. On the other hand, version B is likely to be extremely difficult. The difficulty of an item is greatly affected by the plausibility of the distractors. If the examinee knows nothing about the domain being tested, any distractor might be equally plausible. In general, however, people taking tests have some knowledge of the domain and are not fooled by ridiculous distractors. On the other hand, examinees usually have an imperfect knowledge of the domain and therefore may be fooled by extremely plausible distractors. Some items are extremely difficult because of one or two extremely popular distractors. In either case, it may be possible to substantially change the Item's difficulty by rewriting some distractors.

Difficulty and Discrimination

The level of difficulty places a direct limit on the discriminating power of an item. If everyone chooses the correct response (p = 1.0), or if everyone chooses an incorrect response (p = .0), item responses cannot possibly be used to discriminate between those who do well on the test and those who do poorly. When the p value is near .0 or near 1.0, the ability to discriminate between individuals is restricted.

Table 4.4 shows the maximum possible value of the item discrimination index (D) for test items at various levels of difficulty. The table shows that items with p values near .50 have the maximum potential to be good discriminators. It is important to keep in mind that a p value near .50 docs not guarantee that an item will be a good discriminator. Extreme p values place a direct statistical limit on the discriminating power of an item. When the p values are near .50, there are no statistical limits on discriminating power. Nevertheless, a poor item with a /; value of .50 is still a poor item.

Extreme p values place a direct statistical limit on the discriminating power of an item. When the p values are near .50, there are no statistical limits on discriminating power. Nevertheless, a poor item with a /; value of .50 is still a poor item.

Table 4.4 Maximum Value of the Item Discrimination Index (d) As A Function of Item Difficulty

Item p value	Maximum D

1.00	.00
.90	.20
.80	.40
.70	.60
.60	.30
.50	1.00
.40	.80
.30	.60
.20	.40
.10	.20
.00	.00

It is difficult to write a good test item. There usually is no problem in writing the stem of the question or framing the correct response; the trick is to write good distractors. The lack of discriminating power shown by test items often can be attributed to poor distractors. The presence of one or more completely implausible distractors serves to lower the difficult of aji item. As discussed in the previous section, items that are extremely easy have little or no potential for making valid discriminations. The same is true of extremely difficult items.

Distractors should be carefully examined when items show negative D values or negative item-total correlations. When one or more of the distractors looks extremely plausible to the informed reader and when recognition of the correct response depends on some extremely subtle point, it is possible that examinees will be penalized for partial knowledge. Consider, for example, the following item:

Reviews of research on the effectiveness of psychotherapy have concluded that

A. psychotherapy is never effective.

B. psychotherapy is effective, but only in treating somatoform disorders.

C. psychotherapy is effective; there are no major differences in the effectiveness of different types of therapy.

D. psychotherapy is effective: behavioral approaches are consistently more effective than all others.

Choice c is correct. A student who is familiar with research on psychotherapy, but whose professor is strongly committed to praising the virtues of behavior therapy, might mistakenly choose d. Those examinees who have no inkling of the correct answer may choose randomly and do better than those who have some command of the domain being tested but who do not recognize the correct response. This sort of an item will not contribute to the overall quality of the test. Rather, it is a source of measurement error and should be either revised or removed. The revision of one or more distractors may dramatically increase the discriminating power of poor items.

The Item Characteristic Curve

Tests are usually designed to measure some specific attribute, such as verbal ability. The more of this attribute a person has. The more likely the person will answer each test item correctly. The item characteristic curve (ICC) is a graphic presentation of the probability of choosing the correct answer to an item as a function of the level of the attribute being measured by the test. The item characteristic curve serves as the foundation of one of the most powerful theories in modern psychometrics. item response theory (or latent trait theory) The ICC also summarizes much of the information conveyed by item analysis, and suggests how this information might be used to understand the relationships between the attribute being measured and test responses (Lord, 1977; Lord & Novick. 1968; Rasch. 196QJL

Item Characteristic Curves and Item Response Theory

Traditional item analysis procedures supply us with good measures of discrimination and difficulty. If item characteristic curves simply provided alternate measures of these vital item characteristics, they would hardly be worth the trouble. There are, however, some unique advantages associated with ICCs. In particular, they illustrate the basic ideas that underlie item response theory.

Item response theory was constructed to explain and analyze the relationship between characteristics ol the individual (e.g., ability) and responses to individual items (Hulin, Drasgow & Parsons. 19S3; Lord. 1980; Thisscn & Steinberg. 1988; Weiss. 1983). Reliability and validity theories typically concentrate on test scores and only peripherally deal with individual items. Thus, item analysis represents a sort of ad hoe follow-up analysis when used in the context of explaining reliability or validity. Item response theories, on the other hand, are developed precisely for the purpose of understanding how individual differences in trails or attributes affect the behavior of an individual when confronted with a specific item.

Item response theory starts with a set of assumptions about the mathematical relationship between a person's ability and the likelihood that he or she will answer an item correctly. These assumptions form a basis for item characteristic curves, which represent a combination of assumptions regarding underlying relationships and the empirical outcomes of testing. To the extent that these assumptions are true, item response theory allows precise inferences to be made about underlying trails (e.g.. ability) on the basis of observed behavior (e.g., item responses).

One of the principal advantages of item response theory is that it provides measures that are generally sample invariant. That is. The measures that are used to describe an item characteristic curve do not depend on the sample from which test data are drawn. The same is not true for standard item analysis statistics. The same reading test that is difficult (low/; values) for a group of fourth graders may be less difficult and more discriminating for a group of sixth graders but extremely easy for a group of eighth graders. The invariance of measures obtained using item response theory is thought to be important, since it allows characteristics of items to be analyzed without confounding them, as traditional item analyses do. With the characteristics of the people taking the test.

Hem response theory and test-taking behavior. Most proponents of item response theory focus on the technical advantages of this approach, such as sample invariance. or on the fact that this approach allows you to tackle issues that are very difficult to handle in traditional item analysis (applications of this theory to three such issues are described in the section follows). In our opinion, the real advantage of item response theory may be conceptual rather than mathematical. Unlike the traditional approach to item analysis, this theory encourages you to think about why people answer items correctly.

The most widely cited model for responses to typical test items suggests that there are really two things that explain a correct response: luck and ability. The mathematical methods used in constructing ICCs under this model include a parameter that represents the susceptibility of the item to guessing, another that represents the role of ability in item responses, and a third that represents item discriminating power. Item response theory defines difficulty in terms of the level of ability needed to answer the item correctly, with a given level of probability. In contrast, the traditional definition of difficulty says nothing about what makes an item difficult or easy. Thus, traditionally an item is difficult if most people answer incorrectly and easy if most people answer correctly. Item response theory states that a difficult item is one that requires a high level of ability to achieve a correct answer, while an easy item is one that can be answered by people with lower ability levels.

In addition to providing a better definition of item difficulty than the traditional approach, this theory provides a better definition of discriminating power. In item response theory, discriminating power is defined in terms of the relationship between item responses and the construct the test is designed to measure. Thus, if individuals who are very high on spatial ability ate more likely to answer an item on a spatial ability test than individuals who are low cm-that ability, the item shows high discriminating power. The traditional approach links item responses to total test scores rather than to the construct the test is designed to measure. One implication is that traditional item discrimination statistics are only as good as the test itself. If the test is a poor measure of the construct, the relationship between item scores and total test scores may tell you little about the worth of the item.

Practical Applications of Item Response Theory

Although the mathematical complexity of item response theory has to some extent stood in the way of its widespread use. There are several advantages unique to item response theory that explain why this approach has become increasingly popular. First, as mentioned earlier, the theory produces measures of item difficulty, discrimination, and so on, that are invariant across different samples of people who take the test. Second, item response theory can be applied to solving several problems that are difficult to solve using traditional approaches.

ITEM ANALYSIS OF SPEEDED TESTS

Whether or not speed is relevant to the function being measured, item indices computed from a speeded test may be misleading. Except for items that all or nearly all examinees have had time to attempt, the item indices found from a speed test will reflect the position of the item in the test rather than its intrinsic difficulty or discriminative power. Items that appear late in the test will be passed by p relatively small percentage of the total sample, because only a few persons have time to reach these items. Regardless of how easy the item may be, if it occurs late in a speeded test, it will appear difficult. Even if the item merely asked for one's name, the percentage of persons who passed it might be very low if the item were placed toward the end of a speeded test.

Similarly, item discrimination indices tend to be overestimated for those items that have not been reached by all test takers. Because the more proficient individuals tend to work faster, they are more likely to reach one of the later items in a speed test. Thus, regardless of the nature of the item itself, some correlation between the item and the criterion would be obtained if the item occurred late in a speed test.

To avoid some of these difficulties, we could limit the analysis of each item to those persons who have reached the item. This is not a completely satisfactory solution, however, unless the number of persons failing to reach the item is small. Such a procedure would involve the use of a rapidly shrinking number of cases and would thus render the results on the later items quite unreliable. Moreover, the persons on whom the later items are analyzed would probably constitute a selected sample and hence would not be comparable to the larger samples used for the earlier items. As has already been pointed out, the faster performers tend also to be the more proficient. The later items would thus be analyzed on a superior sample of individuals. One effect of such a selective factor would be to lower the apparent difficulty level of the later items, since the percentage passing would be greater in the selected superior group than in the entire sample. It will be noted that this is the opposite error from that introduced when the percentage passing is computed in terms of the entire sample. In that case, the apparent difficulty of items is spuriously raised.

The effect of the above procedure on indices of item discrimination is less obvious, but nonetheless real. It has been observed, for example, that some low-scoring test takers tend to hurry through the test, marking items almost at random in their effort to try all items within the time allowed. This tendency is much less common among high-scoring test takers. As a result, the sample on which a late-appearing item is analyzed is likely to consist of some very poor respondents, who will perform no better than chance on the item, and a larger number of very proficient and fast respondents, who are likely to answer the item correctly. In such a group, the item-criterion correlation will probably be higher than it would be in a more representative sample. In the absence of such random respondents, on the other hand, the sample on which the later items are analyzed will cover a relatively narrow range of ability. Under these conditions, the discrimination indices of the later items will tend to be lower than they would be if computed on the entire unselected sample.

The anticipated effects of speed on indices of item difficulty and item discrimination have been empirically verified, both when item statistics are computed with the entire sample (Wesman, 1949) and when they are computed with only those persons who attempt the item (Mollenkopf, 1950a). In the latter study, comparable groups of high school students were given two forms of a verbal test and two forms of a mathematics test. Each of the two forms contained the same items as the other, but items occurring early in one form were placed late in the other. Each form was administered with a short time limit (speed conditions) and with a very liberal time limit (power conditions). Various intercomparisons were thus possible between forms and timing conditions. The results clearly showed that the position of an item in the speed tests affected its indices of difficulty and discrimination. When the same item occurred later in a speeded test, it was passed by a greater percentage of those attempting it, and it yielded a higher item-criterion correlation.

The difficulties encountered in the item analysis of speeded tests are fundamentally similar to those discussed in chapter 4 in connection with the reliability of speeded tests. Various solutions, both empirical and statistical, have been developed for meeting these difficulties. One empirical solution is to administer the test with a long time limit to the group on which item analysis is to be carried out. This solution is satisfactory provided that speed itself is not an important aspect of the ability to be measured by the test. Apart from the technical problems presented by specific tests, however, it is well to keep in mind that item analysis data obtained with speeded tests are suspect and call for careful scrutiny.

CROSS-VALIDATION

Meaning of Cross-Validation. It is essential that test validity be computed on a different sample of persons from that on which the items were selected. This independent determination of the validity of the entire test is known as cross-validation. Any validity coefficient computed on the same sample that was used for item-selection purposes will capitalize on random sampling errors within that particular sample and will consequently be spuriously high. In fact, a high validity coefficient could result under such circumstances even when the test has no validity at all in predicting the particular criterion.

EXPLORATIONS IN ITEM DEVELOPMENT

The rapid expansion of computer utilization in the 1980s and 1990s, in combination with progress in cognitive psychology, stimulated extensive research on innovative approaches to item construction. Traditionally, item writing has been more an art than a science. Even under the best conditions, item writers are given instructions that specify little more than item form and content coverage. It is still common practice to rely on empirical pretesting of items to assess their difficulty level and discriminative power. Is there any way to predict these item statistics, before pretesting, simply from an analysis of the physical or sethantic properties of the stimuli? Better yet, can items be constructed so as to have the desired difficulty and discriminative values? Can systematic manipulation of stimulus characteristics predetermine the cognitive demands of test items? These are the questions that are being investigated in ongoing research, through both experimental and mathematical procedures (Bejar, 1985, 1991; Carroll, 1987; Embretson, 1985a, 1985b, 1991, 1994,1995; Freedle, 1990).

The cognitive demands of test stimuli can be explored through the techniques of task decomposition developed within cognitive psychology. By these procedures, the relationships of different item features to speed and error of performance can be investigated. Several such studies have been conducted with spatial items (Embretson, 1994; Pellegrino, Mumaw, 6k Shute, 1985). For example, the stimuli presented in a spatial analogies test can be classified with respect to: (1) complexity or number of separate elements that must be identified (e.g., shape, size, position); and (2) transformations, or number of ways the stimulus is altered within the pair to be evaluated. In certain types of spatial visualization problems, which require the test taker to choose the parts that can be assembled I to form a given whole, the parts may be merely separated, or displaced, or rotated, or altered in a combination of these ways.

Other studies have been concerned with the semantic characteristics of verbal stimuli. For example, in verbal reasoning tests, items can be constructed according to known logical principled (Colberg, 1985; Colberg, Nester, 6k Trattner, 1985; Scheuneman, Geritz, 6k Embretson, 1991; K. Sheehan 6k Mislevy, 1989; Shye, 1988). Such a procedure could ensure that only one of the response options is truly correct and that different logical relations are represented in a predetermined proportion in the item sample. This procedure would also make possible the manipulation of the logical complexity of the item, whose relation to difficulty level can then be empirically investigated. Some researchers have experimented with the construction of letter series designed to test inductive reasoning (Butterfield et al., 1985). A detailed set of rules was first developed for the systematic construction of such letter series. Hypotheses were then formulated about what people do in trying to understand a series. The hypotheses were tested through empirical studies of the difficulty of series completion items.

Embretson (1994) presents a thoroughgoing analysis and updating of the process of item development. This process begins with a definition of the constructs to be assessed and proceeds to the design of a cognitive model for the test. The detailed features of this cognitive model provide the specifications for item writing. Empirical validation of items follows to ascertain how well the items actually fit the cognitive model in its practical applications. The complete procedure is illustrated in the development of the Spatial Learning Ability Test, which measures not only initial spatial ability but also its modifiability following standardized instruction.

Research on the prediction of item difficulty from the physical and semantic properties of the stimuli not only facilitates the production of effective tests by item writers but may also lead to the construction of items by computers. It is certainly possible to incorporate detailed item specifications in computer programs (see, e.g., Butterfield et al., 1985; Efhbretson, 1994). Undoubtedly, the potential advantages of these evolving test construction procedures are impressive. We must, however, guard against expecting too much from any one approach. It is quite likely, for example, that a test may measure some clearly identified cognitive constructs fully and effectively and yet not have high predictive validity for certain important practical uses. For this reason it is essential to consider both aspects of construct validation, which Embretson (1983) designates as construct representation and nomothetic span. Task decomposition provides information on construct representation; nomothetic span requires the investigation of the relations of test scores to a network of other, external variables, including criterion measures. A second caution against overgeneralization pertains to the need for relevant content knowledge in order to perform effectively in any subject-matter area or field of expertise. Processes are often linked to content; they cannot be effectively evaluated in the absence of the appropriate content.

In conclusion, the innovative procedures cited in this section, when properly applied, can contribute significantly to the systematic and controlled construction of test items. Moreover, by identifying the constructs measured by a test, these procedures can greatly enhance our understanding of the reasons why particular tests predict performance in criterion situations. A related benefit pertains to the diagnostic use of tests, insofar as the source of the individual's strengths and weaknesses can be linked to particular cognitive processes. These are worthy goals, but their practical implementation still requires considerable research on remaining unsolved problems (see, e.g., Wainer, 1993a). Much research is now in progress on the development of items that permit identification of the cognitive processes employed by individual respondents in solving particular items (Will-son, 1994). Analysis of the type of errors made by individuals provides promising leads for this purpose (Kulikowich 6k Alexander, 1994).

Lesson 05

RELIABILITY: THE CONSISTENCY OF TEST SCORES

Neither physical measurements nor psychological tests are completely consistent: if some attribute of a person is measured twice, the two scores are likely to differ. For example, if a person's height is measured twice in the same day. a value of may be obtained the first time, and a value of 5'10Vih" the second. A person Hiking the Scholastic Aptitude Test (SAT) twice might obtain a score of 1060 in the fall and 990 in the spring. A person taking two different forms of a test of general intelligence might obtain an IQ of 110 on one test and 114 on the other. Thus, test scores and other measures generally show some inconsistency. On the other hand, most test scores are not completely random. For example, one is unlikely to obtain a reading of 98.6° using one fever thermometer and a reading of 103° using another. Similarly, a child is unlikely to obtain a score of 95 percent on a well-constructed reading test and a score of 40 percent on a second, similar test. Methods of studying, defining, and estimating the consistency or inconsistency of test scores form the central focus of research and theory dealing with the reliability of test scores.

As stated in Chapter 3, measurement is the process of assigning numbers to persons so that some attribute of each person is accurately reflected in some attribute of the numbers. The reliability or consistency of test scores is critically important in determining whether a test can provide good measurement. For example, suppose you take a test of spatial visualization ability that provides a different score every time you take the test— sometimes your scores are very high, sometimes low and sometimes moderate. Since spatial visualization ability is a fairly stable characteristic, this test cannot possibly be a reliable measure—the scores vary substantially, even though the attribute the test is designed to measure does not. In other words, in this case the numbers do not reflect the attribute they are being used to measure.

The practical importance of consistency in test scores is a direct result of the fact that tests are used to make important decisions about people. For example, many high 1 schools require that students pass a competency test before they are allowed to graduate. Imagine what would happen if test scores were so inconsistent that many students who received low scores on one form of the test received high scores on another form of the same test. The decision to either grant or withhold a diploma might depend more on which form of the test the student took than his or her mastery of the high school curriculum.

This chapter presents a basic definition of test reliability and describes methods used in estimating the reliability or consistency of test scores. First, we describe sources of consistency and inconsistency in test scores. Next, we present a short outline of the theory of test reliability. Finally, we discuss methods of estimating the reliability of test scores, particularly as they relate to the different sources of consistency or inconsistency described in the first section of this chapter. Questions regarding the use of information about test reliability and factors affecting reliability are discussed in Chapter 6.

Sources of Consistency and Inconsistency in Test Scores

In understanding factors that affect the consistency of test scores, it is useful to ask. "Why do test scores vary at all?" For example, if I give a spelling test to a group of fourth graders, what factors are likely to lead to variability in test scores? Thorndike (1949) has prepared a list of possible sources of variability in scores on a particular test. This list, presented in Table 5.1, is a useful starting place for understanding factors that may affect the consistency or inconsistency of test scores.

The first category of factors that affect test scores lists some lasting and general characteristics of individuals. For example, we would expect some children to do consistently better than others on a spelling test because they are good spellers or because they are skillful in following instructions and in taking tests. The second category lists lasting but specific characteristics of individuals. For example, some children who are generally poor spellers might nevertheless know how to spell many of the particular words included in the test. If these children were given another test, made up of different words, they might receive very different scores. The third category lists temporary but general characteristics of individuals. For example, a child who is ill or very tired might do poorly this time around but might receive much higher scores if he or she is tested when healthy and well rested. The fourth category in the table lists temporary and specific characteristics of individuals. For example, the test may contain the words Baltimore Milwaukee and Seattle. A child who took the test shortly after looking at the sports section of the newspaper might have a temporary advantage on such a test. The fifth category lists some aspects of the testing situation that could lead to inconsistency in test scores. For example, if half the class took the test in a noisy, poorly lit room, we might expect their scores to be lower than they would have obtained under normal conditions. Finally, the sixth category in the table lists some chance factors that may affect test scores. Some of the variability in scores will be due strictly to luck.

Table 5.1 Possible Sources of Variability in Scores on a Particular Test

I. Lasting and general characteristics of the individual
A. Level of ability on one or more general traits, which operate in a number of tests
B. General skills and techniques of taking tests ("test-wiseness" or "test naivete"]
C. General ability to comprehend instructions
II. Lasting but specific characteristics of the individual
A. Specific to the test as a whole
Individual level of ability on traits required in this test but not in others
Knowledge and skills specific to particular form of test items
Stable response sets (e.g. to mark A options more frequently than other options of multiple-choice items, to mark true-false items "true" when undecided)
B. Specific to particular test items
The "chance" element determining whether the individual does or does not know a particular fact
Item types with which various examinees are unequally familiar (cf. item IIA2 above)
III. Temporary but general characteristics of the individual (factors affecting performance on many or all tests at a particular time)
A. Health
B. Fatigue
C. Motivation
D. Emotional strain
E. "Test-wiseness" (partly lasting; cf. Item I.B. above)
F. Understanding of mechanics of testing
G. External conditions of heat, light, ventilation, etc.
IV. Temporary and specific characteristics of the individual
A. Specific to a test as a whole
Comprehension of the specific test task
Specific tricks or techniques of dealing with the particular test materials
Level of practice on the specific skills Involved (especially in psychomotor tests)
Momentary "set" for a particular test
B. Specific to particular test items
Fluctuations and idiosyncrasies of human memory
Unpredictable fluctuations in attention or accuracy, superimposed upon the general level of performance characteristic of the individual
V. Systematic or change factors affecting the administration of the test or the appraisal of test performance
A. Conditions of testing adherence to time limits, freedom from distractions, clarity of instructions, etc.
B. Interaction of personality, sex, or race of examiner with that of examinee that facilitates or inhibits performance
C. Unreliability or bias in grading or rating performance
VI. Variance not otherwise accounted for (chance)
A. Luck in selection of answers by sheer guessing
B. Momentary distraction

In most testing applications, we are interested in lasting and general characteristics of persons, such as spelling ability. Thus, in most cases the first category in Table 5-1 represents a source of consistency in test scores, and all of the remaining categories represent sources of unwanted inconsistency. However, this breakdown is not always quite so simple. For example, we might be interested in measuring a specific characteristic, such as a child's ability to spell Baltimore on May 30, 1993. In this case, we might conclude that elements from categories I. II, III, and IV would all contribute to consistency in measurement and that inconsistency in the child's performance on that particular item at the particular time the child is tested would largely be determined by elements from categories V and VI. The determination of whether each of the factors listed in the table contributes to consistency or inconsistency in measurement thus depends largely on what one is trying to measure. As will be seen in the sections that follow, the definition of reliability, as well as the methods used to estimate the reliability of test scores, ultimately depends on one's definition of precisely what attribute is being measured and of the sources of inconsistency in the measurement of that attribute.

GENERAL MODEL OF RELIABILITY

A perfect measure would consistently assign numbers to the attributes of persons according to some well-specified rule (e.g.. if John is more anxious than Teresa, he should receive a higher score on an anxiety test than she does). In practice, our measures are never perfectly consistent. Theories of test reliability have been developed to estimate the effects of inconsistency on the accuracy of psychological measurement. The basic starting point for almost all theories of test reliability is the idea that test scores reflect the influence of two sorts of factors:

1. factors that contribute to consistency: stable characteristics of the individual or the attribute one is trying to measure
2. factors that contribute to inconsistency: features of the individual or the situation that can affect test scores but have nothing to do with the attribute being measured

This conceptual breakdown is typically represented by the simple equation

Observed Test Score = True Score + Errors of Score Measurement

Or, more succinctly,

$$X = T + e$$

Where

X = score on the test
T = true score
e = error of measurement

There is an important distinction between the concept of true score as applied in Formula 5-2 and the notion of ultimate truth or perfect measurement. Thus, true scores on a measure of anxiety are not an indication of a person's "true" or "real" level of anxiety. Rather, the true score represents a combination of all the factors that lead to consistency in the measurement of anxiety (Cronbach, Glescr. Nanda & Rajaratnam, 1972; Stanley. 1971). As a result, the components that make up the true score part of a test will vary, depending upon what is being measured. Consider, for example, a test of mechanical aptitude given to a U.S. Air Force mechanic in a noisy, poorly lit room. As indicated in Table 5-2, the stressful conditions of measurement could be considered as either part of the true score or part of the error component, depending on whether the test is intended to measure mechanical aptitude in general or mechanical performance under stressful conditions.

Errors in measurement represent discrepancies between scores obtained on tests and the corresponding true scores. Thus,

$$e = X - T$$

The goal of reliability theory is to estimate errors in measurement and to suggest ways of improving tests so that errors are minimized.

The central assumption of reliability theory is that measurement errors are essentially random. This does not mean that errors arise from random or mysterious processes. On the contrary, a sizable negative error in a score a person received on the Graduate Record Examination (GRE) could easily be accounted for if it were known that the person (a) had stayed up all night, (b) had a hangover, (c) was sitting next to a noisy air conditioner during the test, and (d) used the wrong part of the form to mark the answers. For any individual, an error in measurement is not a completely random event. However, across a large number of individuals the causes of measurement error are assumed to be so varied and complex that measurement errors act as random variables. Thus, a theory that assumes that measurement errors are essentially random may provide a pretty good description of their effects.

If errors have the essential characteristics of random variables, then it is reasonable to assume that errors are equally likely to be positive or negative and that they are not correlated with true scores or with errors on other tests. That is, it is assumed that

Table 5.2 Different Definitions of True Score When an Aptitude Test Is Given Under Stressful Conditions

	Test is used to measure	
	Mechanical performance in general	**Mechanical performance under stressful conditions**
Individuals' Mechanical Aptitude	True Score	True Score
Stressful Conditions of Measurement	Error	True Score
All Other Irrelevant Sources of Variability	Error	Error

1. Mean error of measurement = 0.
2. True scores and errors are uncorrelated: $r_{Te} = 0$.
3. Errors on different measures are uncorrelated: r_{12}: = 0.

=

On the basis of these three assumptions, an extensive theory of test reliability has been developed (GuIIiksen. 1950; Lord & Novick. 1968). Several results can be derived from this theory that have important implications for measurement. For example Table 5-1 listed sources of variability in test scores that might contribute to the consistency or inconsistency of measurement. Reliability theory provides a similar breakdown by showing that the variance of obtained scores is simply the sum of the variance of true scores plus the variance of errors of measurement. That is,

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2$$

In effect Formula 5-4 suggests that test scores vary as the result of two factors: (a) variability in true scores and (b) variability due to errors of measurement. If errors are responsible for much of the variability observed in test scores, test scores will be inconsistent; if the test is given again, scores may not remain stable. On the other hand, if errors of measurement have little effect on test scores, the test reflects mainly those consistent aspects of performance we have labeled true score.

The reliability coefficient (rxx) provides an index of the relative influence of true and error scores on obtained test scores.1 In its general form, the reliability coefficient is defined as the ratio of true score variance to the total variance of test scores. That is

$$r_{xx} = \frac{\sigma_T^2}{\sigma_X^2}$$

or equivalently:

$$rxx = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}$$

There are several interpretations of the reliability coefficient. Perhaps the most straightforward interpretation is that rxx indicates the proportion of variance in test scores that is due to or accounted for by, variability in true scores.

Simple Methods of Estimating Reliability

The goal of estimating reliability is to determine how much of the variability in test scores is due to errors in measurement and how much is due to variability in true scores. The parallel test model suggests a strategy for accomplishing this goal. According to the parallel test model, it might be possible to develop two forms of a test that are equivalent in the sense that a person's true score on form A would be identical to his or her true score on form B. If both forms of the test were administered to a number of people, differences between scores on form A and form B could be due only to errors in measurement. Thus, if there were large differences between scores on the two forms, one would conclude that measurement errors were a major source of variability in test scores. On the other hand, if scores on both tests were highly similar, one would conclude that measurement errors were small and that the test was highly reliable.

The parallel test model provides a conceptual solution for estimating reliability but does not necessarily provide a practical solution. The reason for this is that strictly parallel tests are difficult to develop. Four practical strategies have been developed that incorporate many features of the parallel test method and that provide workable methods of estimating test reliability:

1. test-retest methods
2. alternate forms methods
3. split-half methods
4. internal consistency methods

The methods described here are most likely to be useful in situations where one is interested in measuring lasting, general characteristics of individuals, such as abilities or trails. In these situations, the attributes of the persons being measured represent the sole source of true score variance, and all other factors that affect measurement combine to represent the error component. As will be discussed in a later section, this sort of testing application is typical, but not universal. In many situations, the tester must identify explicitly the purpose and the uses of measurement in order to determine which factors affect true scores and error. In more complex testing applications, reliability theories that break a measure down into true score and random error components may not be sufficient: a more general and complex theory is presented later in this chapter. Nevertheless, for many testing applications, simple methods based upon classical reliability theory may be quite useful. Several of these methods are presented below.

Test-Retest Method

The test-retest method is one of the oldest and. at least at first glance, one of the most sensible methods of estimating the reliability of test scores. Reliability is concerned with the consistency of test scores; the test-retest method directly assesses the degree to which test scores are consistent from one test administration to the next The test-retest method involves (a) administering a test to a group of individuals, (b) readministering that same test to the same group al some later time, and (c) correlating the first set of scores with the second. The correlation between scores on the first test and scores on the retest is used to estimate the reliability of the test.

The rationale behind this method of estimating reliability is disarmingly simple. Since the same test is administered twice and every test is parallel with itself, differences between scores on the test and scores on the retest should be due solely to measurement error. This sort of argument is quite probably true for many physical measurements. For example, if a tape measure is used to measure the length of a room and is then used to measure the same room a week later, any difference between the two measurements is likely to be entirely due to measurement error. Unfortunately, this argument is often inappropriate for psychological measurement, since it is often impossible to consider the second administration of a test a parallel measure to the first. Thus it may be inaccurate to treat a test-retest correlation as a measure of reliability.

Limitations of Test Retest Method

There are several reasons why the second administration of a psychological test might yield systematically different scores than the first administration firstly; the characteristic or attribute that is being measured may change between the first test and the retest. Consider, for example, a reading test that is administered in September to a class of third graders and then readministered in June. We would expect some change in children's reading ability over that span of time: a low test-retest, correlation might reflect real changes in the attribute measured by the test.

Second one experience of taking the test itself can change a person's true score; this is referred to as reactivity for example, students who take a geography test may look up answers they were unsure of after taking the test, thus changing their true knowledge of geography. Likewise, the process of completing an anxiety inventory could serve to increase a person's level of anxiety. Thus, the first test may serve as a catalyst that causes substantial change in true scores from one administration to the next.

Third one must be concerned with carry-over effects, particularly if the interval between test and retest is short When retested, people may remember their original answers, which could affect their answers the second time around.

In addition to the theoretical problems inherent in the test-retest method, there is a practical limitation to this method of estimating reliability. The test-retest method requires two test administrations. Since testing can be time consuming and expensive, retesting solely for the purpose of estimating reliability may be impractical. It is common to distinguish between reliability, which is the ratio of true to observed variance, and temporal stability, which refers to the consistency of test scores over time. If true scores are likely to change over time, this distinction is both theoretically and practically important. Thus, test-retest correlations are often thought of as stability coefficients rather than reliability coefficients. However, even when this distinction is drawn, there are problems with the test-retest technique. For example it is not clear whether carry-over effects should be regarded as sources of measurement error or as sources of real stability (or instability) in measurement. Nor is it clear whether reactivity effects always contribute to true scores or to error./ It could be argued that carry-over and reactivity effects are a natural aspect of the attribute being measured and should be taken into account in estimating stability. On the other hand, these effects may inflate (or deflate) one's estimate of the true stability of test scores, thus yielding inaccurate estimates of both reliability and stability.

The test-retest method is most useful when one is interested in the long-term stability of a measure. For example, research on the accuracy of personnel selection tests is concerned with the ability of the test to predict long-term job performance. Research in this area has therefore focused on the temporal stability of job performance measures (Schmidt & Hunter, 1977). The argument here is that short-term variations in job performance represent error when the-purpose of the research is to predict performance over the long term.

Alternate Forms Method

(The alternate forms method of estimating reliability is, on the surface, the closest approximation to the method suggested by the parallel tests model. The key to this method is the development of alternate test forms that are, to the highest degree possible, equivalent in terms of content, response processes and statistical Characteristics, For example, alternate forms exist for several tests of general intelligence; these alternate forms commonly are regarded as equivalent tests'. The alternate forms method of estimating test reliability involves (a) administering one form of the test (e.g., form A) to a group of individuals; (b) at some later time, administering an alternate form of the same test (e.g., form B) to the same group of individuals; and (c) correlating scores on form A with scores on form B. The correlation between scores on the two alternate forms is used to estimate the reliability of the test?

The alternate forms method provides a partial solution to many of the problems inherent in test-retest methods. For example, since the two forms of the test are different carry-over effects are less of a problem, although /remembering answers previously given to a similar set of questions may affect responses to the present set of questions] Reactivity effects are also partially controlled; although taking the first test may change responses to the second test, it is reasonable to assume that the effect will not be as strong with alternate forms as with two administrations of the same test. The alternate forms method also may allow one to partially control for real changes over time in the attribute being measured? Since carry-over effects are less of a problem here than in the test-retest method! it may not be necessary*to use a long interval between test administrations/ It is feasible to administer the second form immediately after the first, which cannot be done in most test-retest studies.

Although the alternate forms method avoids many of the problems inherent in the test-retest method, there are still many drawbacks to this technique. For example, since two separate test administrations are required, the alternate forms method could be as expensive and impractical as the test-retest method! Second/ it may be difficult and expensive to develop several alternate forms of a test! It is questionable that the expense involved in developing alternate forms and in administering two tests rather than one is justified solely for the purpose of assessing the reliability of the test. In addition/it may be difficult, if not impossible, to guarantee that two alternate forms of a test are, in fact, parallel measures Thus if alternate forms are poorly constructed, one might obtain low reliability estimates strictly as a function of the lack of equivalence between the two alternate form's.

Split-Half Methods

Split-half methods of estimating reliability provide a simple solution to the two practical problems that plague the alternate forms method: (a) the difficulty in developing alternate forms and (b) the need for two separate test administrations. The reasoning behind split-half methods is quite straightforward. The simplest way to create two alternate forms of a test is to split the existing test in half and use the two halves as alternate forms. The split-half method of estimating reliability thus involves (a) administering a test to a group of individuals, (b) splitting the test in half, and (c) correlating scores on one half of the test with scores on the other half. The correlation between these two split halves is used in estimating the reliability of the test.3

The split-half method avoids many of the theoretical and practical problems inherent in test-retest and alternate forms methods. First this method allows reliability to be estimated without administering two different tests or administering the same test twice.] Thus, whenever a multi-item test is administered; the split-half method could be used to estimate reliability. Since there is only one test administration carry-over effects, reactivity effects, and especially the effects of change over time on true scores are minimized. Inconsistencies in scores obtained on two different halves of a test are therefore likely to reflect inconsistencies in responses to the test itself) rather than changes in the individual that may have occurred between administrations of two alternate forms, or between a test and retest.

There are several ways of splitting a test to estimate reliability. For example, a 40-itcni vocabulary test could be split into two subtests, the first one made up of items 1 through 20 and the second make up of items 21 through 40. One might suspect, however, that responses to the first half would be systematically different from responses to the second half, so that this split would not provide a very good estimate of reliability. For example, many vocabulary tests start with the easiest items, and then become progressively more difficult. In addition, a person taking the test might be more fatigued during the second half of the test. In splitting a test, the two halves would need to be as similar as possible, both in terms of their content and in terms of the probable state of the respondent. The simplest method of approximating this goal is to adopt an odd-even split, in which the odd-numbered items form one half of the test, and the even-numbered items form the other. This guarantees that each half will contain an equal number of items from the beginning, middle, and end of the original test.

The fact that there are many ways a test could potentially be split is the greatest weakness of the split-half method. Consider, for example, the six-item test shown in Table. There are ten different ways that this could be split,4 and each split yields a somewhat different reliability estimate^ In other words, the correlation between half A and half B may vary, depending on how the test is split. The question is which is the reliability of the test? Although the idea of forming an odd-even split generally make sense, this particular method of splitting the test is somewhat arbitrary; here is no guarantee that the reliability coefficient obtained from this particular split will be the most reasonable or accurate estimate of the relative contribution of true scores and errors of measurement to the variability of test scores. Since it is difficult to make a strong argument in favor of one particular method of splitting a given test over another, it is difficult to decide which split-half reliability estimate to use.

Table 5.3 Ten Possible Ways of Splitting A Six-Item Test

Test items, half A	Test items, half B	Reliability estimate
1, 2, 3	4, 5, 6	.64
1, 2, 4	3, 5, 6	.68
1, 2, 5	3, 4, 6	.82
1, 2, 6	3, 4, 5	.79
1, 3, 4	2, 5, 6	.88

1, 4, 5	2, 3, 6	.81
1, 5, 6	2, 3, 4	.82
2, 3, 5	1, 4, 6	.72
2, 4, 5	1, 3, 6	.71
2, 4, 6	1, 3, 5	.74

Internal Consistency Methods

Internal consistency methods of estimating test reliability appear to be quite different from the methods presented so far (internal consistency methods estimate the reliability' of a test based solely upon the number of items in the test (k) and the average inter correlation among test items (r.Y) These two factors can be combined in the following formula to estimate the reliability of the test:

$$r_{xx} = \frac{k \; (\overline{r}_{ij})}{1 + (k - 1)\, \overline{r}_{ij}}$$

Thus, (the internal consistency method involves (a) administering a test to a group of individuals, (b) computing the correlations among all items and computing the average of those inter correlations, and (c) using Formula 5-7. or an equivalent formula, to estimate reliability.5 This formula gives a standardized estimate; raw score formulas that take into account the variance of different test items may provide slightly different estimates of internal consistency reliability.

There are both mathematical and conceptual ways of demonstrating the links between internal consistency methods and the methods of estimating reliability discussed so far. First, internal consistency methods are mathematically linked to the split-half method^ In particular, coefficient alpha, which represents the most widely used and most general form of internal consistency estimate, represents the mean reliability coefficient one would obtain from all possible split halves [Cortina (In press) notes that alpha is equal to the mean of the split halves defined by formulas from Rulon (1939) and Flanagan (1937)]. In other words/if every possible split-half reliability coefficient for a 30-item test were computed, the average of those reliabilities would be equal to coefficient alpha. The difference between the split-half method and the internal consistency method is, for the most part, a difference in unit of analysis. Split-half methods compare one half-test to another; internal consistency estimates compare each item to every other item.

In understanding the link between internal consistency and the general concept of reliability, it is useful to note that internal consistency methods suggest a fairly simple answer to the question, "Why is a test reliable?" Remember that/internal consistency estimates are a function of (a) the number of test items and (b) the average inter correlation among those test items. If we think of each test item as an observation of behavior, internal consistency estimates suggest that reliability is a function of (a) the number of observations one makes and (b) the extent to which each item represents an observation of the same thing observed by other test items. For example, if you wanted to determine how good a bowler someone was, you would obtain more reliable information by observing the person bowl many frames than you would by watching the person roll the ball once. You would also obtain a more reliable estimate of that person's skill at bowling from ten observations of the person bowling in typical circumstances than if you watched him bowl three times on a normal alley, then watched him bowl live times in a high-pressure tournament, and then watched him bowl twice on a warped alley/if every item on the test measures essentially the same thing as all other items, and if the number of items is large, internal consistency methods suggest that the test will be reliable.

Reliability Estimates and Error

Each of the four methods of estimating test reliability implies that different sources of variability in test scores might contribute to errors of measurement. Both the split-half and the internal consistency methods define measurement error strictly in terms of consistency or inconsistency in the content of a test. Test-retest and alternate forms methods, both of which require two test administrations, define measurement error in terms of three general factors: (a) the consistency or inconsistency of test content (in the test-retest method, content is always consistent); (b) changes in examinees over time; and (c) the effects of the first test on responses to the second test. Thus, while each method is concerned with reliability, each defines true score and error in a somewhat different fashion.

The principal advantage of internal consistency methods is their practicality. Since only one test administration is required, it is possible to estimate internal consistency reliability every time the test is given. Although split-half methods can be computationally simpler the widespread availability of computers makes it easy to compute coefficient alpha, regardless of the test length or the number of examinees' therefore is _ possible to compute coefficient alpha whenever a test is used in a new situation or population low value for alpha could indicate that scores on the test will not be highly consistent and therefore may form a poor basis for describing the person or for making decisions about the person . However, as noted in the next section, an internal consistency estimate does not necessarily represent the reliability of a test) Indeed, test reliability often depends more on what one is trying to do with test scores than on the scores themselves.

Table 5.4 Sources of Variability That Contribute To Errors in Measurement

		Method of Estimating Reliability		
	Test-retest	**Alternate forms**	**Split-half**	**Internal consistency**
Content Factors		Inconsistency of Test Content Nonparallel Tests	Inconsistency of Test Content Nonparallel Halves	Inconsistency of Test Content
Effects of First Test on Examinee	Reactivity Carry-over	Reactivity Carry-over		
Temporal factors	True Change over Time	True Change over Time		

The Standard Error of Measurement

The reliability coefficient provides a relative measure of the accuracy of test scores'! Thus, a test with a reliability of .90 is more reliable than a test with a reliability of .80. However, the reliability coefficient docs not provide an indication in absolute terms, of how accurate test scores really are. For example, suppose a psychologist measured a child's intelligence and obtained a score of 110. Could we be confident that the score obtained was really higher than the average score expected on the test (100) or would we expect test scores to vary more than 10 points simply because of measurement error? Even if we knew that the test was highly reliable (e.g., a reliability coefficient of (.93), we would not be able to answer this question. The reliability coefficient simply does not reveal, in concrete terms, how much variability should be expected on the basis of measurement error. In order to describe the accuracy of test scores concretely, we need to know more than the reliability of the test; we must also know the size of the standard error of measurement. The standard error of measurement (SEM) is a function of two factors: (a) the reliability of the test (r_{xx}) and (b) the variability of test scores (σ_x). Thus, the standard error of measurement is given by _

$$\text{SEM} = \sigma_x \surd 1 - r_{xx}$$

The standard error of measurement provides a measure of the variability in test scores expected on the basis of measurement errors. For example, imagine the woman with an IQ of 100 was tested repeatedly with a standard intelligence test. Since tests are not perfectly reliable, she would not receive a score of 100 every time; sometimes she would receive a score higher than 100, and sometimes she would receive a lower score. The total distribution of her test scores might look something like the one shown in Figure 6-1. The standard error of measurement corresponds to the standard deviation of this distribution. In other words, the standard error of measurement indicates how much variability in test scores can be expected as a result of measurement error.

The standard error of measurement can be used to form confidence intervals, which in turn provide a concrete indication of how accurate test scores really are. For example

in turn provide a concrete indication of how accurate test scores really are. For example

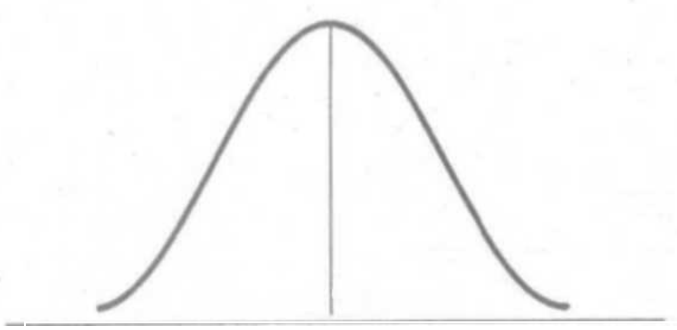


Figure 6-1 Distribution of Scores

85 90 95 100 105 110 115 Obtained by Repeatedly
Testing a Standard Error of Measurement ■ 4.7 Woman with an IQ of 100

Reliability and Validity

Psychological tests are never perfectly consistent or reliable. Our failure to achieve perfect reliability in psychological measurement has implications for the validity of tests, both for the validity of measurement and for the validity of predictions (and decisions) based upon test scores. In a nutshell, lack of reliability places a limit on the validity of inferences drawn from test scores. Thus, a test that is not perfectly reliable cannot be perfectly valid, either as a means of measuring attributes of a person or as a means of predicting scores on a criterion. The rationale here is simple. A valid test is one that consistently assigns to persons numbers that reflect their standing on a particular attribute. An unreliable test, which provides inconsistent scores, cannot possibly be valid.

It is important to note that tests that are reliable are not necessarily valid. For example, consider the five-item "test of general intelligence" shown in Table 5.5. You would expect scores on this test to be almost perfectly consistent. For example, if a person took this test in January and then again in June, he or she would almost certainly give the same answers. However, we would not accept this test as a valid measure of intelligence, nor would we expect this test to provide accurate predictions of scholastic performance. Thus, consistency in test scores is no guarantee of validity in measurement or prediction.

Table 5.5 A "Test of General Intelligence" That Would Be Highly Reliable But Not At All Valid

1. In what month were you born? ___________
2. What is your mother's first name? _____________
3. 1 + 1 = _________

A. How many days are there in a week? _____________
5. Which of the following is a triangle? __________
a) □ b) △ c) O

Basically, there are four factors that affect the reliability of a test:

1. Characteristics of the people taking the test
2. Characteristics of the test itself
3. The intended uses of test scores
4. The method used to estimate reliability

Characteristics of the people taking the test. The first factor that affects the reliability of psychological measurement is the extent to which people taking the test vary on the characteristic or attribute being measured. As stated earlier, tests are designed to measure individual differences. If individuals do not differ very much on a particular attribute or trait, it is difficult to develop a reliable measure of that attribute. For example, imagine that you are a first-grade teacher trying to estimate individual differences in height of your students as they are seated at their desks. Since children at this age tend to be quite similar in height, this task would be quite difficult. On the other hand, a high school teacher would find this task much easier, since at that age there is a greater variation in students' heights. Thus, when individual differences are quite large, it is much easier to measure them.

The standard deviation provides a measure of variability, or the extent to which individuals differ. Thus, the larger the standard deviation, the more likely it is that a test will be reliable. However, a large standard deviation-also implies that measures will, in an absolute sense, be inaccurate. Recall that the formula for the standard error of measurement (SEM) is

$$\text{SEM} = \sigma_x\sqrt{}1 - r_{xx}$$

If individual differences are very small, test reliability probably will be low. However, the standard deviation also will be small, and thus the standard error of measurement will be small. Hence, it might be possible to have a test that is unreliable as a measure of individual differences but nevertheless provides an accurate measure of each person's standing on the attribute measured. The clearest example of this phenomenon would occur if everyone received the same score on a test. For example, if a group of physicists took a third-grade mathematics test, each of them would probably receive a score of 100. Since there are no individual differences in test scores, the test would have a reliability coefficient of 0.0; however, the standard error of measurement would also be 0.0. Thus, the test would be perfectly unreliable and perfectly' accurate at the same time.

Because the variability of the attribute being measured affects test reliability, it is likely that many tests will show different levels of reliability, depending on the population in which they are used. For example, a test of general intelligence will be more reliable for the population in general than for a population of graduates of highly select colleges. It follows, then, that it will often be difficult to cite any figure\is the reliability of a test. Tests will be more reliable in settings when individual differences are extreme and less reliable in settings where individual differences are small.

Characteristics of the Test. Internal consistency formulas for estimating reliability suggest that two factors affect the reliability coefficient: (a) the correlations between items and (b) the number of items. This definition suggests that the reliability of a test can be increased in either of two ways: by increasing the correlations between items or by increasing the number of items. Thus, a test made up of 40 mathematics questions is likely to be more reliable than a 40-item test that covers mathematics, spatial visualization, baseball trivia, and knowledge of French grammar; and an SO-item mathematics test is likely to be more reliable than a similar 40-item mathematics test.

Since most well-developed tests are designed to measure one attribute or trait, the first strategy that of increasing the homogeneity or consistency of the test items is not always easy to implement. Therefore, the typical strategy for increasing the reliability of a test is to lengthen the test. Psychometric theory provides the Spearman-Brown formula, which can be used to predict the effect of lengthening a test on the reliability of that test. If the length is increased by a factor of (e.g., double the length, or n = 2), the reliability of the new, expanded test is estimated by

$$\text{new } r_{xx} = \frac{n \text{ x old } r_{xx}}{1 + (n - 1) \text{ old } r_{xx}}$$

where

old r_{xx} = reliability of the original test
new r_{xx} = reliability of the lengthened test

Table 6-4 illustrates the effects of lengthening a 20-item test with a reliability of .60. As shown by the table, it would be possible to substantially increase the reliability of the test by greatly increasing the number of items. Thus, adding 100 more items would raise the reliability from .60 to .90. It may, however, be exceedingly difficult and prohibitively expensive to add 100 extra items.

Formula 6-5 can also be used to predict the effect of decreasing the length of a test. For example, a 200-item test might he highly reliable (e.g., rl(= .95), but it also might be too long for many applications. A short form of the same test might show acceptable levels of reliability and be much more practical. For example, a 50-item short form of this test would have an estimated reliability coefficient of .80.

In principle, the reliability of most psychological tests is under the control of the test developer. Internal consistency formulas imply that if a test is sufficiently long, it will be reliable, even if the average inter-item correlation is fairly small. For example, a 50-itcm test in which the average inter-item correlation is .20 would have a reliability of .92. A 100-item test in which the average inter- item correlation is as small as .05 still would have a reliability of .84. Thus, in theory, there is really no excuse for an unreliable test. On the other hand a large coefficient alpha is not by itself an indication of a good test if the number of items is large. As Cortina (In press) notes, even a poorly developed test will be reliable if it is sufficiently long. However, though it is theoretically possible to achieve a high level of reliability with any test simply by increasing its length, practical barriers may prevent implementing such a strategy.

Table 5.6 Effects on Reliability of Lengthening a 20-Item Test with a Reliability Coefficient of .60

Test length	n	Estimated reliability
40 items	2	.75
60	3	.81
80	4	.85
100	5	.88
120	6	.90

First, long tests are more time consuming and expensive than short tests. The increase in reliability may not be worth the price, especially if tests are not used to make final decisions about people. Second, it often is quite difficult to write good test items. In addition, the new items added to the test must be substantively and statistically similar to the items already on the test; otherwise, the Spearman-Brown formula will not provide an accurate reliability estimate. Thus, in practice it may be quite laborious to lengthen many tests substantially.

Intended Use of Test Scores. In Chapter 5 we noted that tests might have different levels of reliability for different purposes. For example, suppose that school children take tests of general intelligence at a very early age (e.g., in first grade) and take the same test again many years later (e.g., on entering high school). The test might be highly reliable each time it is given but might show much less temporal stability. Thus, the test would be a more reliable indicator of the children's general mental ability at the time they took each test than it would be as an indicator of their long-term level of intelligence.

In general, the reliability of test scores is related to the breadth (versus narrowness) of the inferences that are made. For example, suppose a person takes a computer-presented test of spatial visualization that involves mentally rotating three-dimensional geometric shapes. This test would provide a more reliable measure of the person's present level of spatial visualization ability than of his or her ability ten years down the road. Test scores could probably be better generalized to other computer-presented tests than to similar paper-and-pencil tests. Finally, test scores might generalize better to other tests involving mental rotation than to tests that involve other spatial visualization tasks.

In sum tests often show different levels of reliability for identifying individual differences at a specific point in time than for identifying individual differences across time. Tests are more reliable for identifying gross distinctions between individuals than they are for very fine distinctions. For example, one would expect a high degree of reliability for a mastery test (scored cither pass or fail) of fifth-grade mathematics given to a sample made up of 50 second graders and 50 college sophomores. Almost everyone in the first group is likely to receive failing scores, and almost everyone in the second group is likely to receive passing scores. On the other hand, it might be difficult to reliably rank-order 100 Ph.D. candidates in terms of their general intelligence; for although there would be some standouts, this group would tend to be homogeneous. Finally, it is easier to make reliable inferences about stable characteristics of individuals than about characteristics that vary unpredictably. For example, it would be easier to develop a reliable measure of a person's basic values than of a person's mood state.

Methods Used to Estimate Reliability. Test-retest, alternate forms split-half, and internal consistency methods of estimating test reliability each imply slightly different definitions of true score and error. For example, changes over time in the attribute being measured are considered sources of measurement error in estimating test-retest reliability. When split-half or internal consistency estimates are employed, the temporal stability of the attribute being measured is riot a relevant factor.

In general, one might expect internal consistency estimates to be higher than alternate forms correlations, which in turn should probably be higher than test-retest reliability estimates. The reason for this is that more factors contribute to measurement error when test-retest or alternate forms methods are used than when internal consistency methods are used. For example, temporal instability and reactivity affect test-retest estimates but have no effect on internal consistency estimates. There are situations, however, where test-retest reliability can be expected to be higher Hum internal consistency estimates. For example, if people remembered all their responses to the first test, it would be possible to obtain perfect test-retest reliability, even if the internal consistency estimate was exactly 0.0.

The method used in estimating reliability should correspond with the way in which test scores are used. For example, if test scores obtained on one occasion are used to make inferences across long periods of time, it is important to estimate me temporal stability of test scores. If a clinic employs several different psychologists to score responses to the Rorschach inkblot test, it may be important to determine the extent to which test scores can be generalized across psychologists. There is no single figure that represents the reliability of a test, so the choice of an appropriate method for estimating and defining reliability is potentially very important.

Special Issues in Reliability

Psychometric theory typically has little to say about the content of a test; a reliability coefficient of .90 is generally interpreted in the same way, regardless of what the test measures or what types of items are included in the test. However, the proper methods of estimating and interpreting test reliability may depend on what the test measures. In particular, some ability and achievement tests measure speed of response, whereas others measure performance irrespective of speed. This distinction has important implications for the estimation and interpretation of test reliability.

A second issue of importance is the reliability of difference scores, or gain scores. In assessing training and educational programs, it is common practice to give a pretest before training, and a post-test after training. The difference between pretest and post-test scores is then used to measure the gain associated with training. Difference scores of this sort may present some special problems in terms of their reliability, as is discussed later in this section.

How Reliable Should Tests Be?

All things being equal, a highly reliable test would always be preferable to a test with little reliability. However, all things are rarely equal; the most reliable test might also be the most expensive or most difficult to administer. Test reliability may be crucial in some settings (e.g., those in which major decisions are made on the basis of tests) but less important in others (e.g. where tests are used only for preliminary screenings). It is impossible to specify any particular figure as the minimum level of reliability needed for all testing applications, but rough guidelines can be established for some of the more common uses of tests.

High levels of reliability are most necessary when (a) tests are used to make final decisions about people and (b) individuals are sorted into many different categories based upon relatively small individual differences. For example, tests are used for placement in the armed forces—individuals can be assigned to any of a number of jobs based largely on their scores in the Armed Services Vocational Aptitude Battery. In this case, measurement error could have a significant effect on decisions (Murphy. 19S4a). If the tests used in placement are unreliable, the decisions made regarding thousands of recruits also will be unreliable.

Lower levels of reliability are acceptable when (a) tests are used for preliminary rather than final decisions and (b) tests are used to sort people into a small number of categories based upon gross individual differences. For example, if several hundred people apply for a small number of places in a graduate program in clinical psychology, test scores might be used for preliminary screening, in which applicants who essentially have no chance of being admitted (e.g.. the bottom 25 percent of the applicant pool) are screened out. If a test were used in this way a high level of reliability would be desirable, but not essential.

Lesson 06

VALIDITY

Validity of Measurement: Content and Construct-Oriented Validation Strategies

Two of the principal problems in psychological measurement are determining whether a test measures what it is supposed to measure and determining whether that test can be used in making accurate decisions. Suppose a psychologist devises a test and claims that it measures reading comprehension and can be used to predict success in graduate school. These claims would not carry much weight unless they were supported by evidence. Hence, the psychologist must present data to show that the claims are accurate, or valid. For example, if test scores are correlated with grades or with professors' evaluations of graduate students, it is reasonable to conclude that the test is a valid predictor of success in graduate school. If test scores are related to other measures of reading comprehension, or if the test provides a representative sample of tasks that measure reading comprehension, then the test probably does indeed measure what it purports to measure.

This chapter opens our discussion of validity. There are many ways in which tests can be used: as a consequence, there are many ways of defining validity. We begin this chapter by discussing in more detail the two major types of validity: (a) the validity of measurement and (b) the validity for decisions. The validity of measurement is the main locus of this chapter. Ways of defining and estimating the validity of decisions are discussed in the chapter that follows.

Validation Strategies

In the 1940s and the early 1950s, research on psychological measurement was characterized by a bewildering array of methods of defining and assessing validity. One of the many contributions of the American Psychological Association's Technical recommendations for Psychological Tests and Diagnostic Techniques (1954) was the development of a fairly simple system for classifying procedures for assessing validity. The Technical Recommendations recognized four essentially different ways of defining validity:

1. Content validity
2. Construct validity
3. Predictive validity
4. Concurrent validity
5. Face Validity
6. Convergent validity
7. Discriminant validity

These four categories are sometimes referred to as the four faces of validity.

For many years it was thought that different types of validity were appropriate for different purposes. For example, for educational testing, content validity might be called for; whereas for personnel testing, predictive validity might be needed. Landy (1987) referred to this approach to matching "types" of validity to specific testing applications as "stamp collecting." Today, it is recognized that these four "faces" represent four different strategies for validating the inferences that are made on the basis of test scores rather than four different types of validity (A PA Standards for Educational and Psychological Testing, 1985). Rather than describing fundamentally different types of validity, researchers now agree that all validation strategies are designed to pursue the same basic goal: understanding the meaning and implication of test scores. Messick (1989) provides a succinct definition of validation as the "scientific inquiry into test score meaning." Nevertheless, there is some value in considering separately two different uses of the term "validity," which correspond to the two different ways in which tests are most often used—for measurement and prediction or for decision making. Both content and construct validation strategies represent approaches for determining whether a test provides a valid measure of a specific attribute. In other words, these approaches define validity in terms of measurement—a test is valid if it measures what it is supposed to measure. Predictive and concurrent approaches examine the validity of predictions or decisions that are based on tests—a test is valid if it can be used to make correct or accurate decisions.

To illustrate the differences between validity of measurement and validity for decisions, consider the case of an organization that decides to use an inventory labeled Leadership Skills Profile to help select managers. First, you might ask whether this inventory really tells you anything about a person's leadership skills (validity of measurement). Second, you might ask whether people who receive high scores on this test turn out to be good Managers (validity for decisions). The difference between these two aspects of validity is that in the first case, you are concerned with what the test measures, whereas in the second, you are interested in using the test to make predictions about a variable that is not directly measured by the test (i.e., success as a manager), but that you think is related to what the test measures.

Guion (1991) notes that validity of measurement is not always necessary or sufficient to guarantee validity for decisions. A very good measure of leadership skills may be a poor predictor of performance as a manager, for leadership and day to-day management are very different things. A poor measure of leadership might nevertheless allow you to identify good managers. It is therefore important to consider the two major aspects of the term "validity" separately.

Assessing the Validity of Measurement

It is a simple matter to determine whether a homemade ruler provides a valid measure of length simply take the ruler to the Bureau of Weights and Measures and compare it with a standard. This strategy won't work for evaluating the validity of a psychological test. The reason for this is simple, but fundamental to an understanding of methods of psychological measurement: For many of the characteristics that psychologists wish to measure (e.g., introversion, intelligence, reading comprehension), there are no universal standards against which test scores can be compared. In other words, if I measure Paula's general intelligence and come up with an IQ of 112, I cannot go to the Bureau of Weights and Standards to find out if I was right. Unfortunately, there is no external standard I can use to check this figure or to check scores from most other psychological tests. Rather than validating test scores against some external standard, psychologists must employ more indirect methods in determining the validity of tests. That is psychologists must collect evidence from a variety of sources to demonstrate that tests measure what they are designed to measure.

In a sense, one's work is never done when attempting to establish the validity of measurement. There is no definitive way of proving that a given test is a measure of general intelligence, of reading comprehension, or of any other trait. Establishing the validity of a specific test is always partially subjective and depends on a researcher's judgment regarding the weight of the available evidence. Nevertheless, although the methods discussed in this chapter are both indirect and partially subjective, judgments regarding the validity of measurement can and should be based solidly on empirical evidence.

Both content and construct-oriented validation strategies involve the accumulation of evidence that suggests that the test actually measures what it purports to measure. Content validity is established by examining the test itself, while construct validity is established by examining the relationship between test scores and a variety of other measures.

Content-oriented Validation Strategies

One way to gather evidence for the validity of measurement is to examine the content of the test. A test that contains 25 addition and subtraction problems is probably a better measure of simple mathematical ability than a test that contains 10 questions about sports and no addition and subtraction problems. Content validity is established by showing that the behaviors sampled by the test are a representative sample of the attribute being measured. Thus, content validity depends both on the test itself and on the processes involved in responding to the test (Guion, 1977). For example, a paper-and-pencil test of job knowledge might not provide a valid measure of a worker's ability to do the job, even if it provides a valid measure of his or her knowledge of what to do in the job.

One can get a rough idea of a test's content validity simply by looking at test items.' If all test items appear to measure what the test is supposed to measure, there is some evidence of content validity. (The evidence is weak, but it is a start. In order to develop more detailed evidence of content validity, it is necessary to introduce the concept of content domain.

Content Domains

Every psychological test is a systematic sample from a particular domain of behaviors. A detailed description of the content domain provides the foundation for assessing content validity.

When you say, "I want to measure X," you have specified a particular content domain. A content domain represents the total set of behaviors that could be used to measure a specific attribute or characteristic of the individuals that are to be tested (Guion, 1977). For example, a test designed to measure performance as a baseball player could be constructed by sampling from the total domain of behaviors (running, fielding, hitting) involved in the game of baseball. The domain covered by a test might be very broad (e.g., reading comprehension), or it might be very narrow (e.g.. addition problems involving decimals to two places, with sums less than 10). Regardless of its size, every content domain has a number of properties that are useful in assessing content validity.

First, as the name implies, a content domain has boundaries. There are a great many possible test items within these boundaries that could validly be used to measure a person's standing on the content domain; a detailed description of the content domain to be measured allows one to determine whether each test item lies within the boundaries of the domain. There is clearly a problem if a large number of behaviors sampled by file test are outside the boundaries of the domain one wishes to measure. Returning to an earlier example, a test that is made up of questions about sports will not provide a valid measure of simple mathematical ability. The test might provide an accurate measure of sports knowledge, but questions about sports fall outside the boundaries of the domain of mathematical ability,

Second, content domains are structured. That is the contents of a content domain can often be classified into several categories. The description of such a content domain presented in Table 6.1 helps clarify the concepts of boundaries and structure of content domains. The domain described in the table has well-defined boundaries and structure. Because this domain is very concrete, it is possible to make some precise statements about the areas include^ in the domain and about the relative importance of each of those areas. This detailed description of the boundaries and structure of the content domain is essential in evaluating content validity.

It should be relatively easy to decide whether specific test items are within or outside the boundaries of the domain describe in table. It should be easy to

Table 6.1 Detailed Description of a Content Domain

1. Domain to be measured: Knowledge of world history as covered in a standard seventh grade course.

2. Areas included in this domain:

A. Issues	B. Areas	C. Time span
1. Social	1. .Europe	1. 18th Century
2. Political	2. Americas	2. 19th Century
3. Cultural	3. Africa & Asia	

3. Relative importance of the areas covered.

		Social	Political	Cultural
Europe	18th Century	5%	10%	3%
	19th Century	5%	8%	2%
Americas	18th Century	6%	17%	2%
	19th Century	9%	13%	5%
Africa & Asia	18th Century	2%	0%	0%
	19th Century	6%	5%	2%
				100%

decide whether a test forms a representative sample of this content domain. As discussed in the section that follows, detailed comparisons between the boundaries and structure of the content domain and the structure of the test are at the heart of content validity.

Unfortunately, many content domains cannot be described in the level of detail shown in the table. It is very difficult to provide detailed descriptions of content domains such as ""mathematical concepts" or "performance on mechanical tasks." These domains include a broad range of behaviors whose boundaries might be difficult to specify. It may therefore be very difficult to decide whether specific test items are within or outside of the boundaries of the domain. If the different areas or classes of behaviors included in the domain cannot be categorized, or if the relative importance of those areas cannot be decided upon, it may be impossible to determine whether the test provides a representative sample of the content domain.

Assessing Content Validity

There is no exact, statistical measure of content validity! Rather, content validity represents a judgment regarding the degree to which a test provides an adequate sample of a particular content domain (Guion, 1977). Judgments about content validity are neither final nor absolute; tests show various levels of content validity, and experts do not always agree in their judgments. Nevertheless, judgments about content validity are not arbitrary. A detailed description of the content domain provides a framework for the careful evaluation of tests and provides a method for systematic evaluation of the validity of measurement.

The basic procedure for assessing content validity consists of three steps:

1. Describe the content domain.
2. Determine the areas of the content domain that are measured by each test item.
3. Compare the structure of the test with the structure of the content domain.

Although this procedure appears to be simple, in practice it is difficult to implement. The principal difficulty is encountered at the first step, the description of the content domain. Outside of the area of classroom testing it is often difficult to describe content domains in the level of detail shown in Table 7-1. For example, consider applying this strategy in assessing a test of vernal ability. Although it might be possible to decide what sort of tasks were or were not included in this content domain (boundaries of the domain), it would be exceedingly difficult to define in any detail the structure of this domain. In other words, it would be difficult to specify the relative importance or frequency of the different tasks that involve verbal ability. In this case, a test developer might have a general definition of the domain he or she wishes to measure but little detailed knowledge of the structure of that domain. It would be difficult to determine whether a specific test provided a representative sample of this domain; therefore, it would be difficult to argue on the basis of content alone that the test provides a valid measure of verbal ability.

A detailed description .of the content domain yields a set of categories that can be used to classify test item}. First, each test item can be classified as being within the boundaries of the domain or outside the boundaries of the domain. This type of classification is simple, but important; a test is not likely to show much content validity if the majority of the test items are clearly outside of the boundaries of the domain it is supposed to measure. Those test items that are within the boundaries of the domain can be further classified according to the areas of the domain they measure. For example, items on a history test designed to measure the domain described in Table 6.1 could be classified as dealing with social, political, or cultural issues, and further classified as dealing with European, American, or African and Asian history. In other words, the categories used to describe the content domain can also be used to classify each item on the testy'"

The final step in assessing content validity is to compare the content and structure of the test with that of the content domain. If none of the test items falls within the boundaries of the domain, then it seems clear that the test will show no content validity. Furthermore, if test items deal with only a specific portion of the domain, the test will not show substantial evidence of content validity.

Consider, for example, the tests described in Table 6.2. The domain we wish to measure is performance as a shortstop, which presumably includes hit dug, base running, and fielding a variety of balls. Each of the two tests described in the table includes behaviors that are clearly within this content domain, yet neither of them provides a valid

Table 6.2 Examples of Tests That Fail to Adequately Sample a Domain

The Domain:	Performance as a shortstop in intermural softball
Test A:	Hitting 40 pitches thrown in batting practice
Test B:	Fielding 70 ground balls that are hit toward second base

sample of the domain. In this example, some combination of test A and test B might provide a better sample of this particular content domain and thus a better, more valid test. In general, a test that appears to provide a representative sample of the major parts of a content domain will be judged to show high levels of content validity. The closer the match between the structure of the test and the structure of the domain, the stronger the evidence of content validity. In order to demonstrate content validity, tests should sample all parts of the content domain and should devote the largest number of items to the larger, more important areas included in the domain, with comparatively fewer items devoted to less important aspects of the domain.

Response processes and content validity. Several researchers, notably Guion (1977) and Sackett (1987), have noted that assessments of content validity focus almost exclusively on the content of test items or assessment exercises. Questions of how stimuli are presented to subjects, how responses are recorded and evaluated, and what is going through the respondent's mind are rarely considered in studies that use content-oriented validity strategies. Failure to consider these issues adequately could be a serious mistake, as can be illustrated through an examination of the underlying logic of content-oriented validation.

The logic of content-oriented strategies of validation is that the test should provide a representative sample of the domain one wishes to measure. By this logic, validation efforts should be concerned more with the extent to which responses to items provide a representative sample of the domain of responses than with the extent to which stimuli are representative. After all, psychological tests are samples of behavior that are used to draw inferences about the domain of behaviors sampled. Therefore, a test that used unrepresentative stimuli but sampled the domain of responses well would be preferable to one that included representative stimuli but unrepresentative responses. An example of the latter is the work sample test, which is often used as a predictor, and sometimes as a measure, of job performance. Work sample tests ask respondents to carry out, under optimal conditions, a standard set of tasks that are part of the job; their performance on each task is carefully observed and scored. In terms of the stimuli employed, these tests are often highly representative and realistic.

However, it is clear that these tests typically measure maximal rather than typical performance. That is, people typically perform their best on these tests and probably do not perform as well on the job, where they are not so closely monitored and not so free of distractions. It has long been argued that tests of maximal performance measure different things than tests of typical performance (i.e., how well a person can do a task versus how well he or she does it; see Cronbach, 1970, for an excellent discussion of the typical-maximal distinction). Research has shown that tests of typical versus maximal performance are not highly correlated (Sackett, Zedeck & Fogli, 19.S8). A content-oriented analysis of work sample tests would not account for this finding unless it considered the effects of a wide range of variables, such as the respondent’s motivation to perform on responses to test items.

Outcome of a Content Validity Study

The principal outcome of a content validity study is a judgment about the adequacy with which the test samples a particular content domain. Lawshe (1975) has proposed a content validity ratio as a measure of the extent to which expert judges agree on content validity, but this statistic measures agreement rather than validity itself. To our knowledge, there is no single statistic that can be used to measure content validity.

Although there is no exact measure of content validity, it is clear that some studies provide more and better evidence for content validity than others. The key to deciding whether it is possible to make systematic and reliable judgments about content validity lies in the description of the content domain. The more detail provided about the boundaries and structure of the content domain, the more confidence that can be placed in judgments about content validity. II a test developer cannot clearly describe the boundaries and the contents of a particular domain; it is difficult to see how he or she could ever convincingly demonstrate content validity. However, even if we assume that the domain is well understood and that we can, with some confidence, establish that the test is a representative sample from the domain, we are still not out of the woods. Even if the right types of items are sampled, the way they are written may be confusing or the response formats used may be inappropriate. Thus, two tests that both show strong evidence of content validity will not necessarily produce identical scores. Although a test that is known to provide a representative sample of a particular domain is very likely to provide a valid and accurate measure of that domain, it is important to remember that a content validity study cannot, by itself, guarantee the validity of measurement.

Content Validity, Reliability, and the Validity of Decisions

You may have noted the strong similarity between our discussion .of reliability and our discussion of content validity. Although reliability studies and content validity studies address somewhat different questions, reliability and content validity are conceptually similar (Cronbach et al., 1972). The difference is mainly one of emphasis. Reliability theory assumes that the test represents a sample from a domain of possible test items: this same assumption provides the basis for studies of content validity. The principal difference between a reliability study and a content validity study lies in the emphasis placed on providing a precise description of the domain. Reliability theory simply assumes that there is a domain and that the test could be lengthened by sampling more items from that domain. In a content validity study, the researcher must describe in detail which domain he or she wishes to measure. Thus, if a test provides a reliable measure of some domain but fails to measure the particular domain that is of interest, one might achieve a high level of reliability with little or no content validity.

It seems clear that content validity is important to understanding test scores. However, there is some controversy over whether content validity can be used to establish the validity of decisions based on test scores. A number of researchers have suggested that a content validity approach might be useful in determining whether specific tests could be used in applications such as personnel selection. The basic argument is as follows: (a) tests are used to predict performance on the job; (b) job performance requires certain abilities and skills; (c) if the tests require the same abilities and skills as those required on the job, then tests could be used to predict job performance; and (d) therefore, the validity of a test for selection decisions can be established by comparing the content of the test with the content of the job. This type of definition of content validity has been widely accepted both by industry and by the federal government. However, most experts agree that content validity is relevant only in determining the validity of measurement (does the test measure what it claims to measure'.'), not in determining the validity of decisions that are made based on test scores.

Carrier, Delessio, and Broun (1990) investigated the hypothesis that judgments about the content validity of tests would allow one to assess the validity of those tests as predictors of important criteria. Their results are sufficiently ambiguous to give comfort to the supporters and the detractors of content-oriented strategies of assessing validity for decisions. They found that expert judgments of content validity were significantly correlated with levels of criterion-related validity, but that these correlations were small. Their results suggest that content-related evidence is useful but not sufficient for assessing the criterion-related validity of psychological tests.

Construct-Oriented Validation Strategies

Psychologists are keenly interested in measuring abstract attributes—happiness, intelligence, motivation, sociability. These things do not exist in the literal, physical sense: it is impossible to gather up a pound of happiness or a handful of intelligence. Nevertheless, they must be measured in order to apply, test, and extend psychological theories and principles.

The problem of measuring abstract attributes is by no means restricted to psychology. Physicists routinely measure unobservable properties of matter. Mass provides a good example of this type of property: mass itself cannot be seen or heard, yet this hypothetical property of objects is clearly important and clearly measurable. Attributes such mass, happiness or intelligence are referred to as constructs. They represent ideas constructed by scientists to help summarize a group of related phenomena or objects. For example, if a person tells the truth in a wide variety of situations, we might label that person as honest. Honesty is a construct; it cannot be directly observed, yet it is a useful concept for understanding, describing, and predicting human behavior.

Tests are often designed to measure psychological constructs. Some tests provide valid measures of important constructs, while others show little or no construct validity. Because constructs are abstract in nature, the process of determining whether a test provides an adequate measure of a specific construct is complex.2 In order to describe the process of construct validation, we must first discuss the nature of psychological constructs

All constructs have two essential properties: They are abstract summaries of some regularity in nature, and they are related to or connected with concrete, observable entities or events. Gravity provides a good example of a construct: when apples fall to earth the construct gravity is used to explain and predict their behavior. It is impossible to see gravity itself; all one sees is the falling apple. Nevertheless, it makes perfect sense to measure gravity and to develop theories that employ the construct gravity. It certainly seems more sensible to deal with this abstract force we call gravity than to develop theories and methods that apply only to falling apples.

Constructs are essential to science. They represent departures from our immediate sensory experiences that are necessary in order to form scientific laws. They allow us to generalize from an experiment involving falling apples to situations involving a variety of falling objects. A construct such as gravity is related to a number of concrete objects and events. Thus, once I learn about gravity, I will be able to predict a wide variety of phenomena.

Constructs tire not restricted to unseen forces, such as gravity, or to processes, such as learning. Rather, any group of similar things or events may serve to define a construct. Thus, most categories that we use to classify and discuss everyday objects or events are in fact constructs, for example, the color red is a construct. There are plenty of red things, some of which plainly vary in color, but the basic idea of red is an abstraction. Poverty, reading ability, and cognitive style are thus all labels for constructs (Cronbach, 1971).

Although constructs are themselves hypothetical abstractions, all constructs are related to real, observable things or events. The distinguishing feature of psychological constructs is that they are always related directly or indirectly, to behavior or experience.

Some constructs such as aggressiveness and achievement motivation as thought of as causes of particular behaviors. Other constructs, such as pleasure or verbal ability or musical talent, refer to the ability to perform a number of related behaviors As discussed in the section that follows, still other psychological constructs show no direct connection with observable behaviors; rather, they are connected with other constructs that are, in turn, connected with behavior or experience.

Psychological Constructs

Psychological measurement is a process based upon concrete, observable behaviors. Hence, a psychological test is nothing more than a sample of behaviors. To determine whether a test provides a good measure or a specific construct, we must translate the abstract construct into concrete, behavioral terms. The process of providing a detailed description of the relationship between specific Behaviors and abstract constructs, referred to as construct explication, is the key to determining the construct validity of a .test. The process of construct explication consists of three steps:

1, Identify the behaviors that relate to the construct to be measured.

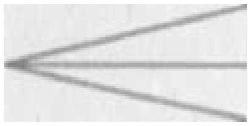
2., Identify other constructs and decide whether they are related or unrelated to the construct to be measured.

3., Identify behaviors that are related to each of these additional constructs and, on the basis of the relations among constructs, determine whether each behavior is related to the construct to be measured.

An example describing the three steps of construct explication is presented in Table 7-3. In the table the construct validity of a test designed to measure aggressiveness in school children is being examined. The first step in the process of construct explication is to describe behaviors that are related to aggressiveness. A child who assault other students

Table 6.3 Steps in Describing the Construct "Aggressiveness in School Children"

1. Identify behaviors related to aggressiveness
   Construct Behavior



2. Identify other constructs and decide whether they are related to aggressiveness

Construct Behavior

Need to power	
Aggressiveness	Assaults other students Pushes to head of line
	Dominates games
Honesty	

2. Identify behaviors that are related to each construct and determine their relation to the construct to be measured

Construct	Behavior
Need to power	Makes decisions in groups Dominates games
Aggressiveness	Assaults other students Pushes to head of line
Honesty	Refrains from cheating Tells truth to the teacher

Note: Constructs and behaviors that are related to one another are connected with a solid line. Unrelated constructs or behaviors are not connected.

who pushes to the head of the line, or who dominates most games might be labeled aggressive. There are many other behaviors that might be considered as examples or manifestations of aggressiveness, and at this stage we should try to provide as many as possible. The more behaviors we are able to list, the clearer the picture will be of what we mean when we say "aggressiveness."

The second step in providing a detailed description of the construct aggressiveness is to identify other constructs that could sensibly be measured in the subject population and to determine whether each of these is related or not related to aggressiveness. For example, it is useful to identify other constructs, such as need for power, that are related to aggressiveness. It is also useful to identify constructs that tire clearly not related to the specific construct to be measured. For example, the statement that the construct "aggressiveness" is unrelated to the construct "honesty" helps to define the boundaries of both aggressiveness and honesty

The third step is to identify behaviors that are related to each of these additional constructs. For example, a child who always makes decisions for groups or who dominates games might exhibit a high need for power. A child who refrains from cheating and who tells the truth to his or her teacher might be labeled as honest. Because we have made some statements about the relationships between the constructs of honesty, need for power, and aggressiveness, it should be possible to state whether each of these behaviors is related or unrelated to aggressiveness. For example, if aggressiveness and need for power are related, it is plausible that some behaviors that indicate high levels of need or power (e.g., "dominates games") will also indicate high levels of aggressiveness. Similarly, if aggressiveness and honesty are unrelated, then knowing that a student refrains from cheating or tells the truth to the teacher reveals nothing about his or her level of aggressiveness.

The end result of the process of construct explication is a detailed description of the relationships among a set of constructs and behaviors. This system of relationships, referred to as a nomological network, provides a definition of what we mean by "aggressiveness" (Cronbach & Meehl, 1955). Because constructs are abstract in nature, it is impossible to provide a concrete, operational definition of a term such as •aggressiveness." The nomological network provides an alternative way of systematically describing constructs. In our example, aggressiveness is defined as a personal characteristic that is related to a large number of behaviors (e.g., assaulting others), but which is not related to other behaviors (e.g., refraining from cheating). Although a construct cannot be directly observed, it can be inferred from observable behaviors. To put this another way we cannot say precisely what aggressiveness is, but we can describe how an aggressive child might act and we can make reliable and meaningful statements about children's level of aggressiveness by observing their behavior. The more detail included in descriptions of the nomological network, the more precision there will be in describing constructs.

Cronbach (1988, 1989) noted that applications of the methods described in Cronbach and Meehl (1955) have proved to be very difficult. He noted in 1988, "There is no hope for developing in the short run the "nomological networks' we once envisioned" (p. 13). Thus, construct validity has a somewhat more limited goal than was once envisioned. Rather than embedding each test in a complex network of associations with many other constructs, most construct validity researchers now pursue the goal of determining what inferences about psychological constructs can and cannot be made on the basis of a test score and under what conditions those inferences are or are not valid (Cronbach. 1988).

Assessing Construct Validity

The goal of construct validation is to determine whether test scores provide a good measure of a specific construct. The process of construct explication provides a definition of the construct in terms of concrete behaviors\\Although construct explication docs not define precisely what a construct such as aggressiveness is, it docs tell how that construct relates to a number of behaviors. A child who shows a high level of aggressiveness is likely to show certain behaviors (e.g., assaulting classmates) and is less likely to show other behaviors than a child who is low on aggressiveness A well-developed nomological network, therefore, provides detailed information about the relationship between the construct and a large number of behaviors. This information can be used to describe the way in which a good measure of a construct can be expected to relate to each of these behaviors. A test shows a high level of construct validity if the pattern of relationships between test scores and behavior measures is similar to the pattern of relationships that can be expected from a perfect measure of the construct. An example will help to clarify this point.

Our explication of the construct “aggressiveness” (see Table 6.3) suggests that the following behaviors are directly related to the construct: (a) assaulting other students, (b) pushing to the head of lines, and (c) dominating games. In other words, we would expect measures of these behaviors to be positively correlated with a good measure of aggressiveness. The behavior "makes decisions in groups" was related to a construct (need for power) that was in turn related to aggressiveness; we might therefore expect measures of this behavior to show a positive correlation with measures of aggressiveness. Finally, the behaviors “refrain from cheating" and "tells truth to teacher" are not at all related to aggressiveness. We therefore might expect measures of these behaviors to be uncorrelated with a measure of aggressiveness. The correlations we would expect between measures of each of the behaviors and measures of aggressiveness are summarized in Table 6.4. detailed description of the construct provides a basis for describing the relationships to be expected between a good measure of that construct and a variety of behaviors Actual test scores can be correlated with behavior measures, and the results can be compared with the pattern of results expected on the basis of our explication of the construct. The stronger the match between the expected correlations and the actual correlations) between test scores and behavior measures, the stronger the evidence of construct validity.

Table 6.5 shows comparisons between expected and actual correlations for two tests designed to measure aggressiveness in school children. Test A appears to be a valid measure of the construct "aggressiveness"; the correlations between test scores and behavior measures are very similar to the correlations one would expect on the basis of our theory of aggressiveness. In contrast, the data suggest that test B is a poor measure of aggressiveness. Behaviors we would expect to correlate strongly with aggressiveness show weak and sometimes negative correlations with test scores. Other behaviors that have nothing to do with aggressiveness show fairly sizable correlations with test B. It appears fair to conclude that test B does not measure aggressiveness as we have defined it.

Table 6.4 The Expected Correlations Between a Good Measure of Aggressiveness And Measures of Specific Behaviors

Behaviors	Relationship with aggressiveness	Expected correlation
Assaulting others	Direct	Strong Positive
Pushing in line	Direct	Strong Positive
Dominating games	Direct	Strong Positive
Making decisions	Indirect—related to need for power	Weak Positive
Refraining from cheating	None	None

Telling truth to teacher	None	None

Table 6.5 Correlations between Test Scores and Behavior Measures for Two Tests

Behaviors	Expected correlations	Actual correlations	
		Test A	Test B
Assaulting others	Strong Positive	.59	-.22
Pushing in line	Strong Positive	.70	.14
Dominating games	Strong Positive	.65	.02
Making decisions in groups	Weak Positive	.30	-.04
Refraining from cheating	None	.09	.56
Telling truth to teacher	None	-.04	.39
		(Test with high level of construct validity)	(Test with low level of construct validity)

Construct validity depends upon a detailed description of the relationship between the construct and a number of different behaviors. The more we know about the construct, the better our chances for determining whether a test provides an adequate measure of that construct. One implication is that it will be easier to determine construct validity for measures of well-defined constructs than for measures of constructs that are loosely defined. If I define a new construct but have only a fuzzy idea of what that construct means, it follows that 1 will never be able to tell whether a given test provides a good measure of that construct.

A very sophisticated version of the validation strategy described here was applied by Mumford, Weeks, Harding, and Fleishman (1988) in their analysis of the relationships among student characteristics, course content, and training outcomes. Their study incorporated measures of six student characteristics, sixteen measures of course content, and seven measures of training outcomes. They articulated hypotheses about the relationships between each of these 29 measures and criterion variables, and they developed an integrated model that described the hypothesized relationships. This model allowed them to describe the relationships among the three constructs (i.e.. characteristics, content, and outcomes) and to draw conclusions about the relative importance of individual versus situational variables in determining training outcomes.

Another example of this approach is in a study by Pulakos, Borman, and Hough (1988), in which experts were first asked to estimate the correlations they would expect between specific predictors and criteria. This allowed the authors to compare the estimated correlations with the correlations they actually found. Results of one study of army enlisted personnel, in which there were eight predictors and three criteria, are presented in Table 6.6. In this study, 83 percent of the obtained correlations were in the predicted direction. Although the observed correlations were typically smaller than the estimated correlations .87 percent of the correlations were of the relative magnitude predicted by the experts (i.e., those predicted to be largest were, in fact, usually the largest). These results present strong evidence for the construct validity of the measures employed.

Table 6.6 Estimated and Obtained Correlations for Army Enlisted Personnel

	Criteria					
	Technical skill		Personal discipline		Military bearing	
Predictors	Est.	Obs.	Est.	Obs.	Est	Obs.

Technical	.38	.21	.11	.00	.09	-.18
Quantitative	.27	.17	.10	.06	.07	-.08
Verbal	.29	.16	.08	.04	.08	-.19
Speed	.16	.09	.07	.04	.06	.07
Achievement Orientation	.50	.23	.36	.03	.25	.17
Dependability	.36	.15	.54	.22	.31	.14
Adjustment	.34	.12	.44	.05	.19	.11
Physical Fitness	.16	-.01	.10	-.11	.54	.27

Methods of Construct Validation

Which methods are most appropriate for studying construct validity depend to a large extent on the construct we wish to measure combination of laboratory experiments, Held experiments questionnaires, and unobtrusive observations might be necessary to provide the data that underly construct validation-Statistics that deal with differences between control and experimental groups, with correlations between various measures, with individual variations, or with change over time might be used to test predictions that are based upon our description of the construct. Data about the reliability of a test under different conditions might contribute to our assessment of construct validity. In fact, it would be fair to say that any type of data or statistic might be useful in determining construct validity.

Although any method might be used to assess construct validity, a few methods, seem to be most common. The most basic method is to correlate scores on the test in question with scores on a number of other tests. Here, the word "test" is used broadly to indicate any type of behavioral measure. We have already discussed this basic method in some depth will return to a specialized application of this method in a later section. Another common method of studying construct validity involves the mathematical technique known as factor analysis (see Chapter 3). Factors are very much like constructs, and factor analysis provides an analytical method for estimating the correlation between a specific variable (a test score) and scores on the factor. Factor analysis also provides a compact summary of information about the relationship among a large number of measures. The description of a construct provides information about the expected relation ships among variables: factor analysis helps determine whether this pattern of relationships docs indeed exist.

A third common method of studying construct validity involves experimental manipulation of the construct that is to be measured. For example, a test designed to measure anxiety should show higher scores for subjects in an experiment who are led to expect shocks than for subjects who expect to fill out an innocuous questionnaire. On the other hand, if a study has nothing to do with anxiety, the control group and the experimental group would not be expected to receive different scores on the test of anxiety. A combination of experiments in which the construct of interest is manipulated and experiments in which that construct is not manipulated provides a powerful method for assessing construct validity.

A study by Flynn (1987) illustrates the use of archival data and reviews of existing research in construct validation. This study reviewed research from 14 nations documenting substantial rises in 10 scores over the last 30 to 40 years. Flynn (1987) noted that there are several correlates of increasing IQ scores, including increases in scientific achievement, superior performance in schools, and an increase in the number of individuals classified as '•geniuses." He then surveyed selected newspapers, magazines, and educational journals in countries exhibiting substantial increases in 10 scores and looked for evidence of these phenomena (e.g., an increasing number of patents and inventions, news stories describing an increase in the number of geniuses). His survey showed no evidence of increases in scientific achievement, inventions, and so forth, to go along with the massive increases in 10 leading him to question whether the IQ tests surveyed really measured intelligence.

Validity for Decisions Criterion-Related Validity

As noted in Chapter 1, a major reason for our interest in tests is that they are used to make important decisions about individuals. Tests do not always lead to correct decisions: but compared to other alternatives, they are thought to represent the most accurate, fair, and economical method of making decisions (Wigdor & Garner, 1982a). In fact, in settings where decisions must be made about large numbers of individuals (e.g.. in screening military recruits), psychological tests often represent the only practical method of making systematic decisions. . The validity of tests as decision-making aids is a topic of great practical importance. The accuracy of decisions is directly linked to the validity of test scores—an invalid test can lead to decisions that are both ineffective, from the decision maker's point of view, and unfair, from the individual's point of view/The simplest method of determining, whether a test can be used validly in making decisions is to correlate test scores with measures of success or of the outcomes of decisions. These measures are referred as criteria; hence the term "criterion-related validity" !The correlation between test scores and criteria provides a quantitative estimate of validity, which in turn can be used to obtain a detailed picture of the effect of testing on decisions.) In particular, measures of criterion-related validity provide means to determine whether tests will serve to reduce particular types of decision errors and whether the gains in the accuracy of decisions are worth the costs of testing! This chapter presents a number of methods of estimating and interpreting the criterion-related validity of a test.

Decisions and Prediction

Mathematical decision theory makes a formal distinction between a decision, which involves choosing a particular course of action, and a prediction, which involves estimating the value of some variable, such as success in school, on the basis of what is known about another variable, such as Scholastic Aptitude Test (SAT) scores. Criterion-related validity deals with the correlation between tests and criteria, or with the degree to which test scores can be used to predict criteria. Although a prediction is formally different from a decision, there are strong conceptual similarities between predictions and decisions. In fact, these are sufficiently strong that we will treat the correlation between test scores and criteria as synonymous with validity for decisions. An example will help to clarify both the similarities and the differences between predictions and decisions.

Assume that you are a personnel manager trying to pick the best three applicants for the job of master machinist out of the five shown in Table 6.7.

Table 6.7 Mechanical Comprehension Scores of Five Applicants

Applicant	Scores on mechanical comprehension test (100 = Perfect score)
A	98
B	82
C	81
D	43
E	29

Applicants A, B, and C have reasonably high test scores, while D and E have low scores. If you had no other information, you would probably predict that A, B, and C would perform well on the job, and you therefore would hire these three and reject D and E.1 Prediction occurs when you try to estimate a person's score on a particular measure, such as a measure of job performance, on the basis of that person's score on some other measure, such as a mechanical comprehension test. Thus, you predict that applicants with high test scores will perform well on the job. You never know how someone will actually perform, but a good test allows you to make fairly accurate predictions. A decision represents some action that you take on the basis of your predictions. You hire someone because you predict he or she will perform well on the job; you select a particular course of psychotherapy because you predict that it will be most beneficial for a particular client; you place a child in a special remedial class because you predict that the class will provide a suitable learning environment for the child. Predictions are not always accurate, and therefore tests do not always lead you to make the correct decisions; however, the more accurate your predictions, the better your decisions will be.

The correlation between test scores and a measure of the outcome of a decision (the criterion) provides an overall measure of the accuracy of predictions. Therefore, the correlation between test scores and criterion scores can be thought of as a measure of the validity of decisions. As will be seen in a later section, thorough investigation of the effect of tests on decisions involves a number of factors in addition to the criterion-related validity of the test. Nevertheless, the validity coefficient, or the correlation between test scores and criterion scores, provides the basic measure of the validity of a test for making decisions.

Criteria

A criterion is a measure that could be used to determine the accuracy of a decision; in psychological testing, criteria typically represent measures of the outcomes that specific treatments or decisions are designed to produce. For example, workers are selected for jobs on the basis of predictions the personnel department makes regarding their future performance on the job; the job applicants who are actually hired are those who, on the basis of test scores or other measures, are predicted to perform at the highest level. Actual measures of performance on the job serve as criteria for evaluating the personnel department's decisions/ If the workers who were hired actually do perform at a higher level than would have those who were not hired, the predictions of the personnel department are confirmed, or validated. In a similar way, measures of grade point average or years to complete a degree might serve as criteria for evaluating selection and placement decisions in the schools. A particular strategy for making college admissions decisions may be a good one if the students who are accepted receive better grades or complete their degrees in a shorter time than the applicants who were rejected would have dime. Again, in a similar way, measures of adjustment or of symptom severity might be used to evaluate decisions regarding the choice of psychotherapy.

An example will help to illustrate the importance of developing appropriate criterion measures. For years, the military has used the Armed Services Vocational Aptitude Battery (ASVAB) in the selection and placement of new recruits. The principal evidence for the validity of the ASVAB has been the finding that ASVAB scores are consistently correlated with success in military training courses. However, the ASVAB was designed and has been used to predict performance on the job, not performance in training. Thus, until recently it was not proven that the ASVAB leads to correct decisions, although research on the validity of similar tests suggests that it does (Schmidt & Hunter. 1981). To correct this, the military undertook large-scale research efforts to develop measures of job performance and to assess the criterion-related validity of the ASVAB.

Unfortunately, the choice of criterion measures used in determining the validity of tests is often made in a careless or haphazard manner (Guion, 1965aWfhe key to choosing criterion measures is to determine the decision maker's goal. The goal of the personnel department is to select productive workers; the appropriate criterion for evaluating its decisions will be a measure of productivity. The goal of admissions directors is to select students who are capable of performing well in classes; measures of classroom performance will supply criteria for evaluating their decisions.

Criterion-related Validation Strategies

There are two general methods for assessing criterion-related validity: predictive and concurrent validation strategies. Predictive validity is recognized as the most accurate method of estimating validity, but is also recognized as presenting the most serious practical and ethical problems. Concurrent validity is a generic term which refers to a variety of more practical procedures for assessing validity. Barrett, Phillips, and Alexander (1981), Guion and Cranny (1982), and Schmitt Gooding, Noe, and Kirsch (1984) note that the conceptual and empirical distinctions between predictive and concurrent validity are not necessarily fundamental. However, the two general approaches do raise different practical, ethical, and statistical issues and are therefore worth discussing separately. We should emphasize here that these two approaches represent different strategies for estimating the same quantity, the correlation between test scores and criterion scores. These two strategies differ in a number of practical and operational details, but not in their fundamental goals.

The Ideal: Predictive Validation Strategies

The goal of a predictive validity study is to determine the correlation between test scores, which are obtained before making decisions, and criterion scores, which are obtained after making decisions. Personnel selection represents a typical setting in which a predictive validity study might be carried out; here, a decision must be made to hire or reject each applicant, and a measure of job performance serves as a criterion for evaluating that decision. In this setting, a predictive validity study consists of two simple steps:

1. Obtain test scores from a group of applicants, but do not use the test, either directly or indirectly, in making hiring decisions.
2. At some later time obtain performance measures for those persons hired and correlate those measures with test scores to obtain the predictive validity coefficient.

The predictive validity approach is thought of as an ideal for two conflicting reasons. From the scientific point of view, it represents the simplest and most accurate strategy for estimating the correlation between test scores and criterion scores in the population of applicants in general. Yet in another sense, this strategy is impractical, so although it is considered an ideal strategy for estimating validity, it is not a realistic one. The distinguishing feature of the predictive validity approach lies in the fact that decisions are made about applicants without using the test for making decisions, either directly or indirectly.

The advantage of the predictive validity approach is that it provides a simple and direct measure of the relationship between scores on the test and performance on the criterion for the population of applicants in general. If the test were used to select applicants, the correlation between test scores and performance measures (which are collected at some later time) would indicate the validity of the test in the population of people with high test scores (those selected using the test), rather than the validity of the test in the population in general. In most decision situations, the goal is to select those most likely to succeed out of the total population of applicants. In order to estimate the validity of a test for this type of decision, an estimate of the correlation between test scores and performance scores must be obtained for all applicants. Most practical methods for estimating the validity of decisions involve correlating test scores and criterion scores in some preselected population (e.g., present workers) and therefore fall short of the ideal predictive validity approach.

Practical objections. The key to the predictive validity approach is the requirement that the population in the validity study be similar to the general population of applicants. The only effective way to cans' out a predictive validity study is either to make the same decision for everyone (e.g., hire all applicants) or to make decisions on a random basis (e.g... flip a coin). It is understandable that decision makers object to tin approach that forces them to abandon, albeit temporarily, any system of selection in order to guarantee that the sample of people selected is representative of the population of applicants. It is impractical to hire people, admit them to school, or assign them to different types of therapy on a random basis. Of course, if the present system has little or no validity, decisions based on a table of random numbers are not likely to be any worse. Still, as a matter of practical politics, it is difficult to envision many people accepting an explicitly random system, such as would be necessary in a true predictive validity approach.

Ethical objections. An incorrect decision has negative consequences for both the individual and the decision maker. For example, an organization that hires a worker who has little chance of success on the job is likely to incur substantial losses in terms of training costs and lost productivity. The individual also incurs some serious costs—failure on the job is a very negative experience and may contribute to depression or to loss of self-esteem, not to mention the monetary loss incurred when the employee is subsequently dismissed. Likewise the psychologist who adopts a predictive validity approach would have to tolerate a number of failures that are probably preventable, with sometimes serious consequences to the clients involved. Thus, the predictive validity approach puts the decision maker in the ethically precarious position of selecting some people who he or she believes are very likely to fail (i.e... those with low test scores).

The Practical Alternative: Concurrent Validation Strategies

The practical alternative to a predictive validity strategy is simply to obtain both test scores and criterion scores in some intact, preselected population and to compute the correlation between the two. Since many research designs of this type call for obtaining test scores and criterion scores at roughly the same time, they are known as concurrent validation strategies. As Guion and Cranny (1982) point out, the delay between obtaining test scores and obtaining criterion is not really the most fundamental difference between predictive and concurrent validity. The most fundamental difference is that a predictive validity coefficient is obtained in a random sample of the population about whom decisions must be made, whereas a concurrent validity coefficient is generally obtained in a preselected sample (e.g.. present employees, students already accepted into college, patients in therapy) that may be systematically different from the population in general.

Guion and Cranny (1982) describe a number of common concurrent strategies for estimating the validity of a test for predicting scores on a specific criterion; three of these are especially common and can be described in simple terms (see also Sussman & Robertson, 1986). First, one can give the test to individuals who have already been selected (e.g., current workers, college freshmen) from the applicant population and obtain the criterion measure at approximately the same time the test is given. The correlation between test scores and criterion measures is a validity estimate. Next, one can select among applicants using the test (X) following up with a criterion measure (Y) at a later time. Finally, it is possible to use data from personnel files as measures X and Y. In all these designs, the correlation between test scores and criterion scores can be used to estimate the validity of the test, yet the population in which both the predictor and the criterion are measured may be systematically different from the population of applicants in general. The population of workers at a plant or students in a college is much more selective than the population who applied for jobs or who applied for school, since there are typically large numbers of applicants who receive low scores on tests and interviews and hence do not pass the selection process. The fact that the population in a concurrent validity study is significantly more selective than the population of applicants in general has a potentially serious effect on the correlation between test scores and criterion scores. The nature and the extent of this effect are discussed later in this section.

Although in applied settings concurrent validity is much more common than predictive validity, the predictive approach is generally preferred over the concurrent validity approach. The correlation between test scores and criterion scores in some preselected, intact population is in theory not likely to be the same as the correlation between test scores and criterion scores in the population in general. Yet it is the latter correlation that is most important in assessing the validity of decisions. In most cases, tests are used to make decisions about very general populations; the validity of a test in a highly select population will not necessarily reveal much about the validity of the test for making selection decisions.

Advantages. Three factors favor a concurrent validity approach. First* concurrent studies are practical. It is not necessary to select randomly or to allow a significant time lag between testing and criterion measurement in order to obtain concurrent validity coefficients. Second, concurrent validity studies are easier to conduct than predictive studies. In a concurrent study, it might be possible to obtain test scores and performance measures and to correlate them in a single day; a predictive study might last for months or even years. Third, although test theory suggests that concurrent validity coefficients seriously underestimate the population validity. Concurrent validities are in fact often similar in size to predictive validities (Barrett, Phillips & Alexander, 1981; Schmitt et al., 1984). Although there may be considerable theoretical differences between a predictive study and a concurrent study, the outcomes of these two types of studies are often sufficiently similar to justify the more practical concurrent validity approach.

Statistical problems. Research designs that use the correlation between test scores and criterion measures in a highly selective population to estimate the validity of a test for making decisions in the population in general give rise to a number of statistical problems. By far the most serious problem in concurrent designs is the range restriction that occurs when people are selected according to their test scores (i.e.. only those with high test scores are selected). Range restriction could have a substantial effect on the correlation between test scores and the criterion.

Range restriction also occurs in criterion measures. In work settings, people who perform very poorly are likely to be fired, whereas those who perform well are likely to be promoted. In school, students with consistently low grades are likely to fail or drop out. In clinical settings, clients are not likely to continue treatments that clearly are not working. As a result, the range of criterion scores is likely to be much smaller in an intact, preselected group (workers at a plant, students already admitted to college) than in the population in general.

Interpreting Validity Coefficients

A criterion-related validity study provides an estimate of the correlation between test scores and criterion measures. Theoretically, this correlation could range in absolute value from 0.0 to 1.0. In practice most validity coefficients tend to be fairly small—a good, carefully chosen test is not likely to show a correlation greater than .5 with an important criterion, and, in fact, validity coefficients greater than .3 are not all that common in applied settings. The correlations would almost certainly be higher if more reliable tests and criterion measures were used. Nevertheless, the levels of criterion-related validity achieved by most tests are rarely in excess of .6 to .7.

The figures shown in Table 6.8 provide a representative picture of the size of the validity coefficients often observed in personnel selection research. These correlations represent average validity coefficients computed across a number of studies, with a total N of over 140,000 (Schmidt Hunter & Pearlman, 1981). At first glance, these average validities look quite discouraging—the average of the correlations is .269. The squared correlation coefficient (r2) indicates the percentage of the variance in the criterion that can be accounted for by the predictor. An average criterion-related validity of .269 indicates that approximately 7 percent of the variability in job proficiency measures and measures of performance in training can be accounted for (on the average) by test scores. Stated another way, 93 percent of the variability in performance cannot be accounted for by tests.

Table 6.8 AVERAGE CRITERION-RELATED VALIDITIES ACROSS A NUMBER OF CLERICAL JOBS

Test	Job proficiency criteria	Training criteria
General Mental Ability	.24	.43
Verbal Ability	.19	.39
Quantitative Ability	.24	.43
Reasoning Ability	.21	.22
Perceptual Speed	.22	.22
Spatial/Mechanical	.14	.21
Clerical Aptitude	.25	.38

It is important to keep in mind that the values presented in Table 6.8 represent uncorrected coefficients. If corrected for attenuation and for range restriction, they could be substantially larger; collected values in the .50*s are often reported for tests of this sort (Hunter & Hunter. 1984). Nevertheless, even a validity of .50 would indicate that tests accounted for only 25 percent of the variance in the criterion (i.e., if r = .50, rz = .25). In most settings where tests are used to make important decisions, they account for relatively small percentages of the variability in criteria. Critics of testing (e.g., Nairn et al.. 1980) have used this point to support their argument that tests are worthless. In fact, things are not as bad as the typically low values of/- and /-: might suggest. The validity of a test as an aid in making decisions should be judged by evaluating the effects of tests on the accuracy of decisions. The validity coefficient does not in itself provide a complete measure of the effects of tests on decisions. In some situations, tests that are extremely accurate predictors lead to only marginal increases in the quality of decisions. In other situations, a test that shows a very low level of criterion-related validity may nevertheless contribute greatly to improving the quality of decisions. The effect of a test on the quality of decisions is affected by a number of factors that may be unrelated to the criterion-related validity of that test.

Tests and Decisions

The validity coefficient is only one of many factors that determines the degree to which a test may improve or detract from the quality of decisions. To fully evaluate the effect of a test on decisions, one must also consider the base rate and the selection ratio of a decision.

The base rate refers to the level of performance on the criterion in the population at large. For example, if 95 percent of all applicants perform successfully in a college course, the base rate is .95. If only 12 percent of the applicants to a psychology program successfully complete their degrees, the base rate is .12. Thus, the base rate indicates the percentage of the population who can be thought of as potential successes.

The selection ratio represents the ratio of positions to applicants. If 3PJ|eople apply for three jobs, there is a 10 percent selection ratio; if 10 prospective students apply for an incoming class of 9 there is a 90 percent selection ratio. Thus, the selection ratio indicates the degree to which the decision maker (e.g.. personnel manager, admissions director) can be selective in his or her decisions.

Outcomes of decisions. A decision represents some action taken with regard to a specific individual. A decision may involve accepting or rejecting a college applicant, or it may involve assigning or not assigning a client to a specific course of psychotherapy. Because predictions are never perfect, each decision may have many possible outcomes. In college admissions, some applicants who are accepted might turn out to be poor students, and others who were rejected might have been excellent students. A clinician might assign some clients to a course of therapy which is less than optimal. Nevertheless, the goal is to make the largest possible number of correct decisions—to accept potential successes and reject potential failures. A valid test should aid in accurately predicting both success and failures and should therefore contribute to the quality of decisions. Figure 6.1 presents a schematic representation of the decision process.

Tests, interviews, application blanks, and the like, present information about each individual. The formal or informal rules used in making decisions comprise -a strategy. For example, a college admissions office might decide to accept all applicants with a B + average and with SAT scores over 550. This strategy leads to a decision accept or reject—for each applicant. Furthermore, there are a number of possible outcomes of this decision. Some students who are accepted will succeed, others will fail. Some students who were rejected would have succeeded, and others would have failed.

The decision to accept an applicant is usually based upon the prediction that the applicant will succeed; the decision to reject an applicant implies a prediction of failure, or at least of a lower level of success. One way to evaluate the accuracy of decisions is to compare predictions with the outcomes of decisions. Figure 6.2 presents a cross-tabulation of predicted criterion scores with actual criterion scores. This type of a table described all the possible outcomes of a decision. One possible outcome is a true positive (TP): a person who is predicted to succeed and who actually does succeed. Another possible outcome is a true negative (TN): a person who is predicted to fail (rejected) and who actually would have failed had he or she had been accepted. True positives and true negatives represent accurate, correct decisions; they represent cases in which people who will succeed are accepted and those who will fail are not accepted. One of the principal goals of psychological testing is to increase the frequency of these correct decisions.

Another possible outcome of a decision is the false positive (FP): some of the people who are accepted turn out to be failures. Since the decision to accept a person who later turns out to be a failure implies the prediction that that person would succeed.

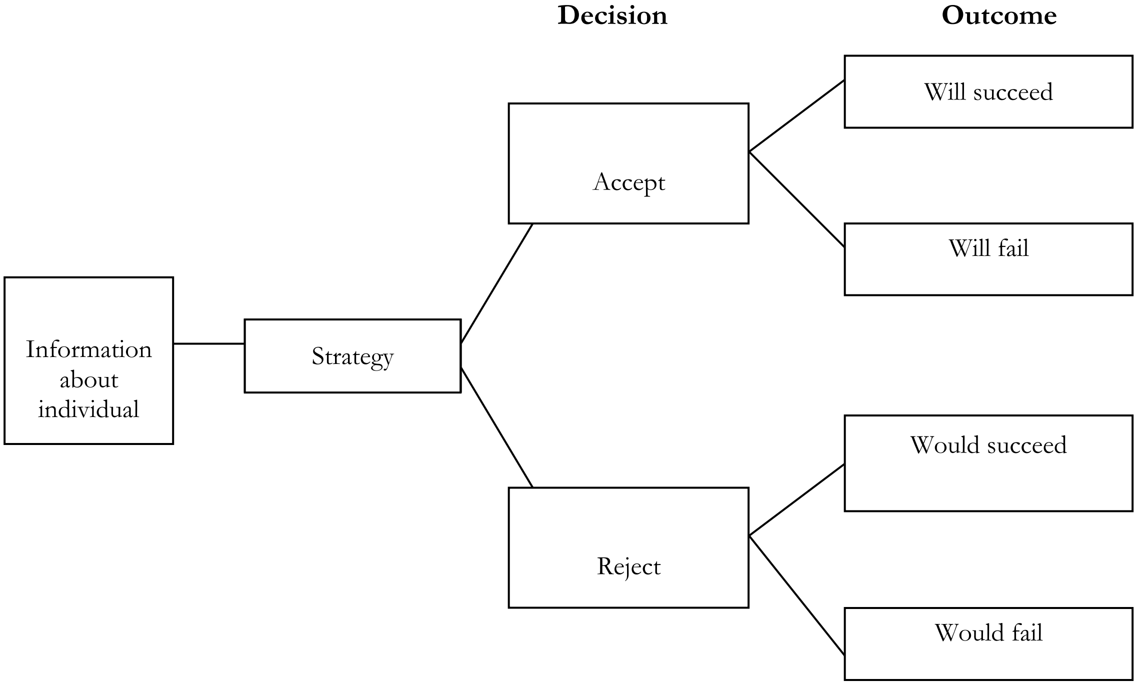


Figure 6.1 The Decision Process Source:

False Negatives (FN)	True Positives (TP)
True Negatives (TN)	False Positives (FP)
Reject (Predict failure)	Accept (Predict success)

Figure 6.2 Possible Outcomes of a Decision

the decision represents an error that is produced as the result of a false prediction. Finally, there are false negative (FN) decisions, in which an applicant, who would have succeeded, given the change, is rejected. False negatives are also decision errors in that they also lead to the wrong course of action for a specific applicant.

Base rates and decisions. One of the factors that strongly affects then it comes of a decision is the base rate (BR), or the number of potential successes in the population of applicants. If 90 percent of those who apply are likely to succeed, it should not be too difficult to select a highly successful group. With a base rate of 90 percent, random selection will on the average lead to successful applicants being chosen nine times out of ten. Indeed, when the base rate is very high, it may be difficult to devise a test that leads to significantly better decisions than would be reached by flipping a coin. If the base rate-is 90 percent, there may not be very much room for improvement.

A very high base rate practically guarantees a large number of true positive decisions. Unfortunately, it may also lead to a number of false negative decisions. Consider a group of 200 people who apply for 100 positions. A base rate of 90 percent means there are 180 potential successes in this group. Since there are only 100 positions, even the best strategy will result in rejecting 80 potential successes (false negative decisions). While an extremely high base rate may help to maximize the number of true positives, it may also lead to a number of decision errors, particularly false negative decisions.

Tests are often used to predict criteria that have very low base rates (Murphy, 1987b). For example, Wiggins (1973) discusses applications of personality tests to screen out draftees because of psychiatric disability. The base rate for severe psychiatric problems is undoubtedly low; it is likely that fewer than 5 percent of the population would be screened out of military service because of psychiatric disorders. Since the base rate is very low. most of the population are, in fact, negatives (no presence of psychiatric disorder). Because positives (those with psychiatric disorders) are very rare, any decision procedure that attempts to isolate this rare group is likely to classify falsely a number of negatives (no disorder) as positives (disorder present). In other words, with a very low base rate, false positive decision errors become much more likely. On the other hand, a low base rate practically guarantees a large number of true negatives. If there are very few psychiatric cases in the population, few people will be screened out by the test, and most of those not screened out will belong to the large group that docs not exhibit psychiatric disorders.

In general, tests are more likely to contribute to the overall quality of decisions when the base rate is around .50. When the base rate is extremely high, there is not much room for improvement in terms of locating true positives and there is little chance of avoiding a substantial number of false negatives. When the base rate is extremely low, a number of false positives is likely. When the base rate is around .50, there is a greater possibility of minimizing decision errors and it may be possible to make very accurate decisions if a test with sufficient validity can be found.

Selection ratio and decisions. The second factor that affects the quality of decisions is the selection ratio (SR), or the ratio of applicants to openings. If 100 people apply for 99 openings at a college, the college cannot be very selective. As a consequence, it doesn't really matter what strategy the college follows, the result will always be pretty much the same. Contrast this with the situation in which 100 people apply for one opening. Here, the validity of the decision strategy has a marked influence on the quality of the decision. A perfectly valid strategy will lead to selection of the best applicant; an invalid strategy could lead to selection of the worst. When the number of openings is large relative to the number of applicants, the selection ratio is high. For example, when ten people apply for eight positions, the selection ratio is equal to .8. In the extreme case where the number of applicants is equal to the number of openings, the selection ratio is equal to 1.0. In this case, there is no choice but to accept everyone.

When the number of openings is small relative to the number of applicants, the selection ratio is small. A small selection ratio indicates that there are few constraints on the decision. For example, if ten people apply for two openings, the selection ratio is .20, which means that eight out of every ten people who apply can be turned down. Of course, if there is no system for selecting among applicants, the freedom to turn down eight out of ten may not represent any real advantage. However, if there is a valid system for making decisions, a low selection ratio allows selection of the "cream of the crop." In fact, when the selection ratio is sufficiently low, a test with very modest validity still can contribute significantly to the accuracy of decisions.

Taylor and Russell (1939) published a series of tables that dramatically illustrate the effect of the selection ratio on the accuracy of decisions. Taylor-Russell tables indicate the proportion of successes that can be expected, given the validity of the test, the base rate, and the selection ratio. One of the Taylor-Russell tables is shown in Table 6.4; validity coefficients are shown at the side of the table, the selection ratio is shown at the top and the expected proportion of successes are shown in the body of the table. Several points should be noted in examining the table. First, if the test has a validity of .00, the expected level of success is exactly equal to the base rate, regardless of the selection ratio. In other words, a test with no validity yields essentially random decisions. Next, at very high selection ratios (e.g., SR = .90), the expected level of success using an extremely accurate test is not much higher than the level of success one would expect with a validity of .00. Finally, at very low selection ratios, a test with a reasonably low level of validity still could lead to substantial increases in the proportion of "successes"; when a highly valid test is used in a situation characterized by a low selection ratio; one is able to select successes essentially 101) percent of the time.

Table 6.9 Taylor-Russell Table Showing the Expected Proportion of Successes with a Base Rate Of .50

	Selection ratio									
Validity	**.05**	**.10**	**.20**	**.30**	**.40**	**.50**	**.60**	**.70**	**.80**	**.90**
.00	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.10	.54	.54	.53	.52 |	.52	.51	.51	.51	.51	.50
.20	.67	.64	.61	.59	.58	.56	.55	.53	.53	.52
.30	.74	.71	.67	.64	.62	.60	.58	.56	.54	.52
.40	.82	.78	.73	.69	.66	.63	.61	.58	.56	.53
.50	.88	.84	.78	.74	.70	.67	.63	.60	.57	.54
.60	.94	.90	.84	.79	.75	.70	.66	.62	.59	.45
.70	.98	.95	.90	.85	.80	.75	.70	.65	.60	.55
.80	1.00	.99	.95	.90	.85	.80	.73	.67	.61	.55
.90	1.00	1.00	.99	.97	.92	.86	.78	.70	.62	.56
1.00	1.00	1.00	1.00	1.00	1.00	1.00	.83	.71	.63	.56

It is important to select successes, but the overall quality of a decision strategy is affected not by the number of successes only but by the combination of all of the possible outcomes of the decision. For example, with a low selection ratio a number of potential successes might also be turned down (false negatives). Taylor-Russell tables do not provide a complete picture of the effects of the base rate or the selection ratio on all the outcomes of a decision; the selection ratio, the base rate, and the validity of the test interact in a complex way to affect all the outcomes of decisions. Nevertheless, the Taylor-Russell tables do provide important information for interpreting validity coefficients.

Tests and the Bottom Line—Applications of Utility Theory

Throughout this book we focus on the impact of psychological tests on decisions that must be made about individuals. The principal justification for the use of psychological tests is that they allow better decisions to be made than would be made without tests. The question is, how much better? Utility theory provides a method of answering this question.

According to utility theory, correct decisions (true positives, true negatives) are valued, while incorrect decisions (false positives, false negatives) are to be avoided, or have a negative value. Utility theory suggests that two things must be known before the impact of psychological tests can be assessed: (a) how many additional correct decisions will result if tests are used and (b) how much value is placed on good decisions. Methods described earlier in this chapter can be applied to the first question—if the base rate, selection ratio, and validity coefficient are known, the effect of tests on decisions can be determined easily. The trick, then, is to determine the value of outcomes.

One way to determine value is to make a judgment as to what will be gained by making a good decision instead of a bad one. This approach has been widely used in personnel selection and has provided estimates of the financial impact of tests on selection decisions. For example, it is possible to estimate, in dollar terms, the value of the products, goods, and services provided by a superior worker, an average worker, and a poor worker. These estimates can be used to determine the standard deviation of job performance, measured in dollar terms (Judicsch, Schmidt & Mount, 1992). Since a valid selection test will lead to an increase in the number of superior workers and a decrease in the number of poor workers hired, it is possible to use judgments of the value of performance to provide estimates of the financial impact of valid selection decisions.

Utility theory provides a method for estimating, in dollar terms, the gain (per year) in productivity that will result if valid tests are used in personnel selection. This gain is estimated by

Productivity gain = Kr_{xy} SD_y Z_s

K number of persons selected

r_{xy} validity coefficient

SD_y standard deviation of the criterion

$\bar{Z}_s$ average (standard) test score among those selected

Utility estimates have suggested that the use of psychological tests in areas such as personnel selection could lead to substantial gains in performance. For example, Hunter and Hunter (1984) estimate that the use of cognitive ability tests in selection for entry-level federal jobs could lead to a productivity gain of over $15 billion per year. Hunter and Schmidt (1982) estimate that the nationwide gain associated with the use of psychological tests in personnel selection could exceed $80 billion per year.4

There are two problems with utility equations such as Formula 8-4. First, they typically overestimate increases in utility (Cronshaw & Alexander. 1985; Murphy, 1986). Second, they ignore the costs associated with incorrect rejection decisions (false negatives; Boudreau, 1991). However, application of this branch of utility theory has greatly increased our understanding of the impact of psychological tests on the quality of decisions made daily by organizations and institutions.

Above mentioned formula suggests that several factors affect the overall quality of decisions. First, of course, is the validity of the test (r_{xy}). Second is the standard deviation of the criterion (SD_y); the larger this standard deviation, the larger the potential gain associated with the use of tests. The rationale here is straightforward. A large standard deviation indicates substantial differences in criterion scores; a small standard deviation indicates that everyone performs at pretty much the same level. If individual differences in performance are large, the quality of selection decisions makes a great deal of difference, but if everyone performs at about the same level, decisions will not affect criterion scores greatly. Finally, the average test score among those selected (Z_s) affects utility. The organization that is able to attract the "cream of the crop" among applicants gains more than another organization that is able to attract only mediocre applicants (Murphy, 1986). The utility score given by above mentioned formula represents a complex combination of the characteristics of the applicant pool, the selection ratio, and the decisions of individual applicants.

Lesson 07

NORMS

The Nature of a Score

Johnny got a score of 15 on his spelling test. What docs this score mean, and how should we interpret it?

Standing alone, the number has no meaning at all and is completely uninterruptible. At the most superficial level, we do not even know whether this number represents a perfect score of 15 out of 15 or a very low percentage of the possible score, such as 15 out of 50. Even if we do know that the score is 15 out of 20, or 75%, what then?

Consider the two 20-word spelling tests in Table 7.1. A score of 15 on Test A would have a vastly different meaning from the same score on Test B. A person who gets only 15 correct on Test A would not be outstanding in a second- or third-grade class. Have a few friends or classmates lake Test B. You will probably find that not many of them can spell 15 of these words correctly. When this test was given to a class of graduate students, only 22% spelled 15 or more of the words correctly. A score of 15 on Test 13 is a good score among graduate students of education or psychology.

As it stands, then, a score of 15 words correct, or even of 75%, has no direct meaning or significance. The score has meaning only when we have some standard with which to compare it.

Table 7.1 Two 20-Word Spelling Tests

Test A	Test B
bar	baroque
cat	catarrh
form	formaldehyde
jar	jardiniere
nap	naphtha
dish	discernible
fat	fatiguing
sack	sacrilegious
rich	ricochet
sit	citrus
feet	feasible
act	accommodation
rate	inaugurate
inch	insignia
rent	deterrent
lip	eucalyptus
air	questionnaire
rim	rhythm
must	ignoramus
red	accrued

Frames of Reference

The way that we derive meaning from a test score depends on the context or frame of reference in which we wish to interpret it. This frame of reference may be described using three basic dimensions. First, there is what we might call a temporal dimension: Is the focus of our concern what a person can do now or what that person is likely to do in the future? Are we interested in describing the current state or in forecasting the future?

A second dimension involves the contrast between what people can do and what they would like to do or would normally do. When we assess a person's capacity, we determine maximum performance, and when we ask about a person's preferences or habits, we assess typical performance. Maximum performance implies a set of tasks that can be judged for correctness; there is a "right" answer. With typical performance there is not a right answer, but we may ask whether one individual's responses are like those of most people or are unusual in some way.

A third dimension is the nature of the standard against which we compare a person's behavior. In some cases, the test itself may provide the standard; in some cases, it is the person's own behavior in other situations or on other tests that provides the standard; and in still other instances, it is the person's behavior in comparison with that of other people. Thus, a given measurement is interpreted as being either oriented in the present or oriented in the future; as measuring either maximum or typical performance; and as relating the person's performance to a standard defined by the test itself, to the person's own scores on this or other measures, or to the performance of other people.

Many instructional decisions in schools call for information about what a student or group of students can do now. Walter is making a good many mistakes in his oral reading. To develop an instructional strategy that will help him overcome this difficulty, we need to determine the cause of his problem. One question we might ask is whether he can match words with their initial consonant sounds. A brief test focused on this specific skill, perhaps presented by the teacher to Walter individually while the other students work on other tasks, can help to determine whether a deficiency in this particular skill is part of Walter's problem. How many children in Walter's class have master)' of the rule on capitalizing proper nouns? A focused test such as the one in Table 7.2 can provide evidence to guide a decision on whether further teaching of this skill is needed. At a broader level, we may ask whether the current program in mathematics in Centerville is producing satisfactory achievement. A survey mathematics test with national or regional norms can permit a comparison of Centerville's students with students in the rest of the country, and this comparison can be combined with other information about Centerville's students and its schools to make a decision on whether progress is satisfactory.

Whenever we ask questions about how much a person can do, there is also the issue of the purpose of our evaluation. There are two fundamental purposes for evaluating capacity in an educational context. One is to reach a summary statement of the person's accomplishments to date, such as teachers do at the end of each marking period. Evaluation for this purpose is called ***summative evaluation.*** It provides a summary of student achievement. By contrast, teachers are often interested in using tests to determine their students' strengths and weaknesses, the areas where they are doing well and those where they are doing poorly. Assessment for this purpose, to guide future instruction, is called ***formative evaluation***. Test results are used to inform or to shape the course of instruction.

Table 7.2 A Focused Test

Test on Capitalizing Proper Nouns

Directions: Read the paragraph. The punctuation is correct, and the words that begin a sentence have been capitalized. No other words have been capitalized. Some need to be. Draw a line under *each word* that should begin with a capital.

We saw Mary yesterday. She said she had gone to Chicago, Illinois, to see her aunt Helen. Her aunt took her for a drive along the shore of Lake Michigan. On the way they passed the Conrad Hilton hotel, where Mary's uncle Joseph works. Mary said she had enjoyed the trip, but she was glad to be back home with her own friends.

The type of maximum performance test that describes what a person has learned to do is called an achievement test. The oral reading test given to Walter, the capitalization test in Table 7.2, and the mathematics test given to the students in Cen-terville are illustrations of sharply contrasting types of achievement tests. The test on initial consonant sounds is concerned with mastery of one specific skill by one student, and no question is raised as to whether Walter's skill in this area is better or worse than that of any other student. The only question is, can he perform this task well enough so that we can rule out this skill as a cause of his difficulty with oral reading?

Similarly, Walter's teacher is concerned with the level of mastery, within this class, of a specific skill in English usage. Tests concerned with level of mastery of such defined skills are often called domain-referenced or criterion-referenced tests because the focus is solely on reaching a standard of performance on a specific skill called for by the test exercises. The test itself and the domain of content it represents provide the standard. Many, perhaps most, assessments needed for instructional decisions are of this sort.

We may contrast these tests with the mathematics survey test given to appraise mathematics achievement in Centerville. Here, the concern is whether Centerville's students are showing satisfactory achievement ***when compared with the students in other towns and school systems like Centerville.*** Performance is evaluated not in relation to the set of tasks per se, but in relation to the performance of some more general reference group. A test used in this way is spoken of as a ***norm-referenced*** test, because the quality of the performance is defined by comparison with the behavior of others. A norm-referenced test may appropriately be used in many situations calling for curricular, guidance, or research decisions. Occasionally throughout this book, we will compare and contrast criterion-referenced and norm-referenced achievement tests with respect to their construction, desired characteristics, and use.

Some decisions that we need to make require information on what a person can learn to do. Will Helen be able to master the techniques of computer programming? How readily will Richard assimilate calculus? Selection and placement decisions typically involve predictions about future learning or performance, based on present characteristics of the individual. A test that is used in this way as a. predictor of future learning is called an aptitude test. Aptitude tests are usually norm referenced.

There are also situations where our decision calls for an estimate of what a person ***is likely to do.*** The selection of bus drivers, police officers, and candidates for many other jobs is best made with an eye to aspects of the person's personality or temperament. We would not want to select someone with a high level of aggression to be driving a large vehicle on confined city streets. Nor would we want people who have difficulty controlling their tempers serving as keepers of the peace. A measure of typical performance can serve as a useful aid in such situations, and these measures usually are also norm referenced.

Note that some of the most effective predictors of future learning are measures of past learning. Thus, for both computer programming and calculus, an effective predictor might be a test measuring competence in high school algebra. Such a test would measure previously learned knowledge and skill, but we would be using that achievement measure to predict future learning. Any test, whatever it is called, assesses a person's present characteristics. We cannot directly measure a person's hypothetical "native" or "inborn" qualities. All we can measure is what that person is able and willing to do in the here and now. That information may then be used to evaluate past learning, as when an algebra test is used to decide whether Roxanne should get an A in her algebra course, or to predict future learning, as when a counselor must decide whether Roxanne has a reasonable probability of successfully completing calculus. The distinction between an aptitude and an achievement test often lies more in the purpose for which the test results are used than in the nature or content of the test itself.

Domains in Criterion- and Norm-Referenced Tests

It is important to realize that all achievement tests relate to a specified domain of content. The mathematics survey rest covers a fairly broad array of topics, while the test' on the rules for capitalization is restricted to a narrowly defined set of behaviors. Thus, it is not really appropriate to differentiate between criterion-referenced and norm-referenced tests by saying that the former derive their meaning from a precisely / specified domain, while the latter do not. A well-constructed, norm-referenced test will represent a very carefully defined domain, but the domain is generally more diverse .than that of a criterion-referenced test, and has only a small number of items covering a given topic or instructional objective. The criterion-referenced test will represent a narrowly defined domain and will therefore cover its referent content more thoroughly than will a norm-referenced test of the same length.

There is a second dimension to using information from an achievement test. In addition to the traditional distinction between criterion-referenced and norm-referenced tests on the breadth of the domain they cover, another dimension relates to the way that the level, or altitude, of performance is represented or used in reaching decisions. A test score from either type of test gets its content meaning from the domain of content that the test represents, but the kind of inference that a teacher or counselor draws from the score can be either absolute or relative, The teacher makes a judgment on the basis of the test score. If the judgment is that when a student or group of students have gained a particular level of proficiency with respect to that content they have mastered the material, then the judgment is an absolute, ***mastery/nonmastery*** one. The decision reached is either that the students have mastered the material or that they have not; degree of mastery is not an issue. Decisions of this type are called ***mastery*** decisions.

The usual definition of a criterion-referenced test is a test that covers a narrow domain and is used for master)' decisions.

By contrast, teachers can also use tests to judge relative achievement of objectives. Relative mastery involves estimating the percentage of the domain that students have mastered. For example, the teacher may decide that students have mastered an objective relating to spelling when they can spell correctly 19 out of 20 words from the domain. But the same teacher might use the information that the average student got a score of ***14*** on the spelling test to indicate that the students had achieved about 70% master)' of the domain. We refer to decisions of this kind as ***relative achievement*** decisions, but the frame of reference is still the domain of content without regard to the performance of anyone other than the current examinees.

The typical norm-referenced test uses neither of these ways to represent the level of performance. Rather, level is referenced to a larger group called a ***norm group,*** or norm sample.

A normative interpretation of a score could lead to the conclusion that the individual was performing at a **very** high level compared with an appropriate reference group, but the same performance might fall far below mastery from the criterion-referenced perspective. Conversely, a ninth grader who has achieved mastery of multiplication facts at the level of 95% accuracy ordinarily would not show a high level of performance when compared with other ninth graders.

CRITERION-REFERENCED EVALUATION

We can approach the problem of a frame of reference for interpreting test results from the two rather different points of view mentioned earlier. One, criterion-referenced evaluation, discussed here, focuses on the tasks themselves, and the other, norm-referenced testing, on the performance of typical people. Consider the 20 spelling words in Test A of Table 7.1 If we knew that these had been chosen from the words taught in a third-grade spelling program and if we had agreed on some grounds (at this point unspecified) that 80% correct represented an acceptable standard for performance in spelling when words are presented by dictation, with illustrative sentences, then we could interpret Ellen's score of 18 correct on the test as indicating that she had reached the criterion of mastery of the words taught in third-grade spelling and Peter's score of 12 correct as indicating that he had not.

Here, we have test content selected from a narrowly defined domain and we have a mastery test interpretation. The test is criterion referenced in that (1) the tasks are drawn from and related to a specific instructional domain, (2) the form of presentation of the tasks and the response to them is set in accordance with the defined objective, and (3) a level of performance acceptable for master)', with which the performance of each student is compared, is defined in advance. That is, criterion-referenced tests relate to a carefully defined domain of content, they focus on achievement of behavioral objectives, and the results are often (but not necessarily) used for mastery judgments.

The "mastery" frame of reference is an appropriate one for some types of educational decisions. For example, decisions on what materials and methods should be used for additional instruction in spelling with Hllen and Peter might revolve around the question of whether they had reached the specified criterion of mastery of the third-grade spelling words. More crucially, in a sequential subject such as mathematics, the decision on whether to begin a unit involving borrowing in subtraction might depend on whether students had reached a criterion of mastery on a 'test of two-place subtraction that did not require borrowing.

By contrast, teachers also use tests to judge the relative achievement of objectives. Relative mastery may involve estimating the percentage of a domain that the students have mastered.

Although the two topics of domain referencing of test content and a mastery/nonmastery decision about achievement historically have been linked, it is important to realize that the)' are quite different and independent ideas that have recently come to be treated together. It is also important to realize that both exist in a sociopolitical context that invests them with normative meaning. What, for example, should a third grader be expected to know about multiplication? The answer to this question depends on what is expected of second and fourth graders, and these expectations put norm-referenced boundaries on what is taught in the third grade. Professional judgment and many years of experience combine to define the reasonable domain of content. A test is then constructed to represent this content.

Given a test that is designed to represent a particular domain of content, the scores from that test may be interpreted strictly with respect to that content, or they may be interpreted in a normative framework by comparing one person's performance with that of others. Domain-referenced interpretation means that the degree of achievement is assessed relative to the test itself and the instructional objectives that gave rise to the test. The evaluation may result in a dichotomous judgment that the person has mastered the material and is ready for further instruction, for certification or licensure, or for whatever decision is the object of the measurement. 0>r, the evaluation may result in a judgment of degree of mastery. The latter approximates what teachers do when they assign grades, while the former is similar to a pass/fail decision or a decision to begin new material.

For the group of tests that are typically called criterion referenced, the standard, then, is provided by the definition of the specific objectives that the test is designed to measure. When the type of decision to be made is a mastery decision, this description of the content, together with the level of performance that the teacher, school, or school system has agreed on as representing an acceptable level of mastery of that objective, provide an absolute standard. Thus, the illustrative content-referenced test of capitalization of proper nouns in Table 7.2 is presumed to provide a representative sample of tasks calling for this specific competence. If we accept the sample of tasks as representative and if we agree that 80% accuracy in performing this task is the minimum acceptable performance, then a score of 10 out of 13 words correctly underlined defines the standard in an absolute sense.

Even the dichotomous or mastery judgment is made in a sociopolitical, hence normative, context. The teacher or school has to decide what constitutes master)', and there are some not-so-subtle social pressures that affect such decisions. Most teachers define the level of achievement necessary for mastery in such a way that an "appropriate" number of students are identified as masters. In practice, this means that over a period of time the teacher develops a fairly accurate idea of how typical students will perform on his or her tests covering a course of instruction. The tests, grading practices, or passing standards are adjusted so that, in the long run, the right number of students pass, which makes the setting of passing standards basically a normative decision! (See Shepard [1984) for a discussion of setting standards in criterion-referenced testing.)

In the usual classroom test used for summative evaluation, such a standard operates indirectly and imperfectly, partly through the teacher's choice of tasks to make up the test and partly through his or her standards for evaluating the responses. Thus, to make up their tests, teachers pick tasks that they consider appropriate to represent the learning of their students. No conscientious teacher would give spelling Test A to an ordinary high school group or Test B to third graders. When the responses vary in quality, as in essay examinations, teachers set standards for grading that correspond to what they consider is reasonable to expect from students like theirs. Quite different answers to the question "What were the causes of the War of 181 2?" would be expected from a ninth grader and from a college history major.

However, the inner standard of the individual teacher tends to be subjective and unstable. Furthermore, it provides no basis for comparing different classes or different areas of ability. Such a yardstick can give no answers to such questions as, Are the children in School A better in reading than those in School B? Is Man' better in reading than in mathematics? Is Johnny doing as well in algebra as most ninth graders? We need some broader, more uniform, objective, and stable standard of reference if we are to be able to interpret those psychological and educational measurements that undertake to appraise some trait or to survey competence in some broad area of the school curriculum. Most of this chapter is devoted to describing and evaluating several normative reference frames that have been used to give a standard meaning to test scores.

NORM -REFERENCED EVALUATION

The other frame of reference for interpreting test performance is based not on a somewhat arbitral-)' standard defined by a particular selection of content and interpreted as representing master)-of that content domain but rather is based on the performance of other people. This represents a norm-referenced interpretation. Thus, the scores of Ellen and Peter can be viewed in relation to the performance of a large performance group of typical third graders or of students in different school grades. Their performance is viewed not in terms of mastery versus nonmastery or in terms of relative master)' of the subject matter, but instead as above average. Or below average are sought to refine that scale of relative performance so that all degree of excellence can be expressed in quantitative terms.

In seeking a scale to represent degrees of excellence, we would like to report results in units that have the following properties:

1. Uniform meaning from test to test, so that a basis of comparison is provided through which we may compare different tests—for example, different reading tests, a reading test with an arithmetic test, or an achievement test with a scholastic aptitude test

2. Units of uniform size, so that a gain of 10 points on one part of the scale signifies the same thing as a gain of 10 points on any other part of the scale
3. A true-zero point of ***just none of*** the quality in question, so that we can legitimately think of scores as representing ***twice as much as*** or ***two-thirds as much as***

The different types of norm-referenced scales that have been developed for tests represent marked progress toward the first two of these objectives. The third can probably never be reached for the traits with which we are concerned in psychological and educational measurement. We can put five 1-lb loaves of bread on one side of a pair of scales, and they will balance the contents of one 5-Ib bag of flour placed on the other side. "No weight" is ***truly*** "no weight," and units of weight can be added so that 2 lb is twice 1 lb. but we do not have that type of zero point or that type of adding in the case of educational and psychological measurement. If you put together two below-average students, you will not get a genius, and a pair of bad spellers cannot jointly win a spelling bee. In some cases, this deficit is the result of the particular way we have chosen to measure the trait, but for many psychological and educational traits, the deficit is a result of how we conceptualize the trait itself.

Basically, a raw point score on a test is given normative meaning only by referring it to some type of group or groups. A score on the typical test is not high or low or good or bad in any absolute sense; it is higher or lower or better or worse than other scores. There are two general ways that we may relate one person's score to a more general framework. One way is to compare the person with a graded series of groups and see which one he or she matches. Each group in the series usually represents a particular school grade or a particular chronological age. The other way is to find where in a particular group the person falls in terms of the percentage of the group surpassed or in terms of position relative to the group's mean and standard deviation. Thus, we find four main patterns for interpreting the score of an individual. These are shown schematically in Table 33. We shall consider each in turn, evaluating its advantages and disadvantages.

Table 7.3 Main Types of Norms for Educational and Psychological Tests

Type of Norm	**Type of Comparison**	**Type of Group**
Grade norms	Individual matched to group whose performance he or she equals	Successive grade groups
Age norms	Same as above	Successive age groups
Percentile norms	Percentage of group surpassed	Single age or grade group
Standard score norms	Number of standard deviations individual falls above or below average of group	Same as above

Grade norms

For any trait that shows a progressive and relatively uniform increase from one school grade to the next, we can prepare a set of grade norms. The norm for any grade, in this sense, is the average score obtained by individuals in the grade. Because school participation and the related cognitive growth are both more or less continuous, grade norms typically are expressed with one decimal place. The whole number gives the grade, and the decimal is the month within grade. Thus, a grade equivalent of ***5.4*** is read as performance corresponding to that of the average child in the fourth month of fifth grade.

In simplest outline, the process of establishing grade norms involves giving the test to a representative sample of pupils in each of a number of consecutive grades, calculating the average score at each level, and then establishing *grade equivalents* for the in-between scores. Thus, a reading comprehension test, such as that from the Iowa Tests of Basic Skills (ITBS)—Form J, Level 9, might be given in November to pupils in grades 2. 3, ***4,*** and 5, with the following results.

Grade Level	Average Raw Score
2.3	13
3.3	22
4.3	31
5.3	37

The testing establishes grade equivalents for raw scores of 13, 22, 31, and 37. However, grade equivalents are also needed for the in-between scores. These are usually determined arithmetically by interpolation, although sometimes intermediate points may be established by actually testing at other times during the school year. After interpolation, we have the following table.*

Raw Score	Grade Equivalent	Raw Score	Grade Equivalent
10	1.9	24	3.5
11	2.0	25	3.6
12	2.2	26	3.7
13	2.3	27	3.8
14	2.5	28	3.9
15	2.6	29	4.0
16	2.8	30	4.1
17	2.9	31	4.3
18	3.0	32	4.4
19	3.1	33	4.5
20	3.2	34	4.7
21	3.2	35	4.9
22	3.3	36	5.1
23	3.4	37	5.3

Because raw scores on this particular test can range from 0 to 49, some way is needed to establish grade equivalents for the more extreme scores. Establishing such grade equivalents is often done by equating scores on the level of the test on which we are working with scores from lower and higher levels of the same test series, forms that have been given to earlier and later grades. In this way, grade equivalents may be extended down as low as the first month of kindergarten (denoted K.l) and up as high as the end of the first year in college (denoted 13-9), and a complete table to translate raw scores to grade equivalents can be prepared. (The reading test of this particular edition of the ITBS actually is a multilevel test that uses six overlapping sets of passages and items in a single booklet. In this way, some of the same items are used for three different levels of the test, and the projection of grade equivalents is simplified and made more accurate.)

If Jennifer got a raw score of 28 on this test, it would give her a grade equivalent of 3-9, and this score could be translated as "performing as well as the average child who has completed 9 months of third grade." Such an interpretation has the advantage of connecting the test score to familiar milestones of educational development. However, this seductively simple interpretation of a child's performance has a number of drawbacks as well.

A first major question about grade norms is whether we can think of them its"" providing precisely on even approximately equal units. In what sense is the growth in paragraph reading from grade 3.2 to 4.2 equal to the growth from grade 6.2 to 7.2? Grounds for assuming equality are clearly tenuous. When the skill is one that has been taught throughout the school years, there may be some reason to expect a year's learning at one level to be about equal to a year's learning at some other. And there is evidence that during elementary school (and possibly junior high), grade-score units are near enough to equal to be serviceable. However, even in this range and for areas where instruction has been continuous, the equality is only approximate. If, on the other hand, we are concerned with a subject like Spanish, in which instruction typically does not begin until secondary school, or in something like biology, for which instruction is concentrated in a single grade, grade equivalents become almost completely meaningless. In addition, instruction in man)' skills, such as the basic skills in reading in arithmetic computation, tapers off and largely stops by high school, so grade units have little or no meaning at this level. For this reason many achievement batteries show a grade equivalent of 10.0+ or 11.0+ as representing the whole upper range of scores. When grade equivalents such as 12.5 are reported, these do not really represent the average performance of students tested in the middle of the 12th grade, but rather, they are an artificial and fictitious extrapolation of the score scale, used to provide some converted score to be reported for the most capable eighth and ninth graders.

A further note of caution must be introduced with respect to the interpretation of grade norms. Consider a bright and educationally advanced child in the third grade. Suppose we find that on a standardized mathematics test this child gets a score with the grade equivalent of 5.9. This score does ***not*** mean that this child has a mastery of the mathematics taught in the fifth grade. The ***score*** is as high as that earned by the average child at the end of fifth grade, but this higher score almost certainly has been obtained in part by superior mastery of third-grade work. The average child falls well short of a perfect score on the topics that have been taught at his or her own grade level. The able child can get a number of additional points (and consequently a higher grade equivalent) merely by complete master)' of this "at-grade" material. This warning is worth remembering. The fact that a child has a grade equiv**alent** of 5.9 need not mean that the child is ready to move ahead into sixth grade work. The grade equivalent is only the reflection of a score and does not tell in what way that score was obtained. Reference to the content of the questions the child answered correctly would be needed to reach a judgment that the child had sufficient master)' of fifth-grade material to be able to move into the sixth grade. Thus, grade equivalents should not be used to make master)' decisions.

Finally, there is reason to question the comparability of grade equivalents from one school subject to another. Does being a year ahead (or behind) one's grade level in language usage represent the same amount of advancement (or retardation) as the same deviation in arithmetic concepts? A good deal of evidence exists, which we consider later in this chapter that it does not. Growth in different school subjects proceeds at different rates, depending on in-school emphasis and out-of-school learning. For this reason, the glib comparison of a pupil's grade equivalent in different school subjects can result in quite misleading conclusions.

To summarize, grade norms, which relate the performance of an individual to that of the average child at each grade level, are useful primarily in providing a framework for interpreting the academic accomplishment of children in the elementary· For this purpose, they are relatively convenient and meaningful, even though we cannot place great confidence in the equality of grade units or their exact equivalence from one subject to another.

Grade norms are relatively easy to determine because they are based on the administrative groups already established in the school organization. In the directly academic areas of achievement, the concept of grade level is perhaps more meaningful than is age level, for it is in relation to grade placement that a child's performance is likely to be interpreted and acted on. Outside the school setting, grade norms have little meaning.

Developmental Standard Scores

We have noted several problems with grade equivalents as normative representations of a child's performance, particularly that there is an implicit assumption that the amount of growth in the ability being tested is equal from one year to the next. Because this assumption clearly is violated for many abilities, test publishers have developed a type of score scale that is anchored to school grades but provides a better approximation to an equal interval scale, the *Developmental Standard Score Scale*.

Developmental standard scores (DSSs) are based on normalized score distributions within each grade (see the discussion of normalizing transformations later in this chapter). Scale values for two grades are chosen arbitrarily to define the scale metric, and the within-grade means and standard deviations are then used to locate other grade equivalents on this scale. For example, the Iowa Tests of Basic Skills authors have chosen to fix a scale value of 200 as equivalent to the median performance of fourth graders and a value of 250 for eighth graders tested in the spring. The relationship between grade equivalents and DSSs reported in the test manual is as follows:

Grade	K	1	2	3	4	5	6	7	8	9	10
DSS	130	150	168	185	200	214	227	239	250	260	268

One fact is quite clear from comparing grade equivalents and DSSs, equal changes in grade equivalents do not correspond to equal changes in DSS. The DSS scale is constructed to have equal intervals (a 10-unit change has the same meaning everywhere on the scale). The comparison shows that there is a bigger change from year to year during the early years of school than there is in later years, 18 points from first to second grade, 10 points from eighth to ninth.

The main drawback of DSSs is that, unlike grade equivalents, they have no inherent meaning. The values chosen for the anchor points are quite arbitrary. Meaning is given only by their relationship to the grade equivalent scale. It would be appropriate, for example, to say that a student who received a DSS of 255 was performing at the level of students in about December of their ninth-grade year. Because of their complexity and lack of obvious meaning, developmental standard scores are hard to interpret correctly and should be used with caution, even though they are reported by many test publishers.

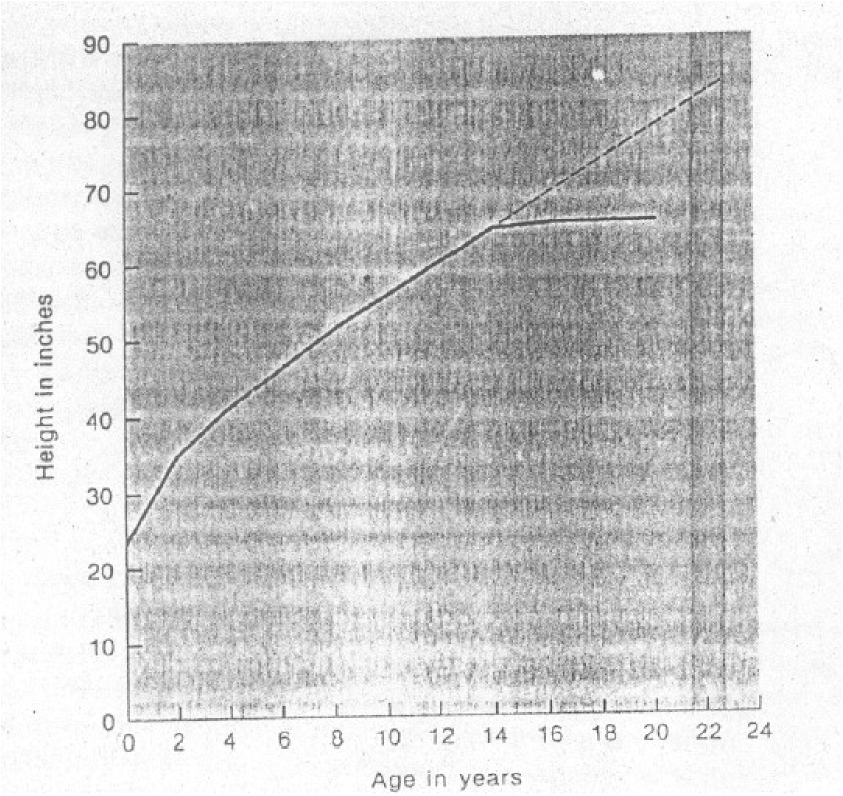
Age Norms

If a trait is one that may be expected to show continuous and relatively uniform growth with age, it may be appropriate to convert the score into an ***age score,*** or ***age equivalent,*** as a type of common score scale. During childhood we can observe continuous growth in height and weight, in various indices of anatomical maturity, and in a wide range of perceptual, motor, and cognitive performances. It makes a crude type of sense to describe an 8-year-old as being as tall as the average 10-year-old • and having the strength of grip of the average 9-year-old, as well as the speaking vocabulary of the average 6-year-old. In the early development of intelligence and aptitude tests, raw scores were typically converted into age equivalents, and the term "mental age" was added to the vocabulary of the mental tester and the general public alike, with occasionally unfortunate results.

An age equivalent is, of course, the average score earned by individuals of a given age and is obtained by testing representative samples of 8-ycar-olds, 9 years old, 10 years old and so fort. In this respect, it parallels the grade equivalent described trr7rTe-p7evious section. And, as in the case of grade equivalents, a major issue is whether we can reasonably think of a year's growth as representing a standard and uniform unit. Is growth from age 5 to age 6 equal to growth from age 10 to age 11? And is growth in any 1 year equivalent to growth in any other year on our scale? As we move up the age scale, we soon reach a point where we see that the year's growth unit is clearly not appropriate. There comes a point, some time in the teens or early 20s, when growth in almost any trait that we can measure slows down and finally stops. In Figure 7.1, which illustrates the normal growth of height for girls, the slowdown takes place quite abruptly after age 14. A year's growth after 14 seems clearly to be much less than a year's growth earlier on the scale. At about age 14 or 15, the concept of height-age ceases to have any meaning. The same problem of a flattening growth curve is found, varying only in the age at which it occurs—and in abruptness, for any trait that we can measure.

The problem introduced by the flattening growth curve is most apparent when we consider the individual who falls far above average. What age equivalent shall we

Figure 7.1
Girls' age norms for height



assign to a girl who is 5 ft 10 in. (70 in.) tall? The average woman never gets that tall at any age. If we are to assign any age value, we must invent some hypothetical extension of our growth curve, such as the dashed line in Figure 7.1. This line assumes that growth after 14 continues at about the same rate that was typical up to age 14. On this extrapolated curve, the height of 5 ft 10 in. would be assigned a height-age of about 16 years and 6 months. Rut this is a completely artificial and arbitrary age equivalent. It does ***not*** correspond to the average height of 16 1/2-year-olds. It does not correspond to the average height at ***any*** age. It merely signifies "taller than average." Unfortunately, there is no cue to be gotten from these extrapolated age equivalents that suggests their arbitrary nature.

Age norms, which are bused on the characteristics of the average person at each age level, provide a readily comprehended framework for interpreting the status of a particular individual. However the equality of age units is open to serious -question, and as one goes up to adolescence and adulthood, age ceases to have any meaning as a unit in which to express level of performance. Age norm are most appropriate for infancy and childhood and for characteristics that grow as a part of the general development of the individual, such as height, weight, or dentition. General mental development, such as the cognitive characteristics embodied in the concept of mental age, show a sufficiently universal pattern to be useful normative indicators of status, but, in general, age norms should not be used for cognitive characteristics beyond the elementary school years, because the patterns of growth of these functions depend too heavily on formal school experiences or haw not been found to show the pattern of growth necessary for age norms to be appropriate.

Percentile Norms

We have just seen that in the case of age and grade norms, meaning is given to the individual's score by determining the age or grade group in which the person would be exactly average. But, often such a comparison group is inappropriate or some other group would be more useful. For example, we are frequently concerned with the performance of people who are no longer in the elementary grades where grade norms have meaning. Or, we may be interested in personality or attitude characteristics for which age or grade norms are wholly unusable. Or, the type of information that we seek may require that we specific the group of interest more narrowly than is practical for age or grade norms For example, we may be interested in people who are all the same age or are all in the same grade.

Each individual belongs to many different groups. An individual who is 18 years old belongs to some of the following groups, but not to others: all 18-year-olds, 18-year-olds in the 12th grade, 18-ycar-okls applying to college, 18-year-olds not applying to college. 18 years old applying to Ivy League colleges, 18-year-olds attending public (or parochial) schools and 18-year-olds attending school in California. For some purposes it is desirable or necessary to define the comparison group more narrow!}' than is possible with grade or age norms. One universally applicable system of norms is the percentile norm system.

The typical percentile norm, or ***percentile rank*** uses the same information that we used to compute percentiles in Chapter 2, hut the procedure is slightly different. ***Percentile ranks are calculated to correspond to obtainable score values.*** If a test has 10 items, it can yield 11 different raw scores, the whole scores from zero to 10. There are only 1 possible values that percentile ranks could assume for this test, one for each obtainable score, but it would still be possible to calculate any number of percentiles. For example, one could compute, using the procedures described in Chapter 2, the 67.4th percentile as well as the 67th and 68th. But only the 11 obtainable scores would have corresponding percentile ranks. The normative interpretation of test scores more often uses percentile ranks than percentiles, because test results come in a limited number of whole score units.

The procedure for determining percentile ranks starts with a frequency distribution such as the one shown in Table 7.4. We assume, as we did for percentiles, that the underlying trait that the test measures is continuous, that each observable score falls at the midpoint of an interval on this continuum, and that the people who obtained a given raw score are spread throughout the interval. Because each raw score-falls at the middle of an interval, half of the people in the interval are considered to be below the midpoint and half above. Even if only one person falls in a particular interval, we assume that half of that person falls above the midpoint of the interval and half falls below.

To find the percentile rank of a raw score, we count the number of people who are below the score and divide by the total number of people. The number of people below a raw score value includes all of the people who obtained lower scores plus half of the people who received the score in question (the latter group because they are assumed to be in the bottom half of the interval and, therefore, below the raw score). For example, to calculate the percentile rank of a raw score of 4 in Table 7.4, we would take the eight people who got scores below 4 and half of the seven people at 4. The result is (8 + 3 5)/60 = 11.5/60 = 0.1917. In reporting percentile ranks it is conventional to round the answer to two places and multiply by 100 to remove the decimal point except at the extremes of the scale. The percentile rank that corresponds to a raw score of 4 is therefore 19.

Table: 7.4 Determining Percentile Ranks for a 10-Item Test

Raw Score	Frequency	Cumulative Frequency	Percentile Rank
10	1	60	99
9	3	59	96
8	5	56	89
7	12	51	75
6	15	39	52
5	9	24	32
4	7	15	19
3	4	8	10
2	2	4	5
1	1	2	2
0	1	1	1

The major procedural difference between calculating percentiles such as the median and percentile ranks such as those in Table 7.4 is where one starts. To calculate percentiles, we specify a percent of interest, such as the 25th or 60th, and determine the answer, a point on the score scale, by the procedures described in Chapter 2. The values that correspond to these percentages need not be, and seldom are, whole points of score. When calculating percentile ranks, we start with a point on the score scale, an obtainable score value, and find as the answer the percentage of the group that falls below the chosen point of score.

Percentile norms are very widely adaptable and applicable. They can be used wherever an appropriate normative group can be obtained to serve as a yardstick. They are appropriate for young and old and for educational and industrial situations. To surpass 90% of a reference comparison group signifies a comparable degree of excellence whether the function being measured is how rapidly one can solve simultaneous equations or how far one can spit. Percentile norms are widely used and their meaning is readily understood. Were it not for the two points we next consider, they would provide a framework very nearly ideal for interpreting test scores.

The first issue that faces us in the case of percentile norms is specifying the norming group. On what type of group should the norms be based? Clearly, we will need different norm groups for different ages and grades in the population. A 9-year-old must be evaluated in terms of 9-year-old norms; a sixth grader, in terms of sixth-grade norms; an applicant for a job as real estate agent, in terms of norms for real estate agent applicants. The appropriate norm group is in every case the relevant group to which the individual belongs and in terms of which his or her status is to be evaluated. It makes no sense, for example, to evaluate the performance of medical school applicants on a biology test-by comparing their scores with norms based on high school seniors. If the test is to be used by a medical school, the user must find or develop norms for medical school applicants.

Hence, If percentile norms are to be used, multiple sets of norms are usually needed. There must be norms appropriate for each distinct type of group or situation in which the test is to be used. This requirement is recognized by the better test publishers, and they provide norms not only for age and grade groups but also for special types of educational or occupational populations. However, there are limits to the number of distinct populations for which a test publisher can produce norms, so published percentile norms will often need to be supplemented by the test user, who can build up norm groups particularly suited to local needs. Thus, a given school system will often find it valuable to develop local percentile norms for its own pupils. (Most test publishers will assist school districts with the development of local norms.) Such norms will permit scores for individual pupils to be interpreted in relation to the local group, a comparison that may be more significant for local decisions than is comparison with national, regional, or state norms. Likewise, an employer who uses a test with a particular category of job applicant may well find it useful to accumulate results over a period of time and prepare norms for this particular group of people. These strictly local norms will greatly facilitate evaluating a new applicant.

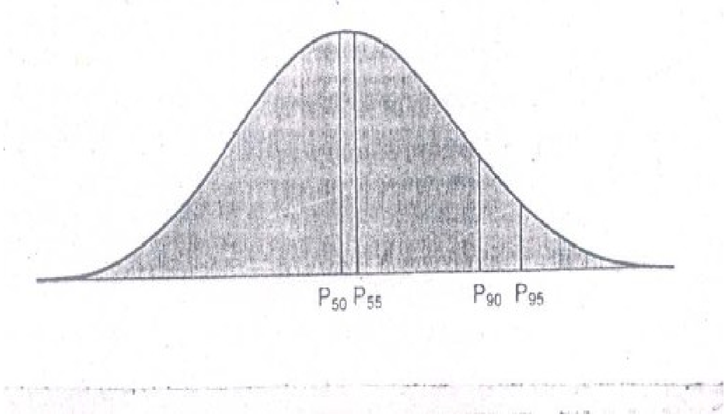
Thus, the possibility of specifying many different norm groups for different uses of a test constitutes both a problem, in the sense of greater complexity, and a strength, in that more accurate comparisons can be made.

The second percentile-norm issue relates to the question of equality of units. Can we think of five percentile points as representing the same amount of the trait throughout the percentile scale? Is the difference between the 50th and 55th percentile equivalent to the difference between the 90th and 95th? To answer this question, we must notice the way in which the test scores for a group of people usually pile up. We saw one histogram of scores in Figure 2.1 of Chapter 2. This picture is fairly representative of the way the scores fall in many situations. Cases pile up around the middle score values and tail off at either end. The ideal model of this type of score distribution, the normal curve, was also considered in connection with the standard deviation in Chapter 2 (see Table 2.5 and Figure 2.7) and is shown in Figure 7.2. The exact normal curve is an idealized mathematical model, but many types of test results distribute themselves in a manner that approximates a normal curve. You will notice the piling up of most cases in the middle, the tailing off at both ends, and the generally symmetrical pattern.

In Figure 7.2, four points have been marked: the 50th, 55th, 90th, and 95th percentiles. The baseline represents a trait that has been measured in a scale with equal units. The units could be items correct on a test. Note that near the median, 5% of the cases (the 5% lying between the 50th and 55th percentiles) fall in a tall narrow pile. Toward the tail of the distribution, 5% of cases (the 5% between the 90th and 95th percentiles) make a relatively broad low bar. In the second instance, 5% of the cases spread out over a considerably wider range of the trait than in the first. The same number of percentile points corresponds to about three limes as much of the score scale when we are around the 90-95th percentiles as when we are near the median. The farther out in the tail we go, the more extreme the situation becomes.

Thus, percentile units are typically and systematically unequal, relative to the raw score units. The difference between being first or second in a group of 100 is many times as great as the difference between being 50th and 51st. Equal percentile differences do not, in general, represent equal differences in amount of the trait in question. Any interpretation of percentile ranks must take into account the fact that such a scale has been pulled out at both ends and squeezed in the middle. Mary, who falls at the 50th percentile in arithmetic and at the 55th in reading, shows a trivial difference in these two abilities, whereas Alice, with respective percentiles of 90 and 95, shows a larger difference.

Figure 7.2: normative curve showing selected percentile points



One of the consequences of this inequality of units in the percentile scale is that percentiles cannot he treated with many of the procedures of mathematics. For example, we cannot add two percentile ranks together and get a meaningful result. The sum or average of the percentiles of two raw scores will not yield the same result as determining the percentile rank of the sum or average of the two raw scores directly. A separate table of percentile equivalents would be needed for every combination of raw scores that we might wish to use.

Standard Scores

Because the units of a score system based on percentile ranks are so clearly unequal, we are led to look for some other unit that does have the same meaning throughout its whole range of values. ***Standard-score scales*** have been developed to serve this purpose.

In Chapter 2 we became acquainted with the standard deviation ***(SD)*** as a measure of the spread, or scatter, of a group of scores. The standard deviation is a function of the deviations of individual scores away from the mean. -Any score may be expressed in terms of the number of standard deviations it is away from the mean. The mean ***mathematics*** score for ninth graders on the Tests of Achievement and Proficiency is 24.1 and the standard deviation is 9.8, so a person who gets a score of 30 falls

$$30 - 24.1/9.8 = 0.60$$

SD units above the mean. A score of 15 would be 0.93 ***SD*** units ***below*** the mean. In standard deviation units, we would call these scores +0.60 and —0.93, respectively.

Scores that are reported as deviations away from the group mean in standard deviation units are called ***standard scores,*** or ***z*** scores. A ***z*** score can be found in any score distribution by first subtracting the groups mean ***(M)*** from the raw score ***(X)*** of interest and then dividing this deviation by the standard deviation:

$$z = (X - M)/SD$$

If this is done for every score in the original distribution, the new distribution of ***z*** scores will have a mean of zero, and the standard deviation of the new distribution will be 1.0. About half of the ***z*** scores will be negative, indicating that the people with these scores fell below the mean. Most of the ***z*** scores (about 99%) will fall between -3.0 and +3.0.

Suppose we have given the Tests of Achievement and Proficiency—Form G during the fall to the pupils in a ninth-grade class, and two pupils have the following scores on mathematics and reading comprehension.

Pupil	Mathematics	Reading Comprehension
Henry	30	48
Joe	37	42

Let us see how we can use standard scores to compare performance of an individual on two tests or the performance of the two individuals on a single test.

The mean and standard deviation for mathematics and reading comprehension are as follows:

	Mathematics	Heading Comprehension
Mean	22.7	338
SD	9.4	11.1

On mathematics, Henry is 7.3 points above the mean. His ***z*** score is 7.3/9-1 = +0.78. On reading comprehension, he is 14.2 points above the mean, or ***z*** = 14.2/1 1.1 = + 1.28. Henry is about one-half of a standard deviation better in reading comprehension than in mathematics. For Joe, the corresponding calculations for mathematics give

(37 - 22.7)/9.4 = +1.52

and for reading comprehension give

(42 - 33-H)/l 1.1 = +0.74

Thus, Henry has done about as well on mathematics as Joe has done on reading comprehension, while Joe's mathematics score is about one-quarter of a standard deviation better than Henry's score on reading comprehension.

Each pupil's level of excellence is expressed as a number of standard deviation units above or below the mean of the comparison group. The ***z*** scores provide a standard unit of measure having essentially the same meaning from one test to another. For aid in interpreting the degree of excellence represented by a standard score, see-Table 2.5.

Converted Standard Scores

In all, z scores are quite satisfactory except for two matters of convenience: (1) They require use of plus and minus signs, which may be miscopied or overlooked, and (2) they get us involved with decimal points, which may be misplaced. Also, people do not generally like to think of themselves as negative or fractional quantities. We can get rid of the need to use decimal points by multiplying every ***z*** score by some constant, such as 10. We can get rid of minus signs by adding a convenient constant amount, such as 50. Then, for Henry's scores on mathematics and reading comprehension we would have

	Mathematics	**Reading Comprehension**
Mean of distribution of scores	22.7	33 8
SD of distribution	9.4	11.1
Henry's raw score	30	48
Henry'sz score	+0.78	+1.28
z score X 10	8	13
Plus a constant amount (50)	58	63

(It is conventional to round such converted scores to the nearest whole number, consistent with the objective of making them easy to use.)

Converted standard scores are based on a simple equation that changes the size of the units and the location of the mean. In symbolic form, the equation for the above transformation is where *z* is the standard score defined earlier and ***C*** is the converted standard score.

$$C= 10(z) + 50$$

The use of 50 and 10 for the mean and the standard deviation, respectively, of a linear transformation is an arbitrary decision. We could have used values other than 50 and 10 in setting up the conversion into convenient standard scores. The army has used a standard score scale with a mean of 100 and a standard deviation of 20 for reporting its test results. The College Entrance Examination Board has long used a scale with a mean of 500 and a standard deviation of 100 for reporting scores on the Scholastic Aptitude Test, the Graduate Record Examination, and other tests produced under its auspices. The navy has used the 50 and 10 system; many intelligence tests use a mean of 100 and a standard deviation of 15 or 16.

The scale of scores following a conversion such as this one is stretched out or squeezed together (depending on whether the original standard deviation is smaller or larger than the new one), but the stretching is uniform all along the scale. The ***size*** of the units is changed, but it is changed uniformly throughout the score scale. If the raw score scale represented equal units to begin with, the new scale still docs, but nothing has been done to make unequal units more nearly equal. Because the above equation is an equation for a straight line, this type of transformation of scores is called a ***linear conversion,*** or a ***linear***

transformation. (It is necessary here to add a note on terminology. We'use the symbol ***z*** to stand for standard scores in their original form and ***C*** to stand for ***any*** linear transformation of a ***z*** score. Some authors use the symbol ***T***for the special case of a linear transformation using a mean of 50 and a standard deviation of 10. While details of notation are a matter of personal preference, this use of the symbol 7'is historically incorrect. The symbol ***T*** was first used by William McCall in 1922 to stand for the special kind of ***nonlinear transformation*** described in the next section. Only relatively recently has the distinction been lost, and it seems useful to reinstate the traditional notation to make the important distinction between scores that have been normalized and those that have not.)

Normalizing Transformations

Frequently, standard score scales are developed by combining the percentile ranks corresponding to the raw scores with a linear transformation of the ***z*** scores that are associated with those percentile ranks in the normal distribution, making the assumption that the trait being measured has a normal distribution. (This is called an ***area conversion*** of scores. Because the complete transformation cannot be expressed by a straight line, or linear equation, it is also called a nonlinear transformation.) Thus, in the mathematics test, we might find that 35% of ninth grade boys fall below a score of 17. In the table of the- normal distribution (provided as Appendix 2), the z score below which 35% of the cases fall is -0.39. Consequently, we would ***assign*** to a raw score of 17 a standard score of -0.39. Expressing this result on a scale in which the standard deviation is to be 10 and the mean 50, we have

$$T = 10\ (-0.39) + 50 = -4 + 50 = 46$$

As discussed earlier, the designation of 7'score and the symbol '/'have often been used to identify this particular type of normalized standard score scale.

The complete process of preparing a normalized standard score scale by the area conversion method involves finding the percentile rank for each obtainable raw score. The z score below which the specified percentage of the normal distribution falls is then substituted, for the raw score, resulting in a set of z scores that yield a normal distribution for the group on which we have obtained our data. These z scores are then subjected to a linear transformation using whatever mean and standard deviation are desired (SO and 10. respectively, for / scores).

Normal Curve Equivalents

A second type of normalized standard score gaining popularity in education is the scale of normal curve equivalents, or NCE scale. This scale is developed using the same procedures and mean as the '/'scale uses, but the standard deviation is set at 21.06 rather than at 10. The reason for choosing this particular standard deviation is that it gives a scale in which a score of 1 corresponds to a percentile rank of 1 and a score of 99 corresponds to a percentile rank of 99- The relationship between NCEs and percentile ranks is shown in Table 7.5. Most major publishers of educational achievement tests provide tables of NCI-; scores, thus allowing for comparison of relative performance on different tests As these publishers note, however, the tests differ in content, so a common score scale does not imply that one test could be substituted for another.

We have now identified two ways to develop standard score scales based on an arbitral*)' mean and standard deviation. In one the linear transformation method, z scores are ***computed*** from the observed mean and standard deviation and the resulting z scores may he further transformed by first being multiplied by an arbitrary new standard deviation and then added to an arbitrary new mean. This method docs not change the relative distances between scores and leaves the shape of the score distribution unchanged. In the other method; the area or normalizing transformation, percentile ranks are used to ***assign z*** scores to raw scores, based on the percentage of the normal distribution that falls below the z score. These assigned z scores are then transformed with an arbitrary standard deviation and mean to a desired scale. The resulting scores will form a normal distribution, regardless of the shape of the distribution of the raw scores.

Table 7.5 Relationship between Normal Curve Equivalents, Percentile, Ranks, and Stanines

NCE	PR	Stanine	PR
99	99	9	≥96 +

90	97	8	89-95
80	92	7	77-88
70	83	6	60-76
60	65	5	40-59
50	50	4	23-39
40	32	3	11-22
30	17	2	4-10
20	8	1	3 ≤
10	3		
1	1		

Normalized standard scores make sense whenever it seems likely that the group is a complete one that has not been curtailed **by** systematic selection at the upper or lower ends. Furthermore, they make sense whenever it seems likely that the original raw score scale does not represent a scale of equal units but the underlying trait could reasonably be assumed to have a normal distribution. Many test makers **sys**tematically plan to include in their tests marn' items of medium difficulty and few easy or hard items. The effect of this practice is to produce tests that spread out and make fine discriminations among the middle 80% or 90% of test takers, while making coarser discriminations at the extremes. That is, the raw score units in the middle of the distribution correspond to smaller true increments in the ability being measured than do raw score units at the extremes. The "true" distribution of ability is pulled out into a flat-topped distribution of scores. The operation of normalizing the distribution reverses this process.

Stanines

A type of normalized standard score that has become quite popular for educational tests is the *stannic* (a condensation of the phrase standard nine-point scale) score. The stanine scale has a mean of five, and stanine units each represent half of a standard deviation on the basic trait dimension. Stanines tend to play down small differences in score and to express performance in broader categories, so that attention tends to be focused on differences that are large enough to matter. The relationship between the stanine scale and the percentile rank scale is shown in Table 7.5.

The relationships between a number of the different standard score scales (after normalization) and the relationship of each in percentiles and to the normal distribution are shown in Figure 33- Ibis figure presents the- model of the normal curve, and beneath the normal curve are a scale of percentiles and several of the common standard score scales. This figure illustrates the equivalence of scores in the different systems. Thus, a College board standard score of 600 would represent the same level of excellence (in relation to some common reference group) as an army standard score of 120, a Navy standard score (or 7" score) of 60, a stanine score of 7, a percentile rank of **84,** an NCR of 71, or a Wechsler IQ of I I 5. The particular choice of score scale is arbitrary and a matter of convenience. It is unfortunate that all testing agencies have not been able to agree on a common score unit. However, the important thing is that the same score scale and comparable norming groups be used for all tests in a given organization, so that results from different tests may be directly comparable.

Earlier, we discussed the importance of identifying an appropriate norm group, to allow interpretation of a raw score using percentile norms. The same requirement applies with equal force when we wish to express a person's characteristics within a standard score framework. The conversion from raw to standard score must be based on a relevant group of which the individual with whom we are concerned can be considered a member. It makes no more sense to determine an engineering graduate student's standard score on norm data obtained from high school physics students than it does to express the same comparison in percentiles.

In summary, standard scores, like percentile- ranks, base the interpretation of the individual's score on his or her performance in relation to a particular reference group. They differ from percentiles in that they are expressed in units that are presumed to be equal. The basic unit is the standard deviation of the reference group, and the individual's score is expressed as a number of standard deviation units above or below the* mean of the group. Standard score scales may be based on either a linear or an area (normalizing) conversion of the original scores. Different numerical standard score scales have been used by different testing agencies. Standard score scales share with percentile ranks the problem of defining an appropriate reference group.

QUOTIENTS

In the early days of mental testing, after age norms had been used for a few years, it became apparent that there was a need to convert the age score into an index that would express rate of progress. The 8-year-old who had an age equivalent of 10 1/2 years was obviously better than average, but how much better? Some index was needed to take account of chronological age (actual time lived), as well as the age equivalent on the test (score level reached).

One response to the need was the expedient of dividing the test age by the chronological age to yield a quotient. This procedure was applied most extensively with tests of intelligence, where the- age equivalent on the test was called a ***mental age*** and the corresponding quotient was an ***intelligence quotient*** (IQ). In the 1920s it became common practice to multiply this fraction by 100 (to eliminate decimals), thus giving rise to the general form of the scale that is now so well known in education and psychology (see Chapter S).

The notion of the IQ is deeply embedded in the history of the testing ***movement*** and, in fact, in contemporary American language and culture. The expression "IQ test" has become part of our common speech. We are probably stuck with the term. But the way that the IQ is defined has changed. IQs have become, in almost every case, standard scores with a mean of 100 and a standard deviation of about 15, and we should think of them and use them in this way.

In a number of recent tests of intelligence, the scores that are reported are, in fact, normalized standard scores, based on the type of normalizing area transformation discussed earlier in this chapter. These are sometimes referred to as deviation intelligence quotients, or deviation IQs because they are basically standard scores expressed as a deviation above or below a mean of 100. The latest revision of the Stanford-Binet Intelligence Scale has substituted the term ***standard age score*** for IQ to reflect more accurately the true nature of the scores.

Unfortunately, the score scale for reporting IQs docs not have ***exactly*** the same meaning from test to test.

PROFILES

, .

The various types of normative frames of reference we have been considering provide a way of expressing scores from quite different tests in common units, so that the scores can be meaningfully compared. No direct way of comparing a score of 30 words correctly spelled with a score of 20 arithmetic problems solved exists. Rut, if both are expressed in terms of the grade level to which they correspond or in terms of the percentage of some defined common group that gets scores below that point, then a meaningful comparison is possible. A set of different test scores for an individual, expressed in a common unit of measure, is called a ***score profile.*** The separate scores may be presented for comparison in tabular form by listing the converted score values. A record showing such converted scores for several pupils is given in Figure 3.4. The comparison of different subareas of performance is made pictorially clearer by a graphic presentation of the profile. Two ways of plotting profiles are shown in Figures 3.5 and 3-6.

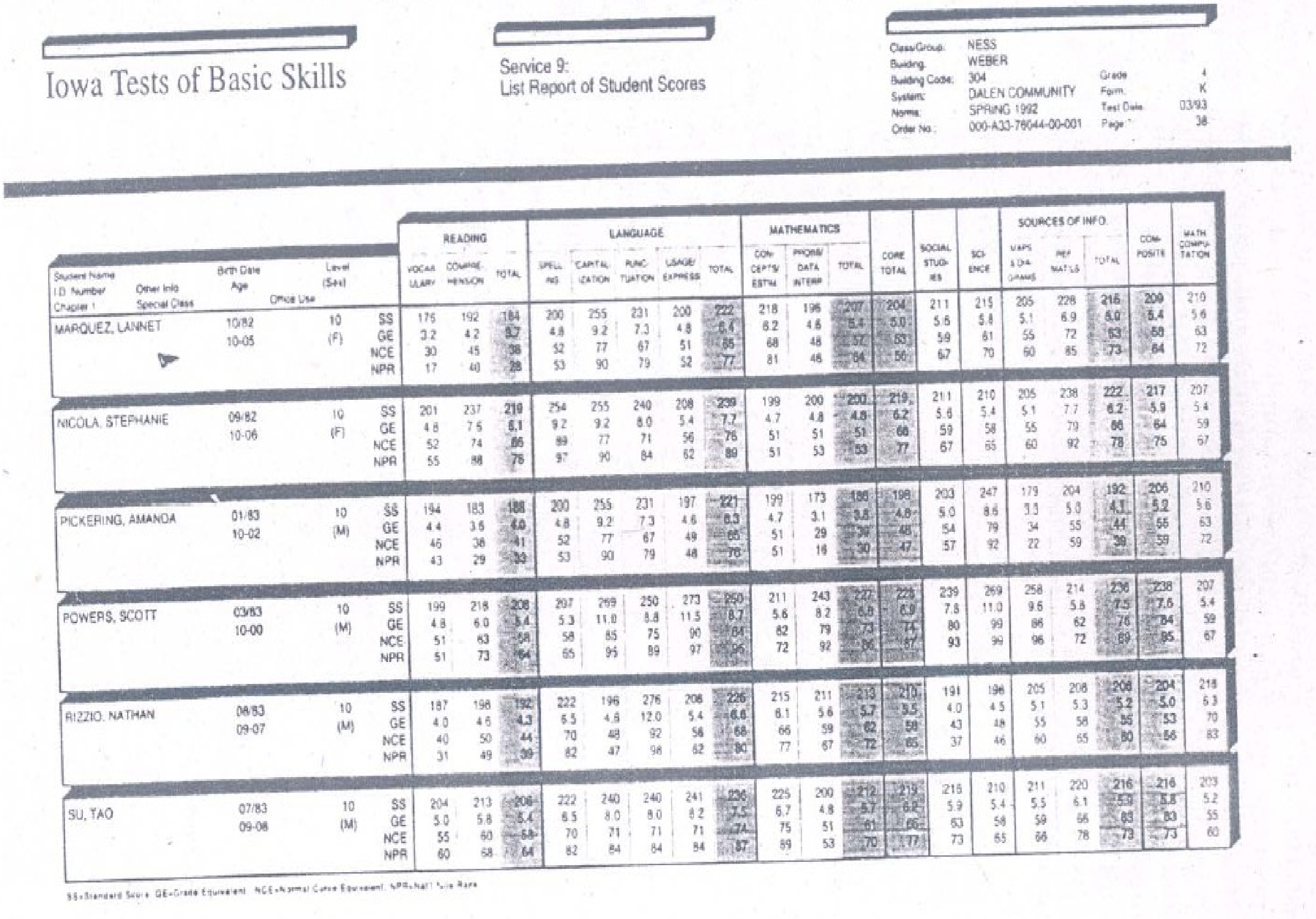
Figures 3.4 and 35 show part of a class record form and an individual profile chart for the ITBS, respectively. The class record illustrates the form in which the data are reported back to the schools by the test publisher's computerized test scoring service. (The precise form that the report of results takes differs from one scoring service to another.) There are four norm-referenced scores reported for each pupil on each test (see Figure 3.4). The first row of the report for each student contains developmental standard scores (called ***SSs*** in this publisher's materials) for the 13 subtests and six composites. The second row of scores are grade equivalents, and because the tests were given after the pupils had spent 7 months in the fourth grade, the norm for the country as a whole would be 4.7. The last two rows for each student contain normal curve equivalents and percentile ranks based on the spring 1992 national norm group. Looking at the scores for Linnet Marquez, we can see that the four score systems give an essentially equivalent picture of her performance. Her grade equivalent of 4.8 is slightly above average, and this is reflected in her NCE and PR scores of 52 and 53. The standard scale score of 200 is also consistent with average performance in the spring of fourth grade. Note that all four reference systems show her to be well above average in capitalization and punctuation.

Figure 3-5 shows data for testings of a student in two successive years. The so-called "developmental scale" referred to toward the left is actually a scale of grade equivalents (GEs). Thus, this pupil had a vocabulary grace equivalent of 5.0 when she was tested the first time. By the next year her grade equivalent on (his test was 5.9. Similar growth of approximately one GE is shown for each of the other subtests, although the level of performance in either year shows considerable variation from one subject to another.

Figure 3.4

List report of pupil scores.

(Copyright © 1993 by The University of Iowa. Reproduced from *Iowa Tests of Basic Skills, Interpretive Guide for Teachers and Counselors, Forms K and L.* Reproduced by permission of the publisher, The Riverside Publishing Company.



Iowa Tests of Basic Skills

Service 9:
List Report of Student Scores

Class/Group: NESS
Building: WEBER
Building Code: 304
System: DALEN COMMUNITY
Norms: SPRING 1992
Order No.: 000-A33-78044-00-001

Grade 4
Form K
Test Date 03/93
Page 38

Student Name I.D. Number Chapter 1	Other Info Special Class	Birth Date Age	Office Use	Level (Sex)		READING			LANGUAGE					MATHEMATICS			CORE TOTAL	SOCIAL STUDIES	SCIENCE	SOURCES OF INFO.			COMPOSITE	MATH COMPUTATION
						VOCABULARY	COMPREHENSION	TOTAL	SPELLING	CAPITALIZATION	PUNCTUATION	USAGE/EXPRESS	TOTAL	CONCEPTS/ESTIM	PROBM/DATA INTERP	TOTAL				MAPS & DIAGRAMS	REF MAT'LS	TOTAL		
MARQUEZ, LANNET		10/82		10	SS	176	192	184	200	255	231	200	222	218	196	207	204	211	215	205	228	216	209	210
		10-05		(F)	GE	3.2	4.2	3.7	4.8	9.2	7.3	4.8	6.4	6.2	4.6	5.4	5.0	5.6	5.8	5.1	6.9	6.0	5.4	5.6
					NCE	30	45	38	52	77	67	51	65	68	48	57	53	59	61	55	72	63	58	63
					NPR	17	40	28	53	90	79	52	77	81	46	64	56	67	70	60	85	73	64	72
NICOLA, STEPHANIE		09/82		10	SS	201	237	219	254	255	240	208	239	199	200	200	219	211	210	205	238	222	217	207
		10-06		(F)	GE	4.8	7.6	6.1	9.2	9.2	8.0	5.4	7.7	4.7	4.8	4.8	6.2	5.6	5.4	5.1	7.7	6.2	5.9	5.4
					NCE	52	74	66	89	77	71	56	76	51	51	51	68	59	58	55	79	66	64	59
					NPR	55	88	78	97	90	84	62	89	51	53	53	77	67	65	60	92	78	75	67
PICKERING, AMANDA		01/83		10	SS	194	183	188	200	255	231	197	221	199	173	186	198	203	247	179	204	192	206	210
		10-02		(M)	GE	4.4	3.6	4.0	4.8	9.2	7.3	4.6	6.3	4.7	3.1	3.8	4.8	5.0	8.6	3.3	5.0	4.1	5.2	5.6
					NCE	46	38	41	52	77	67	49	65	51	29	39	48	54	79	34	55	44	56	63
					NPR	43	29	33	53	90	79	48	76	51	16	30	47	57	92	22	59	39	59	72
POWERS, SCOTT		03/83		10	SS	199	218	208	207	269	250	273	250	211	243	227	228	239	269	258	214	236	238	207
		10-00		(M)	GE	4.8	6.0	5.4	5.3	11.0	8.8	11.5	8.7	5.6	8.2	6.8	6.9	7.8	11.0	9.6	5.8	7.5	7.6	5.4
					NCE	51	63	58	58	85	75	90	84	62	79	73	74	80	99	86	62	76	84	59
					NPR	51	73	64	65	95	89	97	96	72	92	86	87	93	99	96	72	89	95	67
RIZZIO, NATHAN		08/83		10	SS	187	198	192	222	196	276	208	226	215	211	213	210	191	196	205	208	208	204	218
		09-07		(M)	GE	4.0	4.6	4.3	6.5	4.6	12.0	5.4	6.6	6.1	5.6	5.7	5.5	4.0	4.5	5.1	5.3	5.2	5.0	6.3
					NCE	40	50	44	70	48	92	56	68	66	59	62	58	43	48	55	58	56	53	70
					NPR	31	49	39	82	47	98	62	80	77	67	72	65	37	46	60	65	60	56	83
SU, TAO		07/83		10	SS	204	213	208	222	240	240	241	236	225	200	212	219	216	210	211	220	216	216	203
		09-08		(M)	GE	5.0	5.8	5.4	6.5	8.0	8.0	8.2	7.5	6.7	4.8	5.7	6.2	5.9	5.4	5.5	6.1	5.9	5.8	5.2
					NCE	55	60	58	70	71	71	71	74	75	51	61	66	63	58	59	66	63	63	55
					NPR	60	68	64	82	84	84	84	87	89	53	70	77	73	65	66	78	73	73	60

SS=Standard Score GE=Grade Equivalent NCE=Normal Curve Equivalent NPR=Nat'l %ile Rank

The results show her scores generally to have been above the national average. An examination of her profile for the fourth-grade test indicates that she was strongest in capitalization, punctuation, and language usage skills and weakest in mathematics problems. Some of the hazards of paying a great deal of attention to small tips and downs in a profile can be seen in a comparison of her performance on successive testings. Although the profile shows a relatively consistent pattern of highs and lows over the years, relative superiority changes somewhat from one year to the next.

Figure 3.6 shows a second type of profile chart for the ITI5S. Here, the scores for one of the students in the class list (sec Figure 3 *A)* are shown for each of the separate subtests of the batten'. Note that in this case the different tests are represented by separate bars rather than by points connected by a line. The scale used in this case is a percentile scale, but in plotting percentile values, appropriate adjustments in the scale have been made to compensate for the inequality of percentile units. That is, the percentile points have been spaced in the same way that they are in a normal curve, being more widely spaced at the upper and lower extremes than in the middle range. This percentile scale corresponds to the scale called percentile equivalents in Figure 7.3. By this adjustment, the percentile values for an individual are plotted on an equal unit scale. A given linear distance can reasonably be thought to represent the same difference in amount of ability, whether it lies high in the scale, low in the scale, or near the middle of the scale. By the same token, the same distance can be considered equivalent from one test to another.

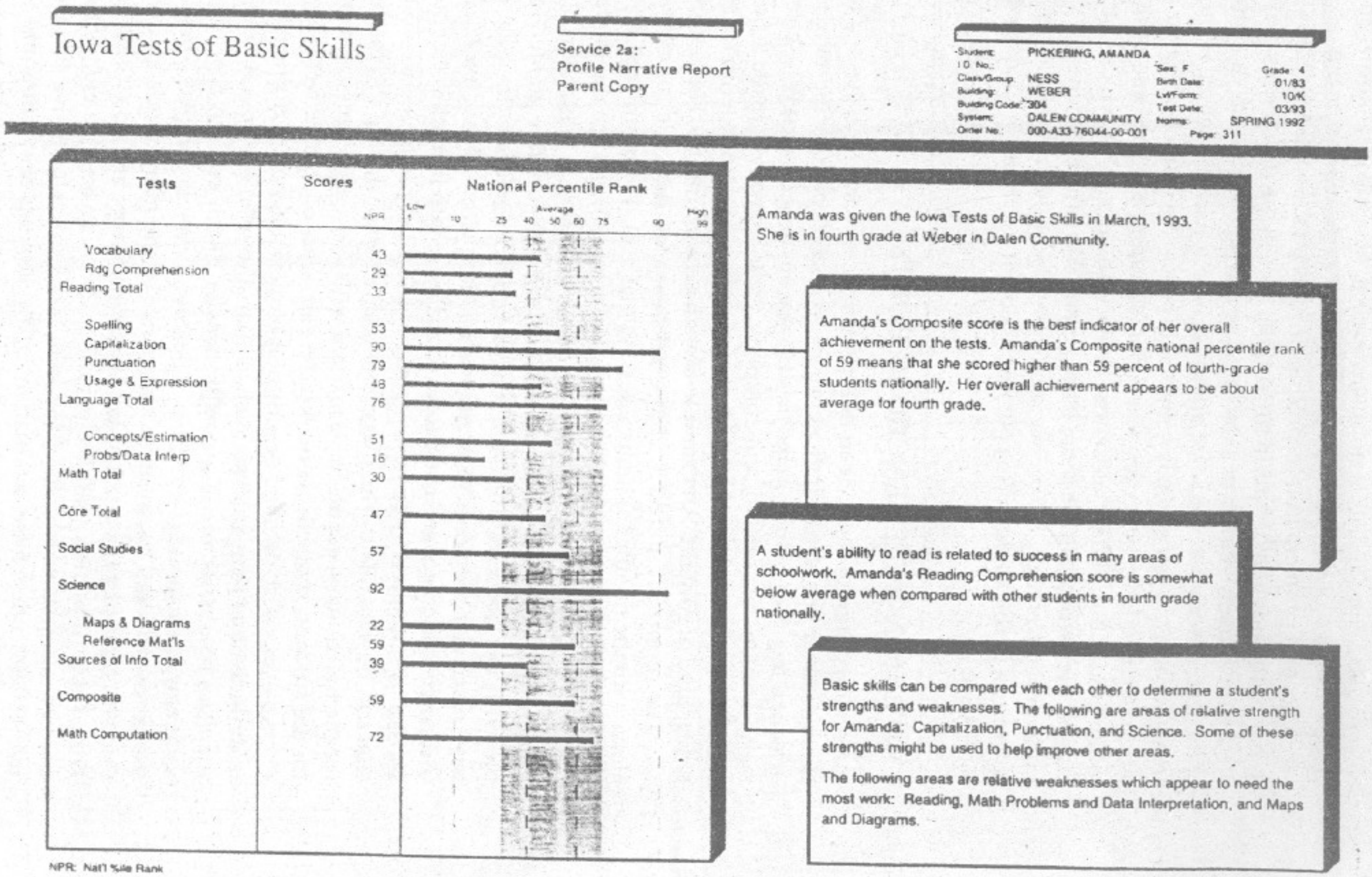
In the profile in Figure 3-6, the middle 50% is shaded to indicate a band of average performance for the norm group. The scores of this student have been plotted as bars that extend from the left side of the chart. For this type of norm, the average of the group constitutes the anchor point of the scale, and the individual scores can be referred to this base level. This type of figure brings out the individual's strengths and weaknesses quite clearly. Note also that the numerical values for this student's percentile ranks in the national norm group are given to (he left of the profile. In addition, this particular test publisher's scoring service provides a narrative interpretation of the profile. Such an interpretation can also help draw the attention of teachers and parents to noteworthy features of the student's performance.

The profile chart is a very effective way of representing an individual's scores, but profiles must be interpreted with caution. First, procedures for plotting profiles assume that the norms for the tests are comparable. For this to be true, age, grade, or percentile scores must be based on equivalent groups for all the- tests. We usually find. In- the case lot the subtests of a test battery. Norms for all the subtests are established at the same lime, on the basis of testing the same group. This guarantee of comparability of norms for the different component tests is one of the most attractive features of an integrated test batten-. If separately developed tests are plotted in a profile, we can usually only hope that the groups on which the norms were established were comparable and that the profile is an unbiased picture of relative achievement in different fields. When it is necessary to use tests from several different sources, one way to be sure of having equivalent norm groups is to develop local norms on a common population and to plot individual profiles in terms of those local norms.

Figure 3.6

Profile narrative report—Parent copy.

(Copyright © 1993 by The University of Iowa. Reproduced from *Iowa Tests of Basic Skills, Interpretive Guide for Teachers and Counselors, Forms K and L.* Reproduced by permission of the publisher, The Riverside Publishing Company.



Iowa Tests of Basic Skills

Service 2a:
Profile Narrative Report
Parent Copy

Student:	PICKERING, AMANDA		
I D No.:		Sex: F	Grade: 4
Class/Group:	NESS	Birth Date:	01/83
Building:	WEBER	Lvl/Form:	10/K
Building Code:	304	Test Date:	03/93
System:	DALEN COMMUNITY	Norms:	SPRING 1992
Order No.:	000-A33-76044-00-001	Page: 311	

Tests	Scores NPR
Vocabulary	43
Rdg Comprehension	29
Reading Total	33
Spelling	53
Capitalization	90
Punctuation	79
Usage & Expression	48
Language Total	76
Concepts/Estimation	51
Probs/Data Interp	16
Math Total	30
Core Total	47
Social Studies	57
Science	92
Maps & Diagrams	22
Reference Mat'ls	59
Sources of Info Total	39
Composite	59
Math Computation	72

NPR: Nat'l %ile Rank

Amanda was given the Iowa Tests of Basic Skills in March, 1993. She is in fourth grade at Weber in Dalen Community.

Amanda's Composite score is the best indicator of her overall achievement on the tests. Amanda's Composite national percentile rank of 59 means that she scored higher than 59 percent of fourth-grade students nationally. Her overall achievement appears to be about average for fourth grade.

A student's ability to read is related to success in many areas of schoolwork. Amanda's Reading Comprehension score is somewhat below average when compared with other students in fourth grade nationally.

Basic skills can be compared with each other to determine a student's strengths and weaknesses. The following are areas of relative strength for Amanda: Capitalization, Punctuation, and Science. Some of these strengths might be used to help improve other areas.

The following areas are relative weaknesses which appear to need the most work: Reading, Math Problems and Data Interpretation, and Maps and Diagrams.

29

A second problem in interpreting profiles is that of deciding how much attention to pay to the ups and downs in the profile. Not all the differences that appear in a profile are meaningful, either in a statistical or in a practical sense. We must decide which of the differences deserve some attention on our part and which do not. This problem arises because no test score is completely exact. No magic size exists at which a score-difference suddenly becomes worthy of attention, and any rule of thumb is at best a rough guide. Hut, differences must be big enough so that we can be reasonably sure (1) that they would still be there if the person were tested again and (2) that they make a practical difference, before we start to interpret them and base action on them. We will return to this topic during our discussion of reliability in Chapter *4*.

CRITERION-REFERENCED REPORTS

Interest in criterion-referenced interpretations of test scores has led test publishers to produce a profile of student performance based on specific item content. A well-designed test will include items that tap various aspects of skill or knowledge development. Modern test scoring and computer technology have made it possible to report a student's performance on subsets of items that are homogeneous with respect lo a particular kind of content. An example of such a report for the I TBS is shown in figure 3-7.

The report presented in Figure 3.7 lists each subtopic for each test of the ITBS, along with the number of items assessing that skill. The number of items the student attempted, the number and percentage of items correct for the student, and the percentages correct for the class and nation are also given, This report allows the teacher to identify specific strengths and weaknesses at a more fine-grained level than is possible with the ordinary norm-referenced report. For example, this student seems to have particular problems with math estimation, although her overall mathematics

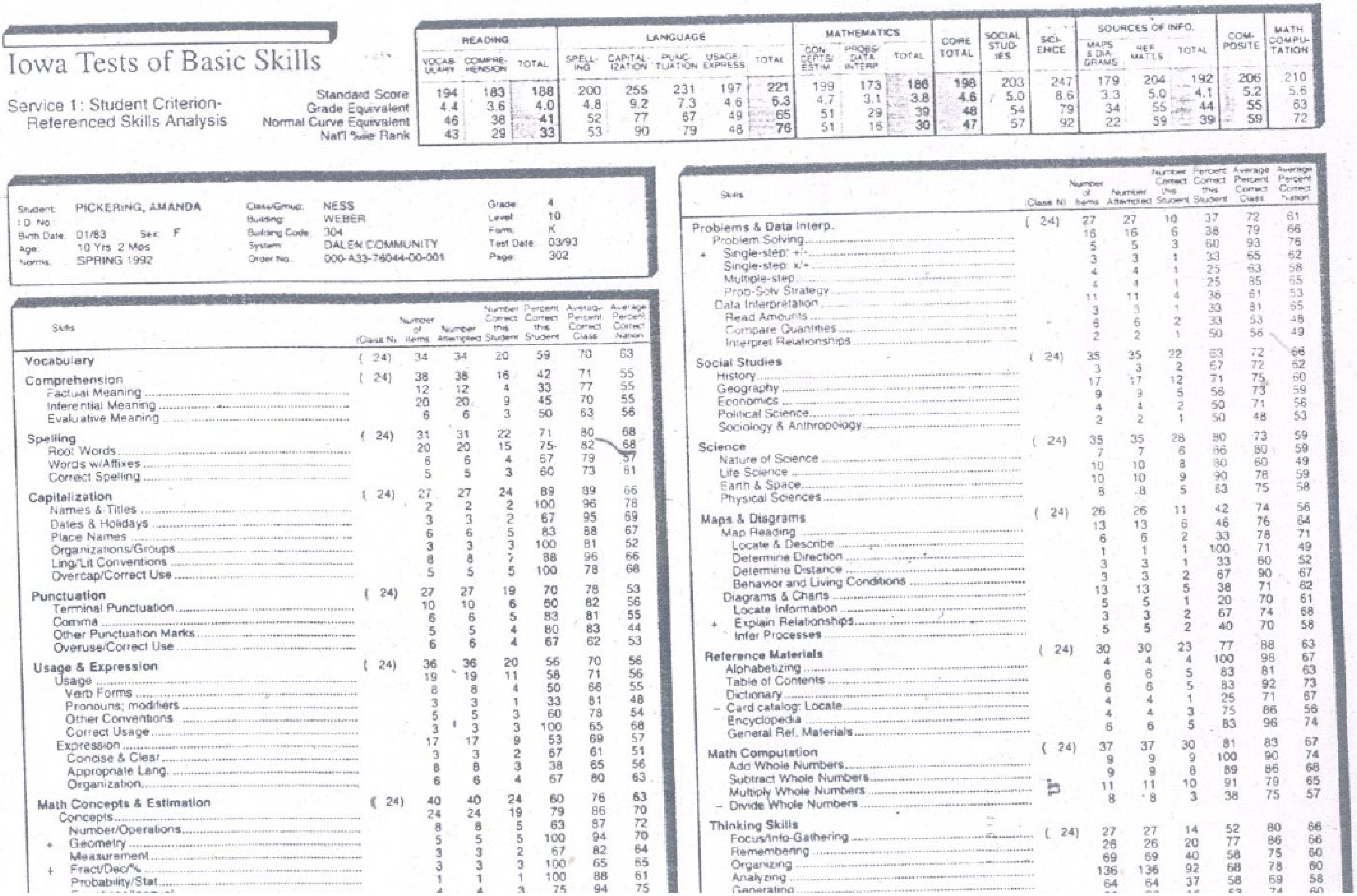
Performance is average, and she shows relative strength in math concepts. {Although each sub skill is measured by too few items to yield a very reliable assessment, the information can be valuable to the classroom teacher in designing the instructional program for the individual student. Skills marked with a (+) represent areas of relative strength for the student, while those marked (-) are areas of relative weakness.

An even more detailed description of this student's performance can be provided in an individual item analysis such. as that shown in Figure 3.8, which shows part of the class results for reading comprehension. Each column represents a student and each row corresponds to an item. The item numbers are given, organized by the skill they measure, and the student's response to the item is indicated if it was incorrect (a blank indicates the student got the item correct and a 0 indicates an omission). From the information on this chart the teacher can see that eight students got item 37 correct, six students omitted the item, six chose alternative B, three chose alternative C, and one chose alternative D. By looking for commonly made errors, the teacher can diagnose particular skill areas where the students need extra work.

Figure 3.7

Student criterion-referenced skills analysis.

(Copyright © 1993 by The University of Iowa. Reproduced from *Iowa Tests of Basic Skills, Interpretive Guide for Teachers and Counselors, Forms K and L.* Reproduced by permission of the publisher, The Riverside Publishing Company.



Iowa Tests of Basic Skills

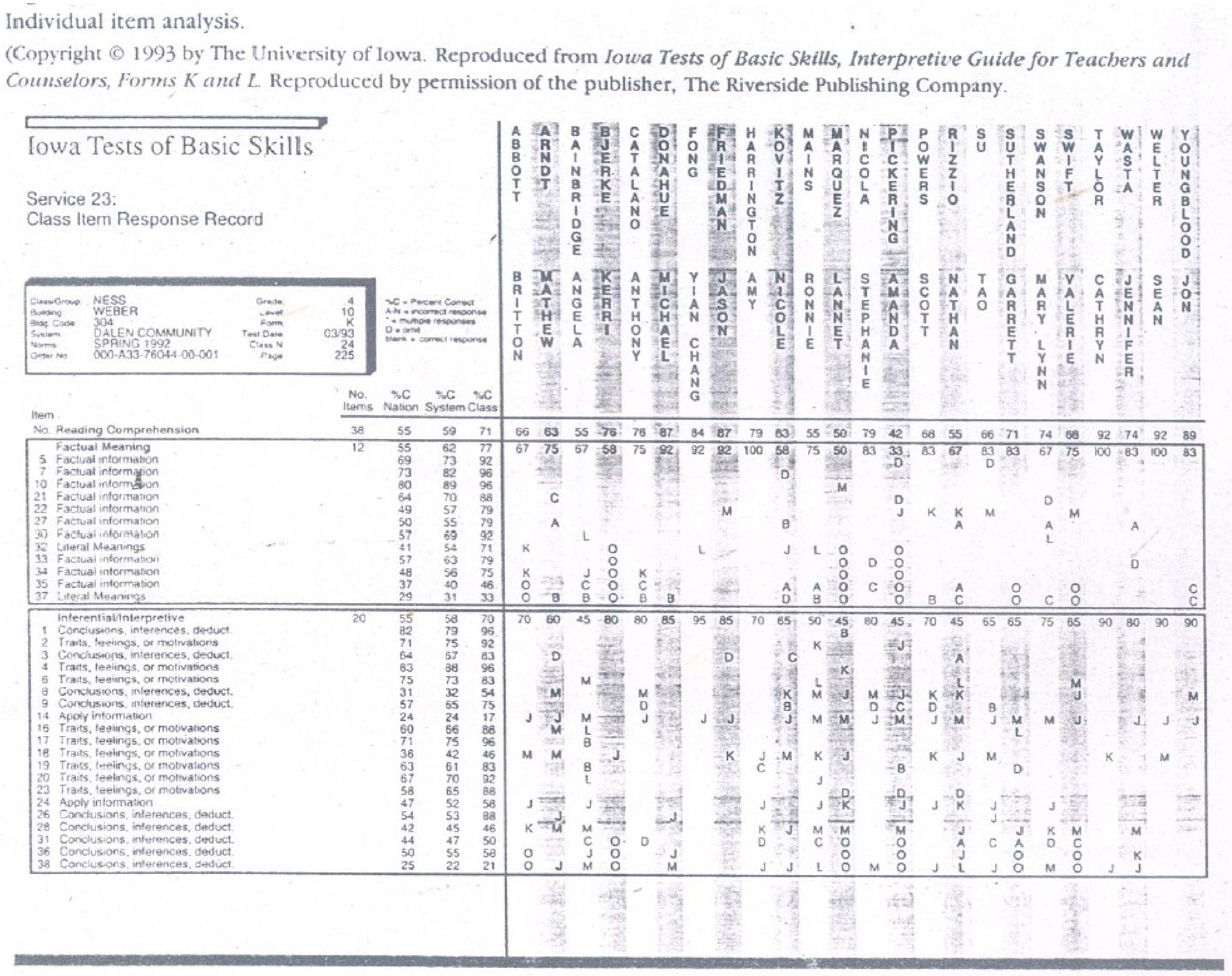
Service 1: Student Criterion-Referenced Skills Analysis

	READING			LANGUAGE					MATHEMATICS			CORE TOTAL	SOCIAL STUDIES	SCIENCE	SOURCES OF INFO.			COMPOSITE	MATH COMPUTATION
	VOCABULARY	COMPREHENSION	TOTAL	SPELLING	CAPITALIZATION	PUNCTUATION	USAGE/EXPRESS	TOTAL	CONCEPTS/ESTIM	PROBS/DATA INTERP	TOTAL				MAPS & DIAGRAMS	REF MATLS	TOTAL		
Standard Score	194	183	188	200	255	231	197	221	199	173	186	198	203	247	179	204	192	206	210
Grade Equivalent	4.4	3.6	4.0	4.8	9.2	7.3	4.6	6.3	4.7	3.1	3.8	4.6	5.0	8.6	3.3	5.0	4.1	5.2	5.6
Normal Curve Equivalent	46	38	41	52	77	67	49	65	51	29	39	48	54	79	34	55	44	55	63
Nat'l %ile Rank	43	29	33	53	90	79	48	76	51	16	30	47	57	92	22	59	39	59	72

Student:	PICKERING, AMANDA	Class/Group:	NESS	Grade:	4
I.D. No.:		Building:	WEBER	Level:	10
Birth Date:	01/83 Sex: F	Building Code:	304	Form:	K
Age:	10 Yrs 2 Mos	System:	DALEN COMMUNITY	Test Date:	03/93
Norms:	SPRING 1992	Order No.:	000-A33-76044-00-001	Page:	302

Skills	(Class N)	Number of Items	Number Attempted	Number Correct this Student	Percent Correct this Student	Average Percent Correct Class	Average Percent Correct Nation
Vocabulary	(24)	34	34	20	59	70	63
Comprehension	(24)	38	38	16	42	71	55
Factual Meaning		12	12	4	33	77	55
Inferential Meaning		20	20	9	45	70	55
Evaluative Meaning		6	6	3	50	63	56
Spelling	(24)	31	31	22	71	80	68
Root Words		20	20	15	75-	82	68
Words w/Affixes		6	6	4	67	79	57
Correct Spelling		5	5	3	60	73	81
Capitalization	(24)	27	27	24	89	89	66
Names & Titles		2	2	2	100	96	78
Dates & Holidays		3	3	2	67	95	69
Place Names		6	6	5	83	88	67
Organizations/Groups		3	3	3	100	81	52
Ling/Lit Conventions		8	8	7	88	96	66
Overcap/Correct Use		5	5	5	100	78	68
Punctuation	(24)	27	27	19	70	78	53
Terminal Punctuation		10	10	6	60	82	56
Comma		6	6	5	83	81	55
Other Punctuation Marks		5	5	4	80	83	44
Overuse/Correct Use		6	6	4	67	62	53
Usage & Expression	(24)	36	36	20	56	70	56
Usage		19	19	11	58	71	56
Verb Forms		8	8	4	50	66	55
Pronouns; modifiers		3	3	1	33	81	48
Other Conventions		5	5	3	60	78	54
Correct Usage		3	3	3	100	65	68
Expression		17	17	9	53	69	57
Concise & Clear		3	3	2	67	61	51
Appropriate Lang.		8	8	3	38	65	56
Organization		6	6	4	67	80	63
Math Concepts & Estimation	(24)	40	40	24	60	76	63
Concepts		24	24	19	79	86	70
Number/Operations		8	8	5	63	87	72
+ Geometry		5	5	5	100	94	70
Measurement		3	3	2	67	82	64
+ Fract/Dec/%		3	3	3	100	65	65
Probability/Stat		1	1	1	100	88	61
[illegible]		4	4	3	75	94	75

Skills	(Class N)	Number of Items	Number Attempted	Number Correct this Student	Percent Correct this Student	Average Percent Correct Class	Average Percent Correct Nation
Problems & Data Interp.	(24)	27	27	10	37	72	61
Problem Solving		16	16	6	38	79	66
+ Single-step: +/-		5	5	3	60	93	76
Single-step: x/÷		3	3	1	33	65	62
Multiple-step		4	4	1	25	63	58
Prob-Solv Strategy		4	4	1	25	85	65
Data Interpretation		11	11	4	36	61	53
Read Amounts		3	3	1	33	81	65
Compare Quantities		6	6	2	33	53	48
Interpret Relationships		2	2	1	50	56	49
Social Studies	(24)	35	35	22	63	72	66
History		3	3	2	67	72	52
Geography		17	17	12	71	75	60
Economics		9	9	5	56	73	59
Political Science		4	4	2	50	71	56
Sociology & Anthropology		2	2	1	50	48	53
Science	(24)	35	35	28	80	73	59
Nature of Science		7	7	6	86	80	59
Life Science		10	10	8	80	60	49
Earth & Space		10	10	9	90	78	59
Physical Sciences		8	.8	5	63	75	58
Maps & Diagrams	(24)	26	26	11	42	74	56
Map Reading		13	13	6	46	76	64
Locate & Describe		6	6	2	33	78	71
Determine Direction		1	1	1	100	71	49
Determine Distance		3	3	1	33	60	52
Behavior and Living Conditions		3	3	2	67	90	67
Diagrams & Charts		13	13	5	38	71	62
Locate Information		5	5	1	20	70	61
+ Explain Relationships		3	3	2	67	74	68
Infer Processes		5	5	2	40	70	58
Reference Materials	(24)	30	30	23	77	88	63
Alphabetizing		4	4	4	100	98	67
Table of Contents		6	6	5	83	81	63
Dictionary		6	6	5	83	92	73
– Card catalog: Locate		4	4	1	25	71	67
Encyclopedia		4	4	3	75	86	56
General Ref. Materials		6	6	5	83	96	74
Math Computation	(24)	37	37	30	81	83	67
Add Whole Numbers		9	9	9	100	90	74
Subtract Whole Numbers		9	9	8	89	86	68
Multiply Whole Numbers		11	11	10	91	79	65
– Divide Whole Numbers		8	8	3	38	75	57
Thinking Skills	(24)	27	27	14	52	80	66
Focus/Info-Gathering		26	26	20	77	86	66
Remembering		69	69	40	58	75	60
Organizing		136	136	92	68	78	60
Analyzing		64	64	37	58	69	58
Generating		[illegible]	[illegible]	[illegible]	53	74	60



Individual item analysis.

(Copyright © 1993 by The University of Iowa. Reproduced from *Iowa Tests of Basic Skills, Interpretive Guide for Teachers and Counselors, Forms K and L.* Reproduced by permission of the publisher, The Riverside Publishing Company.

Iowa Tests of Basic Skills

Service 23:
Class Item Response Record

Class/Group	NESS	Grade	4
Building	WEBER	Level	10
Bldg. Code	304	Form	K
System	DALEN COMMUNITY	Test Date	03/93
Norms	SPRING 1992	Class N	24
Order No	000-A33-76044-00-001	Page	225

%C = Percent Correct
A-N = incorrect response
* = multiple responses
O = omit
blank = correct response

Item No.	Reading Comprehension	No. Items	%C Nation	%C System	%C Class	ABBOTT BRITTON	ARNDT MATHEW	BAINBRIDGE ANGELA	BJERKE KERRI	CATALANO ANTHONY	DONAHUE MICHAEL	FONG YIAN CHANG	FRIEDMAN JASON	HARRINGTON AMY	KOVITZ NICOLE	MAINS RONNIE	MARQUEZ LANNET	NICOLA STEPHANIE	PICKERING AMANDA	POWERS SCOTT	RIZZIO NATHAN	SU TAO	SUTHERLAND GARRETT	SWANSON MARY LYNN	SWIFT VALERIE	TAYLOR CATHRYN	WASTA JENNIFER	WELTER SEAN	YOUNGBLOOD JON
	Reading Comprehension	38	55	59	71	66	63	55	76	76	87	84	87	79	83	55	50	79	42	68	55	66	71	74	68	92	74	92	89
	Factual Meaning	12	55	62	77	67	75	67	58	75	92	92	92	100	58	75	50	83	33	83	67	83	83	67	75	100	83	100	83
5	Factual information		69	73	92														D			D							
7	Factual information		73	82	96										D														
10	Factual information		80	89	96												M												
21	Factual information		64	70	88		C												D					D					
22	Factual information		49	57	79								M						J	K	K	M			M				
27	Factual information		50	55	79		A								B						A			A			A		
30	Factual information		57	69	92			L																L					
32	Literal Meanings		41	54	71	K			O			L			J	L	O		O										
33	Factual information		57	63	79				O								O	D	O								D		
34	Factual information		48	56	75	K		J	O	K							O		O										
35	Factual information		37	40	46	O		C	O	C					A	A	O	C	O		A		O		O				C
37	Literal Meanings		29	31	33	O	B	B	O	B	B				D	B	O		O	B	C		O	C	O				C
	Inferential/Interpretive	20	55	58	70	70	60	45	80	80	85	95	85	70	65	50	45	80	45	70	45	65	65	75	65	90	80	90	90
1	Conclusions, inferences, deduct.		82	79	96												B												
2	Traits, feelings, or motivations		71	75	92											K			J										
3	Conclusions, inferences, deduct.		64	67	83		D						D		C						A								
4	Traits, feelings, or motivations		83	88	96												K												
6	Traits, feelings, or motivations		75	73	83			M								L					L				M				
8	Conclusions, inferences, deduct.		31	32	54		M			M					K	M	J	M	J	K	K				J				M
9	Conclusions, inferences, deduct.		57	65	75					D					B			D	C	D		B							
14	Apply information		24	24	17	J	J	M		J		J	J		J	M	M	J	M	J	M	J	M	M	J		J	J	J
16	Traits, feelings, or motivations		60	66	88		M	L															L						
17	Traits, feelings, or motivations		71	75	96			B																					
18	Traits, feelings, or motivations		36	42	46	M	M		J				K	J	M	K	J			K	J	M				K		M	
19	Traits, feelings, or motivations		63	61	83			B						C					B				D						
20	Traits, feelings, or motivations		67	70	92			L								J													
23	Traits, feelings, or motivations		58	65	88												D		D		D								
24	Apply information		47	52	58	J		J						J		J	K		J	J	K	J		J					
26	Conclusions, inferences, deduct.		54	53	88		J				J											J							
28	Conclusions, inferences, deduct.		42	45	46	K	M	M						K	J	M	M		M		J		J	K	M		M		
31	Conclusions, inferences, deduct.		44	47	50			C	O	D				D		C	O		O		A	C	A	D	C				
36	Conclusions, inferences, deduct.		50	55	58	O		J	O		J						O		O		J		O		O		K		
38	Conclusions, inferences, deduct.		25	22	21	O	J	M	O		M			J	J	L	O	M	O	J	L	J	O	M	O	J	J		

32

Figure 3.8 gives students -by-student detail, but for examining the strengths and weaknesses of the class as a whole, information such as that provided in Figure 3-9 may be more Useful. This report compares the performance of this class with that of the national norm group. The results, shown item by item in terms of percent correct, are displayed both numerically and graphically. The shaded area indicates when the difference is less than 10% and, therefore, probably too small to be of interest. The results for this class show a broad pattern of performance above the norm group (note that the table would continue similar information about items covering other skills and knowledge areas).

Figure 3.7 illustrates quite clearly the way content-based and norm-based frames of reference can coexist in the same test and can supplement each other in score interpretation. The report shows this student's performance, by content area, with reference to the number of items covering that content, the average performance of her class, and the average performance of the grade-equivalent national norm group. Additional reports are available that show, for example, the performance of the class on each item relative to national, system, and building norms (school performance) or that summarize the individual information in Figure 3-7 for the entire class. The publisher's catalog for this test lists over 30 forms of reports that are available. However, it is important to keep in mind that criterion-referenced interpretations of standardized tests are based on very small numbers of items (one or two in some cases) for each content area or objective. Therefore, any conclusions based on such data must be tentative and should be confirmed using other sources of information.

NORMS FOR SCHOOL AVERAGES

Up to this point, we have asked how we can interpret an individual's standing on a test. Sometimes a question arises about the relative performance of a class, a school, a school district, or even the schools of a whole state. The recent emphasis on accountability in education provides ample reason for educators to be concerned about evaluating the performance of students taken as groups. When evaluating the achievement of a school in relation to other schools, it is necessary to have norms for school averages.

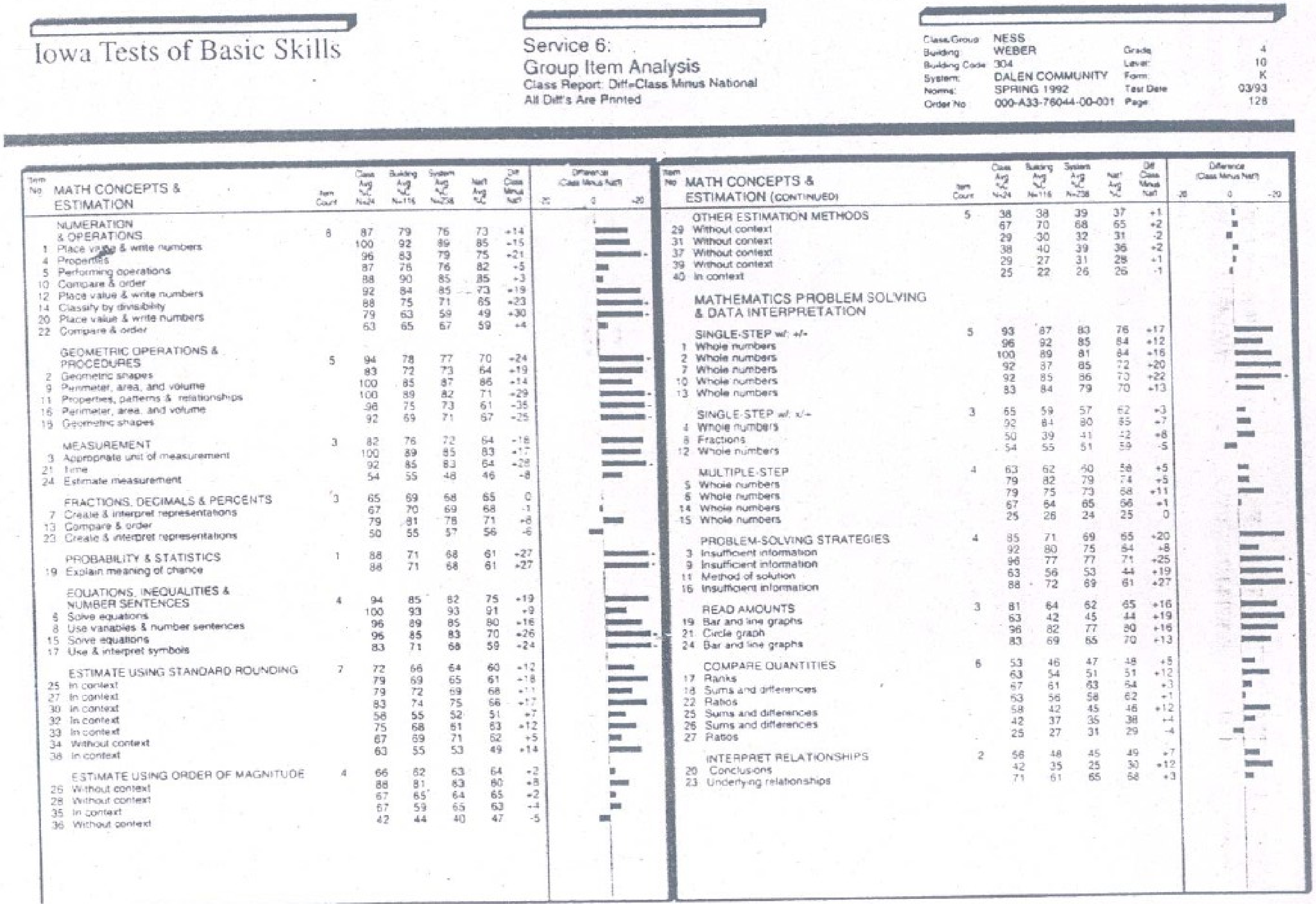
It should be clear that the variation from school **to** school in average ability **or** achievement is much less than the variation from pupil **to** pupil. **No** school **average** comes even close to reaching the level of its ablest student, **and no average** drops anywhere near the performance of the least able. Thus, **a** single pupil **at** the beginning of fifth grade who gets a reading grade equivalent **of** 6.2 might **fall at** the 75th percentile, whereas a school whose ***average*** reading grade equivalent **of** beginning fifth graders is 6.2 might fall at about the 9-1th percentile **of** schools. The relationship between norms for individuals and groups is illustrated more fully **in** Table 38.

The two distributions center at about the same point, but the greater variation among individuals quickly becomes apparent. On this test, an individual grade equivalent of 6.0 ranks at the 60th percentile, but a school in which the average performance is a grade equivalent of 6.0 is at the 85th percentile. The same effect is found for performances that are below average.

Figure 3.9

Group item analysis.

(Copyright © 1993 by The University of Iowa. Reproduced from *Iowa Tests of Basic Skills, Interpretive Guide for Teachers and Counselors, Forms K and L.* Reproduced by permission of the publisher, The Riverside Publishing Company.



Iowa Tests of Basic Skills

Service 6:
Group Item Analysis
Class Report: Diff=Class Minus National
All Diff's Are Printed

Class/Group	NESS		
Building	WEBER	Grade	4
Building Code	304	Level	10
System	DALEN COMMUNITY	Form	K
Norms	SPRING 1992	Test Date	03/93
Order No	000-A33-76044-00-001	Page	128

Item No	MATH CONCEPTS & ESTIMATION	Item Count	Class Avg %C N=24	Building Avg %C N=116	System Avg %C N=238	Nat'l Avg %C	Diff Class Minus Nat'l
	NUMERATION & OPERATIONS	8	87	79	76	73	+14
1	Place value & write numbers		100	92	89	85	+15
4	Properties		96	83	79	75	+21
5	Performing operations		87	76	76	82	+5
10	Compare & order		88	90	85	85	+3
12	Place value & write numbers		92	84	85	73	+19
14	Classify by divisibility		88	75	71	65	+23
20	Place value & write numbers		79	63	59	49	+30
22	Compare & order		63	65	67	59	+4
	GEOMETRIC OPERATIONS & PROCEDURES	5	94	78	77	70	+24
2	Geometric shapes		83	72	73	64	+19
9	Perimeter, area, and volume		100	85	87	86	+14
11	Properties, patterns & relationships		100	89	82	71	+29
16	Perimeter, area, and volume		98	75	73	61	+35
19	Geometric shapes		92	69	71	67	+25
	MEASUREMENT	3	82	76	72	64	+18
3	Appropriate unit of measurement		100	89	85	83	+17
21	Time		92	85	83	64	+28
24	Estimate measurement		54	55	48	46	+8
	FRACTIONS, DECIMALS & PERCENTS	3	65	69	68	65	0
7	Create & interpret representations		67	70	69	68	-1
13	Compare & order		79	81	78	71	+8
23	Create & interpret representations		50	55	57	56	-6
	PROBABILITY & STATISTICS	1	88	71	68	61	+27
19	Explain meaning of chance		88	71	68	61	+27
	EQUATIONS, INEQUALITIES & NUMBER SENTENCES	4	94	85	82	75	+19
5	Solve equations		100	93	93	91	+9
8	Use variables & number sentences		96	89	85	80	+16
15	Solve equations		96	85	83	70	+26
17	Use & interpret symbols		83	71	68	59	+24
	ESTIMATE USING STANDARD ROUNDING	7	72	66	64	60	+12
25	In context		79	69	65	61	+18
27	In context		79	72	69	68	+11
30	In context		83	74	75	66	+17
32	In context		58	55	52	51	+7
33	In context		75	68	61	63	+12
34	Without context		67	69	71	62	+5
38	In context		63	55	53	49	+14
	ESTIMATE USING ORDER OF MAGNITUDE	4	66	62	63	64	+2
26	Without context		88	81	83	80	+8
28	Without context		67	65	64	65	+2
35	In context		67	59	65	63	+4
36	Without context		42	44	40	47	-5

Item No	MATH CONCEPTS & ESTIMATION (CONTINUED)	Item Count	Class Avg %C N=24	Building Avg %C N=116	System Avg %C N=238	Nat'l Avg %C	Diff Class Minus Nat'l
	OTHER ESTIMATION METHODS	5	38	38	39	37	+1
29	Without context		67	70	68	65	+2
31	Without context		29	30	32	31	-2
37	Without context		38	40	39	36	+2
39	Without context		29	27	31	28	+1
40	In context		25	22	26	26	-1
	MATHEMATICS PROBLEM SOLVING & DATA INTERPRETATION						
	SINGLE-STEP w/ +/-	5	93	87	83	76	+17
1	Whole numbers		96	92	85	84	+12
2	Whole numbers		100	89	81	84	+16
7	Whole numbers		92	87	85	72	+20
10	Whole numbers		92	85	86	70	+22
13	Whole numbers		83	84	79	70	+13
	SINGLE-STEP w/ x/÷	3	65	59	57	62	+3
4	Whole numbers		92	84	80	85	+7
8	Fractions		50	39	41	42	+8
12	Whole numbers		54	55	51	59	-5
	MULTIPLE-STEP	4	63	62	60	58	+5
5	Whole numbers		79	82	79	74	+5
6	Whole numbers		79	75	73	68	+11
14	Whole numbers		67	64	65	66	+1
15	Whole numbers		25	26	24	25	0
	PROBLEM-SOLVING STRATEGIES	4	85	71	69	65	+20
3	Insufficient information		92	80	75	84	+8
9	Insufficient information		96	77	77	71	+25
11	Method of solution		63	56	53	44	+19
16	Insufficient information		88	72	69	61	+27
	READ AMOUNTS	3	81	64	62	65	+16
19	Bar and line graphs		63	42	45	44	+19
21	Circle graph		96	82	77	80	+16
24	Bar and line graphs		83	69	65	70	+13
	COMPARE QUANTITIES	6	53	46	47	48	+5
17	Ranks		63	54	51	51	+12
18	Sums and differences		67	61	63	64	+3
22	Ratios		63	56	58	62	+1
25	Sums and differences		58	42	45	46	+12
26	Sums and differences		42	37	35	38	+4
27	Ratios		25	27	31	29	-4
	INTERPRET RELATIONSHIPS	2	56	48	45	49	+7
20	Conclusions		42	35	25	30	+12
23	Underlying relationships		71	61	65	68	+3

Table 7.6 Individual and School Average Norms for the Iowa Tests of Basic

Skills Vocabulary Test (Grade 5)

Grade Equivalent	Individual Percentile Rank	Percentile Rank for School Averages
8.6	99	99
7.0	87	99
6.7	83	98
6.5	79	96
6.0	60	85
5.5	57	65
5.2	50	52
5.0	46	43
4.5	34	23
4.0	24	9
3.5	16	2
3.4	14	1
2.5	5	1
1.7	1	1

When a school principal or an administrator in a central office is concerned with interpreting the average performance in a school, norms for school averages are the appropriate ones to use, and it is reasonable to expect the test publisher to provide them. The better test publishers will also provide item analyses and criterion-referenced reports at the level of the class, building, district, and state.

CAUTIONS IN USING NORMS

For a test that assesses standing on some trait or competence in some area of knowledge, norms provide a basis for interpreting the scores of an individual or a group. Converting the score for any test taken singly into an age or grade equivalent, percentile rank, or standard score permits an interpretation of the level at which the individual is functioning on that particular test. Bringing together the set of scores for an individual in a common unit of measure, and perhaps expressing these scores in a profile, brings out the relative level of performance of the individual in different areas.

The average performance for a class, a grade group in a school, or the children in the same grade throughout a school system may be reported similarly. We can then see the average level of performance within the group on some single function or the relative performance of the group in each of several areas. Norms provide a frame of reference within which the picture may be viewed and bring all parts of the picture into a common frame. Now, what does the picture mean, and what should we do about it?

Obviously, it is not possible, in a few pages, to provide a ready-made interpretation for each set of scores that may be obtained in a practical testing situation. How ever, we can lay out a few general guidelines and principles that may help to forestall some unwise interpretations of test results.

The most general point to keep in mind is that test results, presented in any normative scale, are a ***description of what*** *is,* not a *prescription of what should* **be.** The results make it possible to compare an individual or a class with other individuals and classes with respect to one or more aspects of accomplishment or personality, but they do not in any absolute sense tell us whether the individual is doing "well" or "poorly." They do not provide this information for several reasons.

Normative Scores Give Relative Rather Than Absolute Information. They tell whether an individual pupil's achievement is as high as that of other pupils or whether a class scores as high as other classes. But they do not tell us whether the basic concepts of numbers are being mastered or whether the pupils read well enough to comprehend the instructions for filling out an income lax return. Furthermore, they give us little guidance on how much improvement we might expect from *all* pupils if our educational system operated throughout at higher efficiency.

Remember that by the very nature of relative scores, there will be as many people below average as above. When "the norm" means the average of a reference group, it is a statistical necessity that about half of the group be, to a greater or lesser degree, below average. There has been an enormous amount of foolishness —both in single schools and in statewide legislation—about bringing all pupils "up lo the grade norm." This might conceivably be done temporarily if we had a sudden and enormous improvement in educational effectiveness; however, the next lime new norms were established for the test it would take a higher absolute level of performance to, say, read at the sixth-grade level. So we would be back again with half of the pupils falling at or below average. And if the effectiveness of the schools were to return to the former level, we would be faced with the unhappy prospect of more than half of the students testing "below grade level."

The relative nature of norms has been recognized in the criterion-referenced test movement. When a teacher or a school is concerned with appraising mastery of some *specific* instructional objective, it may be more useful to develop test exercises that appraise that objective, lo agree on some standard as representing an acceptable level of master)', and to determine which students do and which do not have master)' of that specific objective than it would be to know how the students from this school perform relative to those from other schools. In the context described, it is possible for all students to achieve mastery, but some will get there faster than others. Even in a criterion-referenced framework there will still be differences among individuals in their levels of accomplishment.

Output Must Re Evaluated Relative to Input. Test results typically give a picture of output—of the individual or of the group as it exists at the lime of testing, after a period of exposure to educational effort. But what of the input? Where did the group start?

The notion of input is a complex and rather subtle one. Our conception of input should include not only earlier status on the particular ability being measured and individual potential for learning, as far as we are able to appraise this, but also the familial circumstances and environmental supports that make it easier for some children to learn than for others. Parental aspirations for the child, parental skills at teaching and guidance of learning, parental discipline and control, linguistic patterns, and cultural resources in the home are part of the input just as truly as are the biological characteristics of the young organism. Furthermore, peer group and community attitudes are an additional real, though possibly modifiable, part of the input as far as the prospects for learning for a given child are concerned. We must recognize that the adequate appraisal of input is no simple matter, and that, correspondingly, the appraisal of output as "satisfactory" or "unsatisfactory" is something we can do with only modest confidence.

Output Must Be Evaluated Relative to Objectives. The design, content, and norms for published standardized tests are based on their authors' perception **of** common national curricular objectives. The topics included, their relative emphasis, and the levels at which they are introduced reflect that perceived general national pattern. To the extent, then, that a given school system deviates in its objectives and curricular emphases from the national pattern, as interpreted by the test maker, its output at a given grade level can be expected to deviate from the national norms. If computational skills receive little emphasis, it is reasonable **to** find that computational facility will be underdeveloped. If map reading has been delayed beyond the grade level at which it is introduced into the test, it is reasonable to find that relative standing on that part of the test will suffer. Unevenness **of** the local profile, in relation to national norms, should always lead one to inquire whether the low spots represent failures of the local program to achieve its objectives or a planned deviation **of** emphasis from what is more typical of schools nationally. Low performance that results from conscious curricular decisions would be much less cause for alarm than a similar level of performance would be in an area of curricular emphasis. Which of these conditions obtains Will no doubt influence what is done with the finding.

To the extent that individual states have uniform objectives for all districts within their boundaries, well-designed standardized tests measuring achievement of these objectives often are available through contract arrangements with test publishers. Several states now contract with organizations that specialize in test development to have tests constructed according to specifications provided by the state board of education. Such tests usually are intended to be used at particular points in the educational program, such as the transitions from elementary school to middle school, middle school to high school, and near the end of high school.

If these considerations and some of the caveats discussed in the next two chapters are borne in mind, the teacher, principal, superintendent, or school board will be interpreting the reported test results with increased wisdom and restraint

Lesson 08

RESPONSE BIAS, INTELLIGENCE AND THEORIES OF INTELLIGENCE

Response Bias and Item Format

So far we have distinguished between independent and forced-choice formats on the basis of measurement scheme-normative or ipsative. The formats also differ in their approach to the problem of response bias. Because objective tests ask direct questions, test takers potentially can bias or distort their responses. Test takers may try to figure out how a particular type of person would answer the question and respond accordingly. In addition, because objective tests ask test takers to select a response from a set of options, they can produce answers to test items without ever considering item content. They can randomly select answers, always pick "true" for even-numbered questions—engage in all sorts of bizarre answer-generating strategies. Any type of systematic distortion or bias of responses threatens the validity of test results because responses would not reflect true test-taker characteristics (e.g. Cronbach, 1946). The ability of the test to control or identify such distortion varies as a function of item format (e.g., Cronbach, 1950).

Two types of response bias or dissimulation are possible. Test takers could demonstrate a response set, systematically selecting answers in an effort to present themselves in a particular light. For example, test takers could be concerned about social desirability and attempt to present very positive pictures of themselves (Edwards. 1957). To bias responses in this way, the test taker tries to imagine how a well-liked person would answer each item and selects alternatives accordingly. Response set therefore, is a content-dependent bias. Test takers distort their responses systematically according to the content of each item.

A second type of distortion is response style. In this case, test takers adopt systematic strategies for answering items about which they are unsure. The strategy "When in doubt, pick C" on multiple choice achievement tests is a response style. Similarly, the tendency to agree with statements regardless of their content (acquiescence) and lo avoid the use of extreme categories when rating statements (central tendency) are response styles used on personality tests. Ii* contrast to response set, a response style is a content-free bias.

The use of forced-choice items or independent items presents different options for test developers concerned about response sets. The contrast is clearly seen when the design of the Edwards Personal Preference Schedule (EPPS) and the Minnesota Multiphasic Personality Inventory (MMPI—2) are compared.

Edwards was particularly concerned about social desirability because research indicated that people were significantly more likely to select socially desirable statements whenever a choice was available (Edwards. 1957). The design of the EPPS attempts to use the forced-choice format lo control this bias. First, statements were written to represent the different needs being measured. Next, a large sample of people was used to assign a social desirability rating to each statement. The forced-choice pairs were constructed so that each pair contained statements representing different needs equated for degree of social' desirability. Each item, therefore, requires test takers to choose between two equally desirable or equally undesirable alternatives. The construction of items makes it impossible for a test taker to select only statements expressing socially desirable characteristics.

Although the forced-choice format of the EPPS reduces the social desirability response set, research indicates that it is not eliminated (e.g., Feldman & Corah, 1960). It is difficult to write a large number of items that are truly equated on a response set dimension. Furthermore, most objective personality tests do not use the forced-choice format. Many examinees are uncomfortable with the forced-choice format. With independent items, however, test takers rate each statement independently. There is no way to prevent people from selecting socially desirable answers—or from engaging in any other response set. Tests using independent formats can, however, use other techniques to identify the presence of response sets.

The MMPI-2 uses a set of special validity scales to identify possible response biases. These validity scales are summarized in Table 4.2. Note that these scales let Us identify a variety of possible biases in test-taker responses. For example, a high "lie" (L scale) score suggests a need to present oneself as a good person, whereas a high "fake bad" (F scale) implies possible exaggeration of symptoms. Since the "fake good" (K) scale covers several issues, either a high or a low K score leads us to question the validity of the test results. In fact, K scores are used to adjust the scoring of the personality scales themselves to produce a more accurate overall profile.

The comparison of the EPPS and the MM PI-2 illustrates how the problem of response set can be addressed when forced-choice or independent formats are selected. But what of the problem of response style or content-free biases? Again, a variety of strategies is available. To check on possible random responding, items can be repeated within a test and the pattern of answers evaluated to determine the consistency of response. Both the EPPS and the MMPI-2 use this procedure. For example, the EPPS repeats 15 items in random locations throughout the test. Research during test development indicated that most individuals answer at leasl 9 of these items consistently. Therefore, test takers who answer less than 9 of these items consistently are suspected of random responding.

Table 8.1 Validity Scales of the MMPI-2

Scale	Measures	By Noting
? or cannot say	willingness to disclose information	Number of items not answered
L or "lie" scale	Presentation of self as "ideal" or "perfect"	Number of false responses to statements describing ordinary "bad" behavior (e.g., not always telling the truth)
F or 'fake bad" scale	Presentation of self as pathological; random responding; failure to understand questions	Responses to items describing unusual or pathological events (e.g., hearing voices, out-of-body experiences)
K or "fake good" scale	Overly favorable presentation of sell"; defensive-ness (high K); willingness to present self in socially undesirable way (low K)	Similarities between your answers and those given by clinical versus nonclinical samples who produced otherwise normal profiles

To identify the tendency to acquiesce on a true/false or yes/no test, statements tapping a particular attribute can be written in forms that require different responses. A particular behavior or attitude could be presented twice, once stated positively and once stated negatively. If the process is repeated for several behaviors or attitudes, the consistency of content-based responding could be determined. Inconsistent responding in this case would indicate either acquiescence (too many "true" or "yes" answers) or negativism (too many "false" or "no" answers).

Another possibility is to write items so that points are earned on half of them by a true/yes response and on half by a false/no response. Test takers responding independently of content would earn test scores close to 0. Research during the test development stage can identify the average range of scores on such a scale. Later, test takers scoring outside this range would be suspected of a response bias. This procedure is similar to the one used in the development of the MMPI-2's scale.

Response Sets and Response Styles. The tendency to choose response alternatives on the basis of social desirability is only one of several response sets that have been identified in self-report inventory responding (Lanyon 6k Goodstein, 1982, pp. 158-169). Although the voluminous literature on the operation of response sets in personality inventories dates largely from the 1950s, the influence of response sets in both ability and personality tests was observed by earlier investigators (see Block, 1965, chap. 2). One of the response sets that attracted early attention was acquiescence, or the tendency to answer "True" or "Yes." Acquiescence is conceptualized as a continuous variable; at one end of the scale are the consistent "Yeasayers" and at the other end the consistent "Naysayers" (Couch & Keniston, 1960). The implications of this response set for the construction of personality inventories is that the number of items in which a "Yes" or "True" response is keyed positively in any trait scale should equal the number of items in which a "No" or "False" response is keyed positively. This balance can be achieved by the proper selection or rewording of items, as was done in the PRF and is now being done with most new inventories.17

Another response set is deviation, or the tendency to give unusual or uncommon responses. Berg (1967) proposed this hypothesis and demonstrated its operation with nonverbal content in a specially developed test requiring an expression of preference for geometric figures. Scales made up of items likely to be answered in one direction by almost all test takers, such as the Infrequency scale of Jackson's PRF, were intended to identify such deviant response patterns. However, Jackson himself, among others, has pointed out that these scales tend to lack conceptual relevance to external criteria and, therefore, pose a problem especially in contexts like employment settings where the relevance of questionnaire items is considered important. Because of this, the Infrequency scale of the J PI was removed when that inventory was revised (Jackson, 1994a). The tendency to use the extreme choices on a rating scale (e.g., Is and 7s on a seven-point scale) has also been identified as a possible response bias (Paulhus, 1991).

Research on response sets such as social desirability, acquiescence, and deviation has passed through several stages. When first identified, response sets were regarded as a source of irrelevant or error variance to be eliminated from test scores. Later, these response sets came to be regarded as indicators of broad and durable personality characteristics that were worth measuring in their own right (Jackson & Messick, 1958, 1962; J. S. Wiggins, 1962). At this stage, they were commonly described as response styles and an elaborate edifice of empirical data was built around them. Eventually, these data were challenged from many directions (Block, 1965; Heilbrun, 1964; Rorer, 1965). Block (1965), for example, presented strong evidence supporting a content-oriented interpretation of the two major factors generally found to account for most of the common variance in the MMPI scales, which exponents of response sets and response styles had interpreted as social desirability and acquiescence.

The controversy over response sets and content-versus-style in personality assessment has never been fully settled (Edwards, 1990; Hogan 6k Nicholson, 1988; Jackson & Paunonen, 1980).18 The majority of test developers and investigators seem to agree that personality inventory scores are likely to reflect a combination of self-deception, impression management, and realistic self-portrayal and that the weight of each of these components will vary with the individual and the occasion. Some, however, view attempts to improve the trustworthiness of self-report data through special scales and items as possibly counterproductive in that they may reduce the validity of scales especially for normal, as opposed to pathological, samples. Such authors advocate the use of clinical skills in eliciting a patient's cooperation and in interpreting scores, as well as the inclusion of ratings from knowledgeable informants whenever there is reason to suspect serious distortion (see, e.g., Costa & McCrae, 1992a).

Most other workers, especially those involved in the assessment of psychopathology, continue to use so-called "validity" scales, with the awareness that they may also reflect personality styles and characteristics. In fact, some of the newest and technically more advanced instruments for the assessment of psychopathology, such as Jackson's BPI and Morey's PAI, use balanced keying of items and special scales for detecting invalidating response sets. There is also a new variety of such scales, exemplified by the VRIN and TRIN scales of the MMPI-2 and MMP1-A, that make use of specially selected item pairs that are either similar or opposite in content to detect inconsistent or contradictory responding. Because of the way they are constituted, the VRIN and TRIN scales, which are similar to Greene's (1978) Carelessness scale for the original MMPI, are not likely to be confounded by valid personality trait variance (Ozer 6k Reise, 1994).

At any rate, the argument about response sets and styles has stimulated extensive research and has produced several hundred publications. Like many scientific controversies, its net effect has been to sharpen our understanding of methodological problems and thereby improve the construction of personality inventories and their use in both research and applied settings.

Intelligence Testing: Theories and Preschool Assessment

This chapter opens an extended discussion of intelligence testing, a topic so important and immense that we devote the next two chapters to it as well. In order to understand contemporary intelligence testing, the reader will need to assimilate certain definitions, theories, and mainstream assessment practices. The goal of Topic 5A, Theories and the Measurement of Intelligence, is to investigate the various meanings given to the term intelligence and to discuss how definitions and theories have influenced the structure and content of intelligence tests. An important justification for this topic is that an understanding of theories of intelligence is crucial for establishing the construct validity of IQ measures. In Topic 5B, Assessment of Infant and Preschool Abilities, we review the nature and application of prominent infant assessment devices and then investigate a fundamental issue: What is the practical utility of these instruments. We begin with a review of early, traditional, and contemporary theories of intelligence.

Intelligence is one of the most highly searched topics in psychology. Thousands of search articles are published each year on the nature and measurement of intelligence. New journals such as Intelligence and The Journal of Psycho-educational Assessment have flourished in response to the scholarly interest in this topic. Despite this burgeoning research literature, the definition of intelligence remains elusive, wrapped in controversy and mystery. In fact, the discussion that follows will illustrate a major paradox of modern testing: Psychometricians are better at measuring intelligence than conceptualizing it!

Even though defining intelligence has proved to be a frustrating endeavor, there is much to be gained by reviewing historical and contemporary efforts to clarify its meaning. After all, intelligence tests did not materialize out of thin air. Most tests are grounded in a specific theory of intelligence and most test developers offer a definition of the construct as a starting point for their endeavors. For these reasons, we can better understand and evaluate the multifaceted character of contemporary tests if we first review prominent definitions and theories of intelligence.

DEFINITIONS OF INTELLIGENCE

Before we discuss definitions of intelligence, we need to clarify the nature of definition itself. Sternberg (1986) makes a distinction between operational and "real" definitions that is important in this context. An operational definition defines a concept in terms of the way it is measured. Boring (1923) carried this viewpoint to its extreme when he defined intelligence as "what the tests test." Believe it or not, this was a serious proposal, designed largely to short-circuit rampant and divisive disagreements about the definition of intelligence.

Operational definitions of intelligence suffer from two dangerous shortcomings (Sternberg, 1986). First, they are circular. Intelligence tests were invented to measure intelligence, not to define it. The test designers never intended for their instruments to define intelligence. Second, operational definitions block further progress in understanding the nature of intelligence, because they foreclose discussion on the adequacy of theories of intelligence.

This second problem—the potentially stultifying effects of relying upon operational definitions of intelligence—casts doubt upon the common practice of affirming the concurrent validity of new tests by correlating them with old tests. If established tests serve as the principal criterion against which new tests are assessed, then the new tests will be viewed as valid only to the extent that they correlate with the old ones. Such a conservative practice drastically curtails innovation. The operational definition of intelligence does not allow for the possibility that new tests or conceptions of intelligence may be superior to the existing ones.

We must conclude, then, that operational definitions of intelligence leave much to be desired. In contrast, a real definition is one that seeks to tell us the true nature of the thing being defined (Robinson, 1950; Sternberg, 1986). Perhaps the most common way—but by no means the only way—of producing real definitions of intelligence is to ask experts in the field to define it.

Expert Definitions of Intelligence

Intelligence has been given many real definitions by prominent researchers in the field. Following, we list several examples, paraphrased slightly for editorial consistency. The reader will note that many of these definitions appeared in an early but still influential symposium, "Intelligence and Its Measurement," published in the Journal of Educational Psychology (Thorndike, 1921). Other definitions stem from a modern update of this early symposium, What Is Intelligence? edited by Sternberg and Detterman (1986). Intelligence has been defined as the following:

Spearman (1904, 1923): a general ability that involves mainly the education of relations and correlates.

Binet and Simon (1905): the ability to judge well, to understand well, to reason well.

Terman (1916): the capacity to form concepts and to grasp their significance.

Pintner (1921): the ability of the individual to adapt adequately to relatively new situations in life.

Thorndike (1921): the power of good responses from the point of view of truth or fact.

Thurstone (1921): the capacity to inhibit instinctive adjustments, flexibly imagine different responses, and realize modified instinctive adjustments into overt behavior.

Wechsler (1939): The aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with the environment.

Humphreys (1971): the entire repertoire of acquired skills, knowledge, learning sets, and generalization tendencies considered intellectual in nature that are available at any one period of time.

Piaget (1972): a generic term to indicate the superior forms of organization or equilibrium of cognitive structuring used for adaptation to the physical and social environment.

Sternberg (1985a. 1986): the mental capacity to automatize information processing and to emit contextually appropriate behavior in response to novelty; intelligence also includes metacomponents, performance components, and knowledge-acquisition components (discussed later).

Eysenck (1986): error-free transmission of information through the cortex.

Gardner (1986): the ability or skill to solve problems or to fashion products that are valued within one or more cultural settings.

Ceci (1994): multiple innate abilities that serve as a range of possibilities; these abilities develop (or fail to develop, or develop and later atrophy) depending upon motivation and exposure to relevant educational experiences.

Sattler (2001): intelligent behavior reflects the survival skills of the species, beyond those associated with basic physiological processes.

The preceding list of definitions is representative although definitely not exhaustive. For one thing, the list is exclusively Western and omits several cross-cultural conceptions of intelligence. Eastern conceptions of intelligence, for example, emphasize benevolence, humility, freedom from conventional standards of judgment, and doing what is right as essential to intelligence. Many African conceptions of intelligence place heavy emphasis upon social aspects of intelligence such as maintaining harmonious and stable intergroup relations (Sternberg & Kaufman. 1998). The reader can consult Bracken and Fagan (1990). Sternberg (1994), and Sternberg and Detterman (1986) for additional ideas. Certainly, this sampling of vie' is sufficient to demonstrate that there appear to as many definitions of intelligence as there are experts willing to define it!

In spite of this diversity of viewpoints, themes recur again and again in expert definitions of intelligence. Broadly speaking, the experts tend to agree that intelligence is (1) the capacity to learn from experience, and (2) the capacity to adapt to one's environment. That learning and adaptation are both crucial to intelligence stands out poignancy in certain cases of mental disability in which persons fail to possess one or the other capacity in sufficient degree (Case Exhibit 8.1).

How well do intelligence tests capture the experts' view that intelligence consists of learning from experience and adaptation to the environment? The reader should keep this question in mind as we proceed to review major intelligence tests in the topics that follow. Certainly, there is cause for concern: Very few contemporary intelligence tests appear to require the examinee to learn something new or to adapt to a new situation as part and parcel of the examination process. At best, prominent modern tests provide indirect measures of the capacities to learn and adapt. How well they capture these dimensions is an empirical question that must be demonstrated through validational research.

Layperson and Expert Conceptions of Intelligence

Another approach to understanding a construct is to study its popular meaning. This method is more scientific than it may appear. Words have a common meaning to the extent that they help provide an effective portrayal of everyday transactions. If laypersons can agree on its meaning, a construct such as intelligence is in some sense "real" and therefore potentially useful. Thus, asking persons on the street, "What does intelligence mean to you?" has much to recommend it.

Sternberg, Conway, Ketron, and Bernstein 981) conducted a series of studies to investigate inceptions of intelligence held by American adults. In the first study, people in a train station, entering a supermarket, and studying in a college library were asked to list behaviors characteristic of different kinds of intelligence. In a second study— only one discussed here—both laypersons and experts (mainly academic psychologists) rated the importance of these behaviors to their concept of "ideally intelligent" person. The behaviors central to expert and lay conditions of intelligence turned out to be very similar; although not identical. In order of importance, experts saw verbal intelligence, problem-solving ability, and practical intelligence as crucial to intelligence. Laypersons regarded practical problem-solving ability, verbal ability, and social competence to be the key ingredients in intelligence. Of course, opinions were not unanimous; these conceptions represent the consensus view of each group. The components of intelligence and representative descriptors are shown in Table 8.2. In their conception of intelligence, experts placed more emphasis upon verbal ability than problem solving, whereas laypersons reverse these priorities. Nonetheless, experts and laypersons alike consider verbal ability and problem solving to be essential aspects of intelligence. As the reader will see, most intelligence tests also accent these two competencies. Prototypical examples would be vocabulary (verbal-ability) and block design (problem solving) from the Wechsler scales, discussed later. We see then that everyday conceptions of intelligence are, in part, mirrored quite faithfully by the content of modern intelligence tests.

Some disagreement between experts and laypersons is also evident. Experts consider practical intelligence (sizing up situations, determining how to achieve goals, awareness and interest in the world) an essential constituent of intelligence, whereas laypersons identify social competence (accepting others for what they are, admitting mistakes, punctuality, and interest in the world) as a third component. Yet, these two nominations do share one property in common: Contemporary tests generally make no attempt to measure either practical intelligence or social competence. Partly, this reflects the psychometric difficulties encountered in devising test items relevant to these content areas. However, the more influential reason intelligence tests do not measure practical intelligence or social competence is inertia: Test developers have blindly accepted historically incomplete conceptions of intelligence. Until recently, the development of intelligence testing has been a conservative affair, little changed since the days of Binet and the Army Alpha and Beta tests for World War I recruits. There are some signs that testing practices may soon evolve, however, with the development of innovative instruments. For example, Sternberg and colleagues have proposed innovative tests based upon his model of intelligence. Another interesting instrument based upon a new model of intelligence is the Everyday Problem Solving Inventory (Cornelius & Caspi, 1987). In this test, examinees must indicate their typical response to everyday problems such as failing to bring money, checkbook, or credit card when taking a friend to lunch.

Table 8.2 Factors and Sample Items Underlying Conceptions of Intelligence for Laypersons and Experts

Laypersons	Experts
Practical Problem-Solving Ability Reasons logically and well Identifies connections among ideas Sees all aspects of a problem Keeps an open mind	***Verbal Intelligence*** Displays a good vocabulary Reads with high comprehension Displays curiosity Is intellectually curious
Verbal Ability Speaks clearly and articulately Is verbally fluent Converses well Is knowledgeable about a particular field of knowledge	***Problem-Solving Ability*** Able to apply knowledge to problems at hand Makes good decisions Poses problems in an optimal way Displays common sense
Social Competence Accepts others for what they are Admits mistakes Displays interest in the world at large Is on time for appointments	***Practical Intelligence*** Sizes up situations well Determines how to achieve goals Displays awareness to world Displays interest in the world at large

We turn now to a review of major theories of intelligence. A reminder: The justification for reviewing theories is to illustrate how they have influenced the structure and content of intelligence tests. In addition, the construct validity of IQ tests depends upon the extent to which they embody specific theories of intelligence, so a review of theories is pertinent to test validation as well.

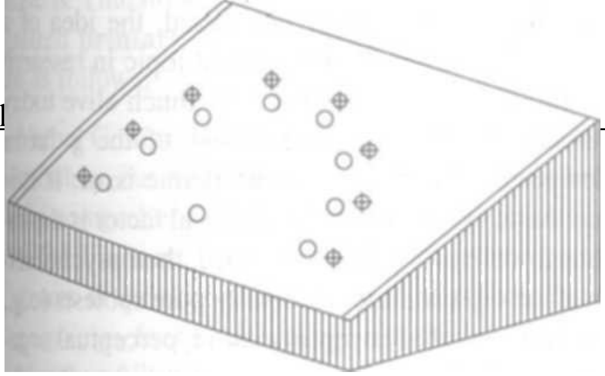


Figure 8.1 The Reaction Time-Movement Time Apparatus

THEORIES OF INTELLIGENCE

Galton and Sensory Keenness

The first theories of intelligence were derived in the Brass Instruments era of psychology at the turn of last century. The reader will recall from Topic 1 A that Sir Francis Galton and his disciple J. McKeen Cattell thought that intelligence was underwritten by keen sensory abilities. This incomplete and misleading assumption was based on a plausible premise:

The only information that reaches us concerning outward events appears to pass through the avenues of our senses; and the more perceptive the senses of difference the larger is the field upon which our judgment and intelligence can act. (Galton, 1883)

The sensory keenness theory of intelligence promoted by Galton and Cattell proved to be largely a psychometric dead end. However, we do see vestiges of this approach in modern chronometric analyses of intelligence such as the Reaction Time-Movement Time (RT-MT) apparatus, an experimental method favored by Jensen (1980) for the culture-reduced study of intelligence (Figure 8.1).

In RT-MT studies, the subject is instructed to place the index finger of the preferred hand on the home button; then an auditory warning signal is sounded, followed (in 1 to 4 seconds) by one of the eight green lights going on, which the subject must turn off as quickly as possible by touching the micro switch button directly below it. RT is the time the subject takes to remove his or her finger from the home button after a green light goes on. MT is the interval between removing the finger from the home button and touching the button that turns off the green light. Jensen (1980) reported that indices of RT and MT correlated as high as .50 with traditional psychometric tests of intelligence.1 R A. Vernon has also reported substantial relationships— as high as .70 for multiple correlations—between speed-of-processing RT-type measures and traditional measures of intelligence (Vernon, 1994; Vernon & Mori, 1990). These findings suggest that speed-of-processing measures such as RT might be a useful addition to standardized intelligence test batteries. In general, test developers have resisted the implications of this line of research.

Spearman and the g Factor

Based on extensive study of the patterns of correlations between various tests of intellectual and sensory ability, Charles Spearman (1904, 1923, 1927) proposed that intelligence consisted of two kinds of factors: a single general factor g and numerous specific factors s_1 s_2, s_3, and so on. As a necessary adjunct to his theory, Spearman helped invent factor analysis to aid his investigation of the nature of intelligence. Spearman used this statistical technique to discern the number of separate underlying factors that must exist to account for the observed correlations between a large number of tests.

In Spearman's view, an examinee's performance on any homogeneous test or subtest of intellectual ability was determined mainly by two influences: g, the pervasive general factor, and s, a factor specific to that test or subtest. (An error factor e could also sway scores, but Spearman sought to minimize this influence by using highly reliable instruments.) Because the specific factor a was different for each intellectual test or subtest and was usually less influential than g in determining performance level, Spearman expressed less interest in studying it. He concentrated mainly on defining the nature of g, which he likened to"energy" or "power" that serves in common the whole cortex. In contrast Spearman considered s, the specific factor, to have a physiological substrate localized in the group of neurons serving the particular kind of mental operation demanded by a test or subtest. Spearman (1923) wrote, "These neural groups would thus function as alternative 'engines' into which the common supply of 'energy' could be alternatively distributed."

Spearman reasoned that some tests were heavily loaded with the g factor, whereas other tests especially purely sensory measures—were representative mainly of a specific factor. Two tests each heavily loaded with g should correlate quite strongly. In contrast, psychological tests not saturated with g should show minimal correlation with one another. Much of Spearman's research was aimed at demonstrating the truth of these basic propositions derived from his theory. We have illustrated these points graphically in Figure 8.2. In this figure, each circle represents an intelligence test, and the degree of overlap between circles indicates the strength of correlation. Notice that tests

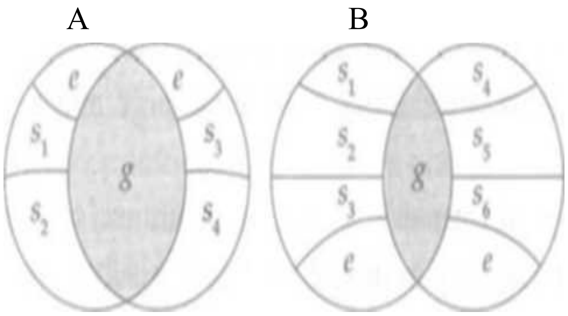


Figure 8.2 Spearman's Two-Factor Theory of Intelligence

A and B, each heavily loaded on g, correlate quite strongly. Tests C and D have weak loadings on g and subsequently do not correlate well.

Spearman (1923) believed that individual differences in g were most directly reflected in the ability to use three principles of cognition: apprehension of experience, education of relations, and education of correlations. Incidentally, the little-used term education refers to the process of figuring things out. These three principles can be explained by examining how we solve analogies of the form A:B:C:? that is, A is to B as C is to? A simple example might be HAMMER: NAIL::SCREWDRIVER:? To solve this analogy, we must first perceive and understand each term based on past experience; that is, we must have apprehension of experience. If we have no idea what a hammer, nail, and screwdriver are, there is little chance we can complete the analogy correctly. Next, we must infer the relation between the first two analogy terms, in this case, HAMMER and NAIL. Using a somewhat stilted phrase, Spearman referred to the ability to infer the relation between two concepts as education of relations. The final step, education of correlates, refers to the ability to apply the inferred principle to the new domain, in this case, applying the rule inferred to produce the correct response, namely, SCREWDRIVER:SCREW.

Although Spearman's physiological speculations have been largely dismissed, the idea of a general factor has been a central topic in research on intelligence and is still very much alive today (Jensen, 1979). The correctness of the g factor viewpoint is more than an academic issue. If it is true that a single, pervasive general factor is the essential wellspring of intelligence, then psychometric efforts to produce factorially pure subtests (e.g., measuring verbal comprehension, perceptual organization, short-term memory, and so on) are largely misguided. To the extent that Spearman is correct, test developers should forego subtest derivation and concentrate on producing a test that best captures the general factor.

The most difficult issue faced by Spearman's two-factor theory is the existence of group factors. As early as 1906, Spearman and his contemporaries noted that relatively dissimilar tests could have correlations higher than the values predicted from their respective^ loadings (Brody & Brody, 1976). This finding raised the possibility that a group of diverse measures might share in common a unitary ability other than g. For example, several tests might share a common unitary memorization factor that was halfway between the g factor and the various .v factors unique to each test. Of course, the existence of group factors is incompatible with Spearman's meticulous two-factor theory.

Thurstone and the Primary Mental Abilities

Thurstone (1931) developed factor-analysis procedures capable of searching correlation matrices for the existence of group factors. His methods permitted a researcher to discover empirically the number of factors present in a matrix and to define each factor in terms of the tests that loaded on it. In his analysis of how scores on different kinds of intellectual tests correlated with each other, Thurstone concluded that several broad group factors— and not a single general factor—could best explain empirical results. At various points in his research career, he proposed approximately a dozen different factors. Only seven of these factors have been frequently corroborated (Thurstone, 1938; Thurstone & Thurstone, 1941) and they have been designated primary mental abilities (PMAs). They are as follows:

- Verbal Comprehension: The best measure is vocabulary, but this ability is also involved in reading comprehension and verbal analogies.
- Word Fluency: Measured by such tests as anagrams or quickly naming words in a given category (e.g., foods beginning with the letter 5).
- Number: Virtually synonymous with the speed and accuracy of simple arithmetic computation.
- Space: Such as the ability to visualize how a three-dimensional object would appear if it was rotated or partially disassembled.
- Associative Memory: Skill at rote memory tasks such as learning to associate pairs of unrelated items.
- Perceptual Speed: Involved in simple clerical tasks such as checking for similarities and differences in visual details.
- Inductive Reasoning: The best measures of this factor involve finding a rule, as in a number series completion test.

Thurstone (1938) published the Primary Mental Abilities Test consisting of separate subtests, each designed to measure one PMA. However, he later acknowledged that his primary mental abilities correlated moderately with each other, proving the existence of one or more second-order factors. Ultimately, Thurstone acknowledged the existence of g as a higher-order factor. By this time, Spearman had admitted the existence of group factors representing special abilities, and it became apparent that the differences between Spearman and Thurstone were largely a matter of emphasis (Brody & Brody, 1976). Spearman continued to believe that g was the major determinant of correlations between test scores and assigned a minor role to group factors. Thurstone reversed these priorities.

P. E. Vernon (1950) provided a rapprochement between these two viewpoints by proposing a hierarchical group factor theory. In his view, g was the single factor at the top of a hierarchy that included two major group factors labeled verbal-educational (V:ed) and practical-mechanical-spatial-physical (k:m). Underneath these two major group factors were several minor group factors resembling the PMAs of Thurstone; specific factors occupied the bottom of the hierarchy.

Thurstone's analysis of PMAs continues to influence test development even today. Schaie (1983, 1985) has revised and modified the Primary Mental Abilities Test and used these measures in an enormously influential longitudinal study of adult intelligence. If intelligence were mainly a matter of g, then the group factors should change at about the same rate with aging. In support of the group factor approach to intellectual testing, Schaie (1983) reports that some PMAs show little age-related decrement (Verbal Comprehension, Word Fluency, Inductive Reasoning), whereas other PMAs decline more rapidly in old age (Space, Number). Thus, there may be practical real-world reasons for reporting group factors and not condensing all of intelligence into a single general factor.

R. Cattell and the Fluid/Crystallized Distinction

Raymond Cattell (1941, 1971) proposed an influential theory of the structure of intelligence that has been revised and extended by John Horn (1968, 1994). As did their predecessors, Cattell and Horn used factor analysis to study the structure of intelligence. But instead of finding a single general factor or a half dozen group factors, Cattell and Horn identified two major factors, which they labeled fluid intelligence (gf) and crystallized intelligence (gc).

Fluid intelligence is a largely nonverbal and relatively culture-reduced form of mental efficiency. It is related to a person's inherent capacity to learn and solve problems. Thus, fluid intelligence is used when a task requires adaptation to a new situation. By contrast, crystallized intelligence represents what one has already learned through the investment of fluid intelligence in cultural settings (e.g.. learning algebra in school). **Crystallized intelligence** is highly culturally dependent and is used for tasks that require a learned or habitual response. Since crystallized intelligence arises when fluid intelligence is applied to cultural products, we would expect these two kinds of intelligence to be correlated. In fact, it is commonly found that measures of crystallized and fluid intelligence correlate moderately (r = .5).

The abilities that make up fluid intelligence nonverbal and not heavily dependent upon exposure to a specific culture. For these reasons, Cattell (1940) believed that measures of fluid intelligence were culture-free. Based on this assumption, he devised the Culture Fair Intelligence Test in an attempt to eliminate cultural bias in testing. Of course, calling a test culture-fair does not make it necessarily so. In fact, the goal of a completely culture-free intelligence test has proved elusive. W< discuss the CFIT in more detail in Topic 6B, Group Tests of Intelligence.

In later versions of the fluid/crystallized theory of intelligence, Cattell (1971) and Horn (1982, 1994) expanded and elaborated on the previously discussed concepts. Today their approach might better be called a theory of many intelligences, the g_f-g_c designation has become so well known that it will not easily be phased out. In the latest revisions, the authors have proposed a hierarchical, interlocking model of intelligence with fluid and crystallized components at the top. These capacities are subserved by identified subcomponents of intelligence, including visual organization, perceptual speed, auditory organization, several memory capacities, and specific sensory reception components as well. The revised model is labyrinthine; interested readers should consult Horn (1994).

Piaget and Adaptation

The Swiss psychologist Jean Piaget (1896-1980) devised a theory of cognitive development that has a number of implications for the design of children's intelligence tests (Ginsburg & Opper, 1988). Piaget (1926, 1952, 1972) used interviews and informal tests with children to develop a series of provocative and revolutionary views about intellectual development. His new perspective included the following points:

- Children's thought is qualitatively different from adults' thought.
- Psychological structures called schemas are the primary basis for gaining new knowledge about the world.
- Four stages of cognitive development can be identified.

We examine each of these points in more detail in the following.

By studying the development of conservation, Piaget concluded that a child's construction of the world is fundamentally different from the adult perspective. **Conservation** refers to the awareness that physical quantities do not change in amount when they are superficially altered in appearance. For example, most adults know that two matching rows of 10 pennies are still equivalent if one row is spread out—adults possess conservation of number. But a young child will be easily misled by the superficial change in appearance and may insist that the second row now has more pennies. In a similar manner, it can be shown that young children do not possess conservation of continuous quantity, substance, weight, or volume.

In order to explain how infants and children gain new knowledge about the world, Piaget suggested that they form psychological structures called schemas. A **schema** is an organized pattern of behavior or a well-defined mental structure that leads to knowing how to do something. Perhaps a few examples will help clarify this difficult concept. Young infants possess schemas that are mainly sensorimotor in nature, such as the grasp-and-pull schema that allows a baby to retrieve a desired object and bring it up to the mouth. As we get older, we add mental structures to our collection of sensorimotor schemas. For example, teenagers usually possess the alphabetizing schema that permits them to find a word in a dictionary by repeatedly applying the simple rule that entries are alphabetical by first letter, then second letter, and so on.

Piaget's genius was in suggesting a mechanism by which schemas evolve toward greater and greater levels of complexity, thereby transforming into the more mature level of intellectual skill observed in most adults. The mechanism by which schemas become more mature is called the process of **equilibration**. To understand equilibration, the reader needs to know three additional Piagetian concepts: assimilation, accommodation, and equilibrium.

Assimilation is the application of a schema to an object, person, or event. For example, assimilation is involved when an infant uses the grasp-and-pull schema to retrieve a baby rattle and bring it to the mouth. If assimilation works to achieve the desired goals of the person, a state of harmony or equilibrium exists. But what happens if the application of the schema doesn't work? Suppose the grasp-and-pull schema is unsuccessful because the baby rattle snags on the vertical side bars of the crib as the infant seeks to bring the toy to the mouth. A state of dynamic tension will then arise, requiring the infant to adjust the schema so that it works. The adjustment of an unsuccessful schema so that it works is called **accommodation**. In our example of the infant using the grasp-and-pull schema to retrieve a baby rattle, the schema might be modified and become the grasp-and-pull-and-turn schema. If the modified schema is successful and allows the infant to bring the rattle to the mouth, a state of equilibrium exists once again. Note the distinction between equilibrium, the state of temporary harmony, and equilibration, the entire process of assimilation, accommodation, and equilibrium. Piaget believed that the striving toward equilibrium was an inherited characteristic of the human species.

Piaget also proposed four stages of cognitive development. According to his view, each stage is qualitatively different from the others and characterized by distinctive patterns of thought (Table 8.3). In the next topic (5B, Assessment of Infant and Preschool Abilities), we discuss an infant test based on a Piagetian analysis of cognitive development. In general, tests based upon these concepts seek to ascertain whether a child has passed certain cognitive milestones (e.g., conservation of volume) proposed by Piaget.

Guilford and the Structure-of-lntellect Model

After World War II, J. P. Guilford (1967, 1985) continued the search for the factors of intelligence that had been initiated by Thurstone. Guilford soon concluded that the number of discernible mental abilities was far in excess of the seven proposed by Thurstone. For one thing, Thurstone had ignored the category of creative thinking entirely, an unwarranted oversight in Guilford's view. Guilford also found that if innovative types of tests were included in the large batteries of tests he administered his subjects, then the pattern of correlations between these tests indicated the existence of literally dozens of new factors of intellect. Furthermore, Guilford noticed that some of these new factors had recurring similarities with respect to the kinds of mental processes involved, the kinds of information featured, or the form that the

Table 8.3 Piaget's Stages of Cognitive Development

Stage and Age Span	Characteristics of Thought
Sensorimotor: birth to 2 years	Infants experience the world mainly through their senses and motor abilities, act as if an object ceases to exist if it is not in sight, but develop object permanence by the end of this stage.
Preoperational:	Conservation concepts not yet developed, but these children do understand 2 to 6 years the idea of a functional relationship—for example, you pull on a cord to open a curtain, and the farther you pull, the more the curtain opens. Ability to mentally symbolize things with words and images also develops.
Concrete Operational:	Children typically develop conservation and demonstrate limited capacities 7 to 12 years of logical reasoning. For example, concept of reversibility develops the knowledge that one action can reverse or negate another.
Formal Operational: years and up	The systematic problem solving that we associate with adult thought usually 12 develops in this stage. There is a greater capacity to generate hypotheses and test them.

items of information took. As a result of these recurring similarities in the newly discovered factors of intellect, he became convinced that these multitudinous factors could be grouped along a small number of main dimensions. Guilford (1967) proposed an elegant structure-of-intellect (SOI) model to summarize his findings. Visually conceived, Guilford's SOI model classifies intellectual abilities along three dimensions called operations, contents, and products.

By operations. Guilford has in mind the kind of intellectual operation required by the test. Most test items emphasize just one of the operations listed here:

Cognition Discovering, knowing, or comprehending

Memory Committing items of information to memory, such as a series of numbers

Divergent production Retrieving from memory items of a specific class, such as naming objects that are both hard and edible

Convergent production Retrieving from memory a correct item, such as a crossword puzzle word

Evaluation Determining how well a certain item of information satisfies specific logical requirements

Contents refer to the nature of the materials o information presented to the examinee. The five content categories are as follows:

Visual Images presented to the eyes

Auditory Sounds presented to the ears

Symbolic Such as mathematical symbols that stand for something

Semantic Meanings, usually of word symbols

Behavioral The ability to comprehend the mental state and behavior of other persons

The third dimension in Guilford's model, prod-is, refers to the different kinds of mental structures that the brain must produce to derive a correct answer. The six kinds of products are as follows:

Unit A single entity having a unique combination of properties or attributes

Class What it is that similar units have in common, such as a set of triangles or high-pitched tones

Relation An observed connection between two items, such as two tones an octave apart

System Three or more items forming a recognizable whole, such as a melody or a plan for a sequence of actions

Transformation A change in an item of information, such as a correction of a misspelling

Implication What an individual item implies, such as to expect thunder following lightning

In total, then, Guilford (1985) identified five of operations, five types of content, and six of products, for a total of 5 x 5 x 6 or 150 factors of intellect. Each combination of an operation e.g. memory), a content (e.g.. symbolic), and a product (e.g., units) represents a different factor of intellect. Guilford claims to have verified over 100 these factors in his research.

The SOI model is often lauded on the grounds it captures the complexities of intelligence. However, this is also a potential Achilles' heel for theory. Consider one factor of intellect, memory for symbolic units. A test that requires the examinee to recall a series of spoken digits (e.g.. Digit Span on the WAIS-III) might capture this factor of intellect quite well. But so might a visual digit span test and perhaps even an analogous test with tactile presentation of symbols, such as vibrating rods applied to the skin. Perhaps we need a separate cube for hearing, vision, and touch; such an expanded model would incorporate 450 factors of intellect, surely an unwieldy number.

Although it seems doubtful that intelligence could involve such a large number of unique abilities, Guilford's atomistic view of intellect nonetheless has caused test developers to rethink and widen their understanding of intelligence. Prior to Guilford's contributions, most tests of intelligence required mainly convergent production—the construction of a single correct answer to a stimulus situation. Guilford raised the intriguing possibility that divergent production—the creation of numerous appropriate responses to a single stimulus situation—is also an essential element of intelligent behavior. Thus, a question such as "List as many consequences as possible if clouds had strings hanging down from them" (divergent production) might assess an aspect of intelligence not measured by traditional tests.

Theory of Simultaneous and Successive Processing

Some modern conceptions of intelligence owe a debt to the neuropsychological investigations of the Russian psychologist Aleksandr Luria (1902-1977). Luria (1966) relied primarily upon individual case studies and clinical observations of brain-injured soldiers to arrive at a general theory of cognitive processing. The heart of his theory is as follows:

> *Analysis shows that there is strong evidence for distinguishing two basic forms of integrative activity of the cerebral cortex by which different aspects of the outside world may be reflected. The first of these forms is the integration of the individual stimuli arriving in the brain into simultaneous and primarily spatial groups, and the second is the integration of individual stimuli arriving consecutively in the brain into temporally organized, successive series. (Luria, 1966)*

Since this approach focuses upon the mechanics by which information is processed, it is often called an information-processing theory.

Simultaneous processing of information is characterized by the execution of several different mental operations simultaneously. Forms of thinking and perception that require spatial analysis, such as drawing a cube, require simultaneous information processing. In drawing, the examinee must simultaneously apprehend the overall shape and guide hand and fingers in the execution of the shape. A sequential approach to drawing a cube (if one were even possible) would be horrifically complex. In effect, the examinee would have to draw individual lines of highly specific lengths and angular orientations, and just hope that everything would line up. In the absence of a simultaneous mental gestalt to guide the drawing, a distorted production is almost guaranteed. Luria discovered that simultaneous processing is associated with the occipital and parietal lobes in the back of the brain.

Successive processing of information is needed for mental activities in which a proper sequence of operations must be followed. This is in sharp contrast to simultaneous processing (such as drawing), for which sequence is unimportant. Successive processing is needed in remembering a series of digits, repeating a string of words (e.g., shoe, ball, egg), and imitating a series of hand movements (fist, palm, fist, fist, palm). Luria localized successive processing to the temporal lobe and the frontal regions adjacent to it.

Most forms of information processing require interplay of simultaneous and successive mechanisms. Das (1994) cites the example of reading an unfamiliar word such as taciturn:

> *The single letters are to be recognized, and that involves simultaneous coding. The reader matches the visual shape of the letter with a mental dictionary and comes up with a name for it. The letter sequences, then, have to be formed (successive coding) and blended together as a syllable (simultaneous). Then the string of syllables has to be made into a word (successive), the word is recognized (simultaneous), and a pronunciation program is then assembled (successive), leading to oral reading (successive and simultaneous).*

Das admits that this may be a simplified view of what occurs when a reader is confronted with a word. The essential point is that higher-level information processing relies upon interplay of specific, anatomically localizable forms of information processing.

The challenge of a simultaneous-successive approach to the assessment of intelligence is to design tasks that tap relatively pure forms of each approach to information processing. Tests that u this strategy are the Kaufman Assessment Battei for Children (K-ABC), discussed in the next topic, and the Das-Naglieri Cognitive Assessment System (Das & Naglieri, 1993). The Das-Naglieri battery includes successive tasks that involve rapid articulation (such as, "Say can, ball, hot as fast as you can 10 times") and simultaneous measures of both verbal and nonverbal tasks. The battery also assesses planning and attention, which leads to the acronym PASS (planning, attention, simultaneous, successive) (Das, Naglieri, & Kirby, 1994).

Information-Processing Theories of Intelligence

Information-processing conceptions of intelligence propose models of how individuals mentally represent and process information. Borrowing from Campione and Brown (1978), Borkowski (1985) has put forward a comprehensive theory that bears a loose analogy to the functioning of a computer. **The architectural system** (hardware) refers to biologically based properties necessary for information processing, such as memory span and speed of encoding/decoding information. Properties of the architectural system include capacity (e.g., number of slots in short-term memory, capacity of long-term memory), durability (rate of information loss), and efficiency of operation (e.g., rate of memory search). The architectural system is considered to be relatively "hard-wired" and impervious to change by the environment.

In addition to the structural component of intelligence, there are various functional components (software). **The executive system**, which refers to environmentally learned components that steer problem solving, provides overall guidance to the functional components. Elements of the executive system include the knowledge base (retrieval of knowledge from long-term memory), schemes (rules of thinking), control processes (rules and strategies such as self-checking and rehearsal), and metacognition (self-awareness of one's own thought processes). Metacognition is the process of thinking about thinking. Flavell (1976), who pioneered research on this topic, explained it as follows:

> *Metacognition refers to one's knowledge concerning one's own cognitive processes or anything related to them, e.g., the learning-relevant properties of information or data. For example, I am engaging in metacognition if I notice that I am having more trouble learning A than B; if it strikes me that 1 should double check C before accepting it as fact.*

The information-processing approach to intelligence has generated a large body of research, especially on the concept of metacognition. A consistent finding in this literature is that individuals who use metacognitive strategies perform at much higher levels than those who do not (Montague & Bos, 1990). For example, in a study of 32 Israeli kindergarten children who were taught metacognition related to mathematics, metacognitive skills explained more of the variance in mathematics performance than general ability (Mevarech, 1995). Metacognition is essential to intelligence and is one of the primary influences on student learning (Wang. Haertel, &Walberg. 1990).

Intelligence as a Biological Construct

Most investigators have studied intelligence in the traditional manner by developing tests of intellect and correlating scores with external criteria (e.g., school grades) or other test results. But a few researchers have sought to discern the nature of intelligence by looking at the properties of the brain itself. For example, Hynd and Willis (1985) provide an excellent survey of the neurological foundations of intelligence.

One important property of the brain required for intelligent behavior is the well-patterned and synchronized electrical activity of brain cells. Neurons must transmit precisely calibrated electrochemical impulses in order for sensation, perception, and higher thought processes to occur. The collective electrical activity of brain cells can be measured by placing electrodes on a person's scalp. The ongoing record of electrical activity shows spontaneous fluctuations over time but also demonstrates predictable patternings in response to certain stimuli. For example, an evoked potential can be measured by noting the pattern of brain waves that occurs in the quarter second or so after a light is flashed in a subject's eyes. An average evoked potential (AEP) is usually obtained from hundreds of such trials for a single individual. In this manner, an extremely consistent and distinctive pattern can be obtained for any individual.

Ertl and Schafer (1969) were among the first researchers to study the brain wave correlates of intelligence. They discovered that the waveform of the AEP has many more peaks and troughs for high-IQ subjects than for low-IQ subjects. Eysenck (1982) published similar findings, which we have reproduced here (Figure 8.3). Two colleagues of Eysenck, A. E. Hendrickson (1982), and D. E. Hendrickson (1982) noticed that the total length of the sinuous waveform of the AEP could be used as a biological index of intelligence. They laid a piece of string over each of the AEP waveforms reported by Ertl and Shafer (1969). The beginnings and ends of the strings were cut, the strings were tightly stretched into straight lines, then measured for length. The researchers were then able to compute the correlation between the string lengths and the published IQ scores. The result was an impressive value of r = .77. This correlation is as high as those reported between any two psychometric tests of intelligence. A purely biological measure of brain function (AEP waves) turns out to be an excellent predictor of intelligence as measured by traditional IQ tests.

In spite of these promising research findings, several investigators remain skeptical about the electrocortical correlates of intelligence. The correlations arise only under certain conditions, and attempts to replicate the results do not always succeed (Eysenck, 1994; Vernon & Mori, 1990). Gale and Edwards (1983) argue that mere correlational studies are not enough; we need a more theory bound orientation that links intelligence as a trait with information processing at the neural level. Efforts to formulate such a theory have been attempted (Deary, Hendrickson, & Burns, 1987). These and similar studies (e.g., Shucard & Horn, 1972) serve as a reminder that intelligence is somehow bound up in the physiological properties of the brain, even though we don't yet understand the precise biological characteristics that account for intelligence.

Haier and his colleagues have pursued a different path in their study of biological intelligence

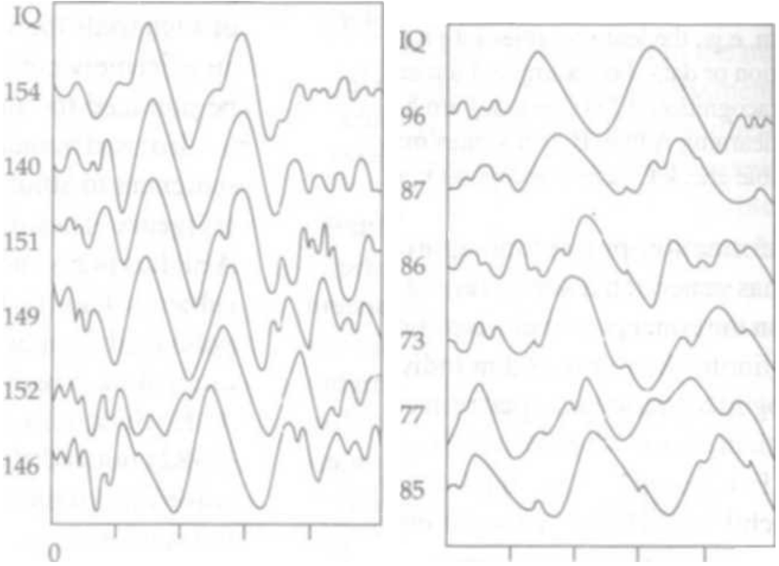


Figure 8.3 Averaged Evoked Potential (AEP) Waveforms for High-and Low-IQ Subjects

(Haier, Nuechterlein, Hazlett, and others, 1988; Haier, Siegel, Tang, and others, 1992). They measured cortical glucose metabolic rates as revealed by positron emission tomography (PET) scan analysis of volunteers solving intellectual problems. Brain cells use glucose and oxygen for fuel, so a PET scan will reveal "hot spots" at the most active brain sites (where glucose is being metabolized). Intriguingly, more-intelligent persons showed less brain activity when solving geometric analogy problems and when playing the Tetris computer game than less-intelligent persons. What remains unclear in this line of research is the causal direction: Are people smart because they use less glucose or do they use less glucose because they are smart? Another possibility is that both high IQ and low glucose metabolism are related to a third causal variable (Sternberg & Kaufman, 1998).

Gardner and the Theory of Multiple Intelligences

Howard Gardner (1983, 1993) has proposed a theory of multiple intelligences based loosely upon the study of brain-behavior relationships. He argues for the existence of several relatively independent human intelligences, although he admits that the exact nature, extent, and number of the intelligences has not yet been definitively established. Gardner (1983) outlines the criteria for an autonomous intelligence as follows:

- **Potential isolation by brain damage**—the faculty can be destroyed, or spared in isolation, by brain injury.
- **Existence of exceptional individuals such as savants**—the faculty is uniquely spared in the midst of general intellectual mediocrity.
- **Identifiable core operations**—the faculty relies upon one or more basic information-processing operations.
- **Distinctive developmental history**—the faculty possesses an identifiable developmental history, perhaps including critical periods and milestones.
- **Evolutionary plausibility**—admittedly speculative, a faculty should have evolutionary antecedents shared with other organisms (e.g., primate social organization).
- **Support from experimental psychology**—the faculty emerges in laboratory studies in cognitive psychology.
- **Support from psychometric findings**—the faculty reveals itself in measurement studies and is susceptible to psychometric measurement.
- **Susceptibility to symbol encoding**—the faculty can be communicated via symbols including (but not limited to) language, picturing, and mathematics.

Based upon these criteria, Gardner (1983, 1993) proposes that the following seven natural intelligences have been substantially confirmed. The seven intelligences are linguistic, logical, mathematical, spatial, musical, bodily-kinesthetic, interpersonal, and intrapersonal. Three of these seven types of intelligence are well known—linguistic (i.e., verbal) intelligence, logical-mathematical intelligence, spatial intelligence—and numerous formal tests have been devised to measure them, so we will not discuss them further here. The other four variations of intelligence are somewhat novel and therefore require more detailed presentation.

Bodily-kinesthetic intelligence includes the types of skills used by athletes, dancers, mime artists, typists, or "primitive" hunters. Although Western cultures are generally loath to consider the body as a form of intelligence, this is not the case in much of the rest of the world, nor was it true in our evolutionary history. Indeed, persons who could skillfully avoid predators, climb trees, hunt animals, and prepare tools were more likely to survive and pass on their genes to succeeding generations.

The personal intelligences include the capacity to have access to one's own feeling life (intrapersonal) as well as the ability to notice and make distinctions about the moods, temperaments, motivations, and intentions of others (interpersonal). Thus, personal intelligence encompasses both an intrapersonal and an interpersonal version. The former is found in great novelists who can write introspectively about their feelings, while the latter is often seen in religious and political leaders (e.g., Mahatma Ghandi or Lyndon Johnson) who can fathom the intentions and desires of others and use this information to influence them and form useful alliances.

Musical intelligence is perhaps the least understood of Gardner's intelligences. Persons with good musical intelligence easily learn to perform an instrument or to write their own compositions. Although knowledge of the structural aspects of melody, rhythm, and timbre is important to musical intelligence, Gardner notes that many experts place the affective or feeling aspects of music at its core. He believes that when the neurological underpinnings of music are finally unraveled, we will have "an explanation of how emotional and motivational factors are intertwined with purely perceptual ones" (Gardner, 1983).

The savant phenomenon provides strong support for the existence of separate intelligences, including musical intelligence. A savant is a mentally deficient individual who has a highly developed talent in a single area such as art, rapid calculation, memory or music.

An example is the extraordinary case of Leslie Lemke, who was born blind, and with mental retardation and cerebral palsy. He was not supposed to live. His adoptive mother had to coax him to suck milk from a bottle. Later, she strapped him to her back to help him learn to walk. In spite of his severe disabilities, Leslie became enamored of the piano and showed incredible precocity at picking out melodies on it. Within a few years, at the age of 18, he could listen to a piece of classical piano music a single time and then play it back flawlessly (Patton, Payne, & Beirne-Smith, 1986). The reader can find additional savant case studies in Miller (1989) and Treffert (1989).

Recently, Gardner (1998) has added three tentative candidates to his list of intelligences. These are naturalistic, spiritual, and existential intelligences. Naturalistic intelligence is the kind shown by people who are able to discern patterns in nature. Charles Darwin would be a prime example of such a person. Gardner believes that the evidence for this kind of intelligence is relatively strong. In contrast, spiritual intelligence (a concern with cosmic and spiritual issues in one's development) and existential intelligence (a concern with ultimate issues, including the meaning of life) are less well proved as independent intelligences. In general, the theory of multiple intelligence is compelling in its simplicity, but there is little empirical investigation of its validity.

Sternberg and the Triarchic Theory of Intelligence

Sternberg (1985b, 1986, 1996) takes a much wider view on the nature of intelligence than most previous theorists. In addition to proposing that certain mental mechanisms are required for intelligent behavior, he also emphasizes that intelligence involves adaptation to the real-world environment. His theory emphasizes what he calls successful intelligence or '"the ability to adapt to, shape, and select environments to accomplish one's goals and those of one's society and culture" (Sternberg & Kaufman, 1998. p. 494).

Sternberg's theory is called triarchic (ruled by three) because it deals with three aspects of intelligence: componential intelligence, experiential intelligence, and contextual intelligence. Each of these types of intelligence has two or more subcomponents. The entire theory is outlined in Table 8.4.

Table 8.4 An Outline of Sternberg's Triarchic Theory of Intelligence

Componential Intelligence

Metacomponents or executive processes (e.g., planning)

Performance components (e.g.. syllogistic reasoning)

Knowledge-acquisition components (e.g., ability to acquire vocabulary words)

Experiential Intelligence
Ability to deal with novelty
Ability to automatize information processing
Contextual Intelligence
Adaptation to real-world environment
Selection of a suitable environment
Shaping of the environment

Componential intelligence consists of the internal mental mechanisms that are responsible for intelligent behavior. The components of intelligence serve three different functions. Metacomponents are the executive processes that direct the activities of all the other components of intelligence. They are responsible for determining the nature of an intellectual problem, selecting a strategy for solving it and making sure that the task is completed. The metacomponents receive constant feedback as to how things are going in problem solving. Persons who are strong on the metacomponential aspect of intelligence are very good at allocating their intellectual resources.

In a problem-solving study using novel forms of analogies, Sternberg (1981) found that higher intelligence is associated with spending relatively more time on global or higher-order planning, and relatively less time on local or lower-order planning. For example, consider this analogy problem:

Man: Skin:: (Dog, Tree):(Bark, Cat)

The examinee must choose the two correct terms on the right that will complete the analogy. (The correct choices are Tree and Bark). Using reaction time measures for a series of such novel or nonentrenched problems, Sternberg (1981) found that persons of higher intelligence spend more time in global planning—forming a macro strategy that applies to this and similar problems—than did persons of lower intelligence. Thus, a crucial aspect of intelligence is knowing when to step back and allocate intellectual effort instead of obtusely attacking a difficult problem.

Performance components are the well-entrenched mental processes that might be used to perform a task or solve a problem. These aspects of intelligence are the ones that are probably measured the best by existing intelligence tests. Examples of performance components include short-term memory and syllogistic reasoning.

Knowledge-acquisition components are the processes used in learning. Sternberg has emphasized that in order to understand what makes some people more skilled than others; we must understand their increased capacity to acquire those skills in the first place. A case in point is vocabulary knowledge, which is learned mainly in context rather than through direct instruction. More-intelligent persons are better able to use surrounding contexts to figure out what a word means; that is, they have greater knowledge-acquisition skills. Their increased vocabulary results, in large measure, from their increased ability to "soak up" the meanings of words they see and hear in their environment. Thus, vocabulary is an excellent measure of intelligence because it reflects people's ability to acquire information in context.

The second aspect of Sternberg's theory involves experiential intelligence. According to the theory, a person with good experiential intelligence is able to deal effectively with novel tasks. This aspect of his theory explains why Sternberg is so critical of most intelligence tests. For the most part, the existing tests measure things already learned by presenting tasks that the subject has already encountered. According to Sternberg, intelligence also involves the capacity to learn and think within new conceptual systems, not just to deal with tasks already encountered. A second aspect of **experiential intelligence** is the ability to automatize or "make routine" tasks that are encountered repeatedly. An example of automatizing that applies to most of us is reading, which is carried out largely without conscious thought. But any task or mental skill can be automatized, if it is practiced enough. Playing music is an example of an extremely high-level skill that can become automatized with enough practice.

The third aspect of Sternberg's theory involves contextual intelligence. **Contextual intelligence** is defined as "mental activity involved in purposive adaptation to, shaping of, and selection of real world environments relevant to one's life" (Sternberg, 1986, p. 33). This aspect of Sternberg's theory appears to acknowledge that human behavior has been shaped by selective pressures during our evolutionary history. Contextual intelligence has three parts: adaptation, selection, and shaping.

Adaptation refers to developing skills required by one's particular environment. Successful adaptation will differ from one culture to the next. In the pygmy cultures of Africa, adaptation might involve the ability to track elephants and kill them with poison-tipped spears. In the Western industrial nations, adaptation might involve presenting oneself favorably in a job interview.

Selection might be called niche finding. This aspect of contextual intelligence involves the ability to leave the environment we are in and to select a different environment more suitable to our talents and needs. Feldman (1982) has illustrated how selection can operate in the career choices of gifted children, thereby determining whether they are highly accomplished as adults. She followed up on the Quiz Kids who were featured in radio and television shows of the 1950s. These were extremely bright children by conventional standards, most with IQs of 140 and higher. A few became highly successful as adults. However, most of them led rather ordinary lives, devoid of the spectacular accomplishments that might have been predicted from their childhood precocity. Those who were most successful had found occupations highly suited to their abilities and interests. In sum, they had selected environmental niches that fitted them well. Sternberg would argue that the ability to select such environments is an important aspect of intelligence.

Shaping is another way to improve the fit between oneself and the environment, especially when selection of a new environment is not practical. In this application of contextual intelligence, we shape the environment itself so that it better tits our needs. An employee who convinces the boss to do things differently has used shaping to make the work environment more suited to his or her talents.

Although Sternberg's triarchic theory is the most comprehensive and ambitious model yet proposed, not all psychometric researchers have rushed to embrace it. Detterman (1984) cautions that we should investigate the basic cognitive components of intelligence before introducing higher-order constructs that may be unnecessary. Rogoff (1984) questions whether the three subtheories (componential, experiential, contextual) are sufficiently linked. Other comments on the triarchic theory can be found in Behavioral and Brain Sciences (1984, pp. 287-304).

Whatever the final verdict on the triarchic theory of intelligence, Sternberg's insistence that intelligence has several components not measured by traditional tests rings true to anyone who has studied or administered these tests. He cites the case of a colleague who was asked to test a number of residents at an institution for those with mental retardation. These residents had just planned and successfully executed an escape from the security-conscious school, a feat requiring high levels of practical intelligence. Yet, when administered the Porteus Maze Test (Por-teus, 1965), a standardized test reputed to involve planning ability, they could not solve even the simplest maze correctly. Sternberg (1986) has made it clear that intelligence just has too many components to be measured by any single test.

Lesson 09

INTELLIGENCE TESTS

Assessment of Infant and Preschool Abilities

1. The infant and preschool period extends from birth to about age 6. An important application of infant and preschool tests is to help answer questions about developmental delay. Most infant tests (ages birth to 2V2) load heavily on sensory and motor skills. Preschool tests (ages 2V2 to 6) tend to tap cognitive skills to a significant degree.
2. The Gesell Developmental Schedules (GDS) gauge the developmental progress of babies from 4 weeks to 60 months of age.
3. The Neonatal Behavioral Assessment 5 (NBAS) assesses the newborn infant's behavior repertoire on 28 behavior items (scored on a 9-" scale), 18 reflexes (scored on a 4-point scale), 7 qualities of responsiveness. The instrument is sensitive to prenatal cocaine exposure and other neurotoxins. The NBAS is also used to sensitize? to the uniqueness of their infants.
4. The Ordinal Scales of Psychological development were designed as a Piagetian based measure of intellectual development (ages 2 weeks to two years) The scales measure development of ob-10 2 Immanence, means-ends, vocal and gestural imitation, operational causality, object relations in space, and schemes for relating to objects.
5. The Bayley Scales of Infant Development-II assess mental and motor development of children from 1 month to 42 months of age. The Bayley is very carefully standardized and highly reliable. Like other infant tests, very low scores predict an intellectually disabled outcome in later childhood, while near-normal and higher scores possess little predictive validity.
6. The Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R) is designed for children ages 3 years to 7 years and 3 months. The WPPSI-R contains three subtests not found on other Wechsler Scales: Sentences (oral memory); Geometric Designs (design copying); and Animal Pegs (coded placement of pegs).
7. The Standford-Binet: Fourth Edition (SB: FE) is a useful instrument for preschool assessment. Although the test is designed to yield four factor scores, Sattler's (1988) two factor solution to the reporting of SB:FE scores (Verbal Comprehension and Nonverbal Reasoning/Visualization) is the preferred approach with preschoolers.
8. The Kaufman Assessment Battery for Children (K-ABC) used for children ages 2:6 •"rough 12:5 years, is a combined measure of intelligence and achievement based upon the distinction between sequential processing (serial or temporal arrangement of stimuli) and simultaneous processing (synthesis and organization of stimuli in an immediate or wholistic fashion).
9. The McCarthy Scales of Children's Abilities are designed for children ages 2:6 to 8:6 years. The 18 subtests produce five different subscores and a General Cognitive Index (GCI) akin to an IQ. The subscores include verbal, perceptual-performance, quantitative, memory, and motor. Unfortunately, these five areas are not confirmed by independent factor analyses.
10. Designed for children ages 2 years 6 months through 17 years 11 months, the Differential Ability Scales consists of 17 cognitive subtests and 3 conormed achievement tests for school-aged children. Initial research indicates that the DAS yields reliable and reasonably independent subtest scores useful in the assessment of learning disability.
11. In general, infant test scores correlate positively but weakly with childhood test scores. Infant test scores must be interpreted with caution. An exception is very low infant test scores on such devices as the Bayley-II, which reliably predict developmental disability in childhood.
12. Tests of recognition memory in infants show promise as predictors of childhood intelligence. For example, in Fagan's studies, indices of simple visual habituation in infancy correlated .57 with picture vocabulary scores at age 7.

PRACTICAL UTILITY OF INFANT AND PRESCHOOL ASSESSMENT

The history of child assessment has shown time and again that, in general, test scores earned in the first year or two of life show minimal predictive validity. For example, in her review of infant intelligence testing, Goodman (1990) concludes:

> *If the successful prediction of adolescent and adult intelligence from early childhood scores is one of the great accomplishments of applied psychology, then the failure to predict intelligence from infancy to early childhood ranks as one of its greatest failures.*

Given this dismal record of repeated failures of predictive validity, we must ask a difficult question: What is the purpose and practical utility of infant assessment? In fact, infant tests do have an important but limited role to play. We return to that issue after a review of predictive studies.

Predictive Validity of Infant and Preschool Tests

With heterogeneous samples of normal children, the general finding is that infant test scores correlate positively but unimpressively with childhood test scores (Goodman, 1990; McCall, 1979). A few studies are more optimistic in tone (e.g., Wilson, 1983), but most researchers agree with McCall's (1976) conclusion:

> *Generally speaking, there is essentially no correlation between performance during the first six months of life with IQ score after age 5; the correlations are predominantly in the 0.20s for assessments made between 7 and 18 months of life when one is predicting IQ at 5-18 years; and it is not until 19-30 months that the infant test predicts 1-IQ in the range of 0.40-0.55.*

McCall (1979) reconfirmed his original conclusion in a later review, which we have summarized here. The reader will notice in Table 9.1 that the correlations between infant and school-age test sco do not exceed .40 until the subjects are at least months of age for the initial testing.

The findings with preschool tests are somewhat more positive in tone. The correlation between preschool test results and later IQ is typically strong, significant, and meaningful. The simplest way to investigate this question is to measure the stability of IQ results in longitudinal studies. In Table 9.2, we have summarized the age-to-age ability of children's IQ scores on the Stanford-Binet from the Fels Longitudinal Study, an early, classic follow-up investigation of children's intellectual and emotional development (Sontag, Baker, & Nelson, 1958). The lowest correlation in this table is .43, and that is between IQ tested at age 4 and again at age 12. What stands out in the table is the robustness of the link between IQ in preschool an later childhood. The older the child at initial testing, the stronger the relationship with later IQ. In fact, the results suggest that IQ becomes reasonably stable, on average, by 8 years of age.

Table 9.1 Summary of Correlations between Infant and Childhood Intelligence Test Scores in Normal Subjects

	Age of Childhood Test (Years)		
Age of Initial Infant Test (Months)	3-4	5-7	8-18
1-6	.21	.09	.06
7-12	.32	.20	.25
13-18	.50	.34	.32
19-30	.59	.39	.49

Table 9.2 Stability of IQ from 3 to 12 Years of Age

	Age at Retesting								
Age at Initial Testing	4	5	6	7	8	9	10	11	12
3	.83	.72	.73	.64	.60	.63	.54	.51	.46
4		.80	.85	.70	.63	.66	.55	.50	.43
5			.87	.83	.79	.80	.70	.63	.62
6				.83	.79	.81	.72	.67	.67
7					.91	.83	.82	.76	.73
8						.92	.90	.84	.83

9	.90	.82	.81
10		.90	.88
11			.90

Collectively, these findings confirm that infant tests generally have poor prognostic value, whereas preschool tests are moderately predictive of later intelligence. This brings us back to the question posed at the beginning of this section: What is the purpose and practical utility of infant assessment?

Practical Utility of the Bayley-ll and Other Infant Scales

The most important and justifiable use of infant tests is in screening for developmental disabilities. Although existing infant tests are poor predictors of childhood intelligence, an exception to this rule ,S Countered for infants who obtain a very low ^ore on the Bayley-II or other screening devices.

r example, infants who score two standard deviations below the mean on the Bayley, particularly on the Mental Scale, have a high probability of testing in the ranges of those with mental retardation later in life (Self & Horowitz, 1979; Goodman, Durieux-Smith, MacMurray, & Bernard, 1990).

With at-risk children, the correlation between infant test scores and later childhood IQ is much stronger than for samples of normal children. Mc-Call (1983) determined that the median correlation between infant test scores and childhood IQ at seven-year follow-up was a healthy .48. The most consistent finding is that a very low score on an infant test—two standard deviations below the mean and lower accurately prognosticates low IQ in childhood (Frankenburg, 1985). For example, studies with the Denver Developmental Screening Test-Revised (since revised and published as the Denver-II) revealed a false positive rate of only 5 to 11 percent, meaning that infants and preschoolers identified as at-risk rarely achieve normal range functioning. Studies with the Bayley Scales also conform to this pattern (e.g., VanderVeer & Schweid, 1974).

New Approaches to Infant Assessment

Lewis has argued that traditional infant tests overlook early information-processing behaviors, such as recognition memory and attentiveness to the environment that might better predict childhood cognitive function (Lewis & Sullivan, 1985). In one study, simple visual habituation to a novel stimulus (measured by the duration of fixation) assessed at three months of age correlated .61 with the Bayley Mental score at 24 months of age (Lewis & Brooks-Gunn, 1981). Using a similar paradigm, Fagan has reported comparable findings (Fagan, 1984; Fagan & Shepherd. 1986). For example, in one study he tested infant recognition memory at four to seven months with the habituation method (Fagan & Mc-Grath, 1981). In this approach, the infants first observed a picture of a baby's face for a short period of time and were then shown the same picture alongside an unfamiliar picture (e.g., picture of a bald-headed man). The investigators kept careful track of which picture the infants looked at most. The logic of the procedure is simple: Staring mainly at the new picture signifies that an infant recognizes the old picture; that is. an infant with good recognition memory prefers to look at something new. Preference for novelty—as measured by visual fixation time on the new picture—thus becomes an index of early recognition memory. Years later, the investigators administered the Peabody Picture Vocabulary Test (PPVT) to gauge early childhood intelligence. Infant, recognition memory scores and early childhood PPVT scores correlated .37 at four years of age and .57 at seven years of age. These correlations probably underestimate the predictive validity of infant memory tests insofar as the index of infant memory was an unreliable procedure based upon a small number of test items. Furthermore, the researchers assessed normal infants, which attenuated the correlations between predictor and criterion.

Infant cognitive measures possess a great deal of promise as predictors of childhood intelligence (Bornstein, 1994; Fagan & Haiken-Vasen In the years ahead, we may witness the ernergence of entirely new types of infant assessment devices based on the measurement of early memory, habituation and attentional capacities instead of sensorymotor abilities. A first step in this direction is Fagan's Test of Infant Intelligence (FTII, Fagan & Shepherd, 1986). a simple instrument based upon the methods previously outlined for measuring infant novelty preference and recognition memory The FTII yields a composite score that is based upon preference for novelty—as measured by visual fixation time on a new picture—averaged over several trials. The procedure shows very high interrater agreement (O'Neill, Jacobson. & Jacobson, 1994).

Initial validity studies of the FTII as a predictor of childhood intelligence are mixed in outcome. In one sample of 200 infants, the FTII scores obtained at 7 to 9 months correlated only .32 with Stanford-Binet IQ at age 3 (DiLalla, Thompson, Plomin, and others, 1990). In another recent study, overall correlations between FTII scores obtained at 7 to 9 months and WPPSI-R IQ at age 5 were around .2 for two Norwegian samples of healthy children (Andersson, 1996). These correlations do not support the use of the test as a screening tool in 0^ risk populations. However, the test may function better when used with at-risk infants. Nonetheless, further research is needed before we abandon traditional infant measures in favor of the Fagan test and similar measures.

Individual and Group Tests

Intelligence testing is one of the major achievements of psychology in the twentieth century. In response to the success of the Binet-Simon scales in the early 1900s, psychologists developed and refined dozens of individual tests of intelligence patterned after this pathbreaking instrument. Explosive growth was also observed in group tests of intelligence, fostered by the enthusiastic acceptance of the Army Alpha and Beta tests during and after World War I. With only a few exceptions, contemporary individual and group tests of intelligence owe their lineage to Binet, Simon and the Army testing program.

The purpose of this chapter is to provide overview of noteworthy approaches to the testing of individual and group intelligence. We survey prominent individual tests in Topic 6A and then close the chapter with a review of group intelligence tests in Topic 6B. Even though this text devotes three full chapters to the fascinating and emotionally charged topic of intelligence testing, we make no pretext that the coverage is comprehensive. An exhaustive analysis of intelligence testing is simply beyond the scope of this or any other basic reference. New and revised tests appear practically every month, and thousands of new research findings are published every year. We have chosen to review tests that are widely used or that illustrate interesting developments in theory or method. Research can find information on additional tests in the Mental Measurements Yearbook series, now published every three or four years by the Buros institute (e.g.. Conoley & Kramer, 1989, 1992; Impara & Plake, 1998; Plake & Impara, 2001). The Encyclopedia of Human Intelligence (Sternberg, 1994) is also a good source of information on individual and group tests of intelligence.

ORIENTATION TO INDIVIDUAL INTELLIGENCE TESTS

The individual intelligence tests reviewed in this topic include the following:

- Wechsler Adult Intelligence Scale-Ill (WAIS-III)
- Wechsler Intelligence Scale for Children-Ill (WISC-III)
- Stanford-Binet: Fifth Edition (SB5)
- Detroit Test of Learning Aptitude-4 (DTLA-4)
- Kaufman Brief Intelligence Test (K-BIT)

Another promising test that we do not review in depth is the Kaufman Adolescent and Adult Intelligence Test (KAIT). Published in 1992, the KAIT is a recent arrival on the testing scene (Dumont & Hagberg, 1994; Shaughnessy & Moore, 1994). Kaufman and Kaufman (1997) list several advantages of the KAIT, including its psychometric foundation in the g_c-g_f distinction proposed by John Horn and his followers. The KAIT also is appealing because of its brevity: The test provides highly reliable indices of intelligence in two-thirds the time needed for most batteries. Along with the preschool tests presented in the previous topic, the previously listed instruments probably account for 98 percent °f the intellectual assessments conducted in the United States.

The Wechsler scales have dominated intelligence testing in recent years, but they are by no means the only viable choices for individual assessment. Many other instruments measure general intelligence just as well—some would say better. Consider the implications of a now familiar observation: For large, heterogeneous samples, scores on any two mainstream instruments (e.g., Wechsler, Stanford-Binet, McCarthy, Kaufman scales) typically correlate 0.80 to 0.90. Often, the correlation between two mainstream instruments is nearly as high as the test-retest correlation for either instrument alone. For purposes of producing a global score, it would appear that any well-normed mainstream intelligence test will suffice.

But producing an overall score is not the only goal of assessment. In addition, the examiner usually desires to gain an understanding of the subject's intellectual functioning. For this purpose, the overall IQ is important, .but there are instances in which the global score may be irrelevant or even misleading. To understand a referral's intellectual functioning, the examiner should also inspect the subtest scores in search of hypotheses that might explain the unique functioning of that individual. Of course, examiners need to undertake subtest analysis cautiously, armed with research-based findings on the nature and meaning of subtest scatter for the test in use (Gregory, 1994b; McLean, Kaufman, & Reynolds, 1989; McDermott, Fantuzzo, & Glutting, 1990).

If the examiner's goal is to understand intellectual functioning and not merely to determine an overall score, the differences between tests become quite real. Every instrument approaches the measurement of intelligence from a different perspective and yields a distinctive set of subtest scores. Furthermore, a test well suited for one referral issue might perform abysmally in another context. For example, the WAIS-III performs admirably in the testing of mild mental retardation, but contains too few simple items for the effective assessment of persons with moderate or severe developmental disability.

A central axiom of assessment is that the choice of a testing instrument should be based on knowledge of its strengths and weaknesses as they pertain to the referral question. Put simply, the skilled examiner does not blindly rely upon a single test for every referral! Instead, the skilled examiner flexibly chooses one or more instruments in light of the perceived assessment needs of the examinee. Each of the tests discussed in this topic has its special merits and also its particular shortcomings. The test user must know these strong and weak facets in order to choose the instruments best suited for each unique referral.

THE WECHSLER SCALES OF INTELLIGENCE

Beginning in the 1930s, David Wechsler, a psychologist at Bellevue Hospital in New York City, conceived a series of elegantly simple instruments that virtually defined intelligence testing in the mid-to late twentieth century. His influence on intelligence testing is exceeded only by the path breaking contributions of Binet and Simon. It is fitting that we begin the survey of individual tests with a historical summary of the Wechsler tradition, followed by a discussion of individual instruments.

Origins of the Wechsler Tests

Wechsler began work on his first test in 1932, seeking to devise an instrument suitable for testing the diverse patients referred to the psychiatric section of Bellevue Hospital in New York (Wechsler, 1932). In describing the development of his first test, he later wrote, "Our aim was not to produce a set of brand new tests but to select, from whatever source available, such a combination of them as would meet the requirements of an effective adult scale" (Wechsler, 1939). In fact, the content of his scales was largely inspired by earlier efforts such as the Binet scales and the Army Alpha and Beta tests (Frank, 1983). Readers who peruse Psychological Examining in the United States Army, a volume edited by Yerkes (1921) just after World War I, might be astonished to discover that Wechsler purloined dozens of test items from this source, many of which have survived to the present day in contemporary revisions of the Wechsler tests. Wechsler was not so much a creative talent as a pragmatist who fashioned a new and useful instrument from the spare parts of earlier, discontinued attempts at intelligence testing.

The first of the Wechsler tests, named the Wechsler-Bellevue Intelligence Scales, was published in 1939. In discussing the rationale for his new test, Wechsler (1941) explained that existing instruments such as the Stanford-Binet were fully inadequate for assessing adult intelligence. The Wechsler Bellevue was designed to rectify several flaws noted in previous tests:

- The test items possessed no appeal for adults.
- Too many questions emphasized mere manipulation of words.
- The instructions emphasized speed at the expense of accuracy.
- The reliance upon mental age was irrelevant to adult testing.

To correct these shortcomings, Wechsler designed his test specifically for adults, added performance items to balance verbal questions, reduced the emphasis upon speeded questions, and invented a new method for obtaining the IQ. Specifically, he replaced the usual formula

$$IQ = \frac{\text{Mental Age}}{\text{Chronological Age}}$$

with a new age-relative formula

$$IQ = \frac{\text{Attained or Actual Score}}{\text{Expected Mean Score for Age}}$$

This new formula was based on the interesting presumption—stated in the form of an axiom—that IQ remains constant with normal aging, even though raw intellectual ability might shift or even decline. The assumption of IQ constancy is basic to Wechsler scales. As Wechsler (1941) put it:

The constancy of the I.Q. is the basic assumption of all scales where relative degrees of intelligence are defined in terms of it. It is not only basic, but absolutely necessary that I.Q. be independent of the age at which they are calculated, because unless the assumption holds, no permanent scheme of intelligence classification is possible.

Although Wechsler's view has been largely accepted by contemporary test developers, it important to stress that the assumption of IQ variance with age is really a statement of values, a philosophical choice, and not necessarily an inherent characteristic of human nature.

Wechsler also hoped to use his test as an aid in psychiatric diagnosis. In pursuit of this goal, he divided his scale into separate verbal and performance sections. This division allowed the examiner to compare an examinee's facility in using words and symbols (verbal subtests) versus the ability to manipulate objects and Perceive visual patterns (performance subtests).[Large differences between verbal ability (V)and performance ability (P) were thought to be diagnostic significance.

Specifically, Wechsler believed that organic brain disease, psychoses, and emotional disorders gave rise to a marked V> P pattern, whereas adolescent psychopaths and persons with mild mental retardation yielded a strong P > V pattern. Subsequent research demonstrated many exceptions to these simple diagnostic rules. Nonetheless, the distinction between verbal and performance skills has proved valid and useful for other purposes, such as analyzing brain-behavior relationships and studying age effects on intelligence. Wechsler's armchair division of subtests into verbal and performance sections ranks as perhaps his most enduring contribution to contemporary intelligence testing (Kaufman. Lichtenberger. & McLean, 2001).

General Features of the Wechsler Tests

Including revisions, David Wechsler and his followers produced 10 intelligence tests in a span of about 60 years. A major reason for the success of these instruments was that each new test or revision remained faithful to the familiar content and format first introduced in the Wechsler-Bellevue. By sticking with a single successful formula, Wechsler ensured that examiners could switch from one Wechsler test to another with minimal retraining. This was not only good psychometrics but also shrewd marketing insofar as it guaranteed several generations of faithful test users.

The various versions and editions of the Wechsler tests possess the following common features:

- Ten to fourteen subtests. The multi-subtest approach allows the examiner to analyze intraindividual strengths and weaknesses rather than just to compute a single global score. As the reader will learn subsequently, the pattern of subtest scores may convey useful information not evident from the overall level of performance.
- A Verbal Scale composed of five or six subtests and a Performance Scale composed also of five or six subtests. With this division, the examiner can assess verbal comprehension and perceptual organization skills separately. The pattern of abilities on these two factors of intelligence may have a bearing on the functional integrity of the left and right hemispheres of the brain, as well as indicating vocational strengths and weaknesses, as discussed in the following.
- A common metric for IQ and Index scores. The mean for IQ and Index scores is 100 and the standard deviation is .15 for all tests and all age groups. In addition, the scaled scores on each subtest have a mean of 10 and a standard deviation of approximately 3, which permits the examiner to analyze the subtest scores of the examinee for relative strengths and weaknesses.

- Common subtests for different ages. For example, the preschool, child, and adult Wechsler tests (WPPSI-R, WISC-III, and WAIS-III) all contain a common core of the same eight subtests (Table 6.1). An examiner who masters the administration of one core subtest on any of the Wechsler tests (such as the Information subtest on the WAIS-III) easily can transfer this skill within the Wechsler family of intellectual measures.

THE WECHSLER SUBTESTS: DESCRIPTION AND ANALYSIS

Wechsler (1939) defined intelligence as "the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment." He also believed that we can only know intelligence by what it enables a person to do. In designing his tests, then, Wechsler selected components to represent a wide array of underlying abilities so as to estimate the global capacity of intelligence. Furthermore, he asked his subjects to do things, not merely to answer questions. The Wechsler subtests are quite diverse and often rely upon what Wechsler referred to as "mental productions."

Table 9.3 Subtest Composition of the Wechsler Scales

	WPPSI-R	**WISC-III**	**WAIS-III**
Verbal Scales			
Information	x	x	x
Digit Span		x	x
Vocabulary	x	x	x
Arithmetic	x	x	x
Comprehension	x	x	x
Similarities	x	x	x
Sentences	x		
Letter-Number Sequencing			x
Performance Scales			
Picture Completion	x	x	x
Picture Arrangement		x	x
Block Design	x	x	x
Matrix Reasoning			x
Object Assembly	x	x	x
Coding/Digit Symbol		x	x
Mazes	x	x	
Geometric Design	x		
Symbol Search		x	x
Animal Pegs	x		

We present here a description of subtests from the WISC-III and WAIS-III. We also analyze the abilities tapped by each subtest and offer research-based comments. The reader is referred to Topic 5B for a description of three subtests unique to the WPPSI-R (Sentences, Geometric Designs, and Animal Pegs). The verbal subtests are listed first.

Information

Factual knowledge of persons, places, and common phenomena is tested here. Questions for children are like the following:

"How many eyes do you have?"
"Who invented the telephone?"
"What causes a solar eclipse?"
"Which is the largest planet?"

Questions for adults are similar, but progress higher levels of difficulty. Difficult questions on I adult Information subtest resemble:

"Which is the most common element in air?"
"What is the population of the world?"
"How does fruit juice get converted to wine?
"Who wrote Madame Bovary?"

Information items test general knowledge normally available to most persons raised in cultural institutions and educational systems Western industrialized nations. Indirectly, this subtest measures learning and memory skills insofar, subjects must retain knowledge gained from for and informal educational opportunities in order I answer the Information items.

Information is usually regarded as one of best measures of general ability among the Wechsler subtests (Kaufman, McLean, & Reynold 1988). For example, the WAIS-III manual reveals that Information typically has the second or highest correlation with Full Scale IQ across the age groups (Tulsky, Zhu, & Ledbetter, 1997). Information consistently loads strongly on the factor identified in factor analyses of the WAIS-subtest correlations (see the following). The factor is labeled Verbal Comprehension. However Information tends to reflect formal education motivation for academic achievement and may therefore yield spuriously high ability estimates for perpetual students and avid readers.

Digit Span

Digit Span consists of two separate sections, Digits Forward and Digits Backward. In Digits Forward, the examiner reads a series of digits at one per second, then asks the subject to repeat them. If the subject answers correctly on two consecutive trials of the same length, the examiner proceeds to the next series, which is one digit longer, up to a maximum length of nine digits. For Digits Backward, a similar procedure is used, except the examinee must repeat the digits in reverse order, up to a maximum length of eight digits. For example, the examiner reads:

“6_l_3-4-2-8-5”

and the subject tries to repeat the numbers in the reverse order:

"5-8-2-4-3-1-6."

Digit Span is a measure of immediate auditory recall for numbers. Facility with numbers, good attention, and freedom from distractibility are required. Performance on this subtest may be affected by anxiety or fatigue, and many clinicians have noted that patients hospitalized for medical or psychiatric reasons frequently perform poorly on Digit Span.

Digits Forward and Digits Backward may assess fundamentally different abilities. Digits Forward seems to require the examinee to access an auditory code in sequential fashion. In contrast, to perform Digits Backward, the examinee must form an internal visual memory trace from the orally presented numerical sequences and then visually scan from end to beginning. Digits Backward is clearly the more complex test: not surprisingly, it loads higher on general intelligence than does Digits Forward (Jensen & Osborne, 1979). Gardner (1981) argues that examiners should supplement standard sporting procedures and list separate subscores for Digit Span. He*1 presents separate means, standard deviations, and percentile ranks on Digits Forward and Backward for children ages 5 to 15.

Vocabulary

The subject is asked to define up to several dozen words of increasing difficulty while the examiner writes down each response verbatim. For example, n an easy item the examiner might ask, "What is cup? and the examinee would get partial credit r answering, "You drink with it" and full credit or answering, "It has a handle, holds liquids, and you drink from it." For adults and bright children, the advanced items on the Wechsler Vocabulary subtests can be very challenging, on a par with tincture, obstreperous, and egregious.

Vocabulary is learned largely in context from reading books and listening to others. It is a rare individual who picks up vocabulary by reading the dictionary or memorizing word lists from the "Building Your Wordpower" section of popular magazines. In the main, a person's vocabulary is a measure of sensitivity to new information and the ability to decipher meanings based on the context in which words are encountered. Precisely because the acquisition of word meaning depends upon contextual inference, the Vocabulary subtest turns out to be the single best measure of overall intelligence on the Wechsler scales (Gregory, 1999). This is a surprise to many laypersons who regard vocabulary as merely synonymous with educational exposure and therefore a mediocre index of general intelligence. However, there is simply no denying the empirical evidence: Vocabulary has the highest subtest correlation with Full Scale IQ on both the WISC-III (combined age groups) and also the WAIS-III (for 12 of the 13 age groups).

Arithmetic

Except for the very easiest items for young people or persons who have mental retardation, the Arithmetic subtest consists of orally presented mathematics problems. The examinee must solve the problems without paper or pencil within a time limit (usually 30 to 60 seconds). The simple items stress fundamental operations of addition or subtraction, for example:

"If you have fifteen apples and give seven away, how many are left?"

The more difficult items require proper conceptualization of the problem and the application of two arithmetic operations, for example:

"John bought a stereo that was marked down 15 percent from the original sales price of $600. How much did John pay for the stereo?"

Although the mathematical requirements of the 11 Arithmetic items are not excessively demanding the necessity of solving the problems mentally within a time limit makes this subtest quite challenging for most examinees. In addition to rudimentary arithmetic skills, successful performance on Arithmetic requires high levels of concentration and the ability to maintain intermediate calculations in short-term memory. In factor analyses of the WISC-III and WAIS-III, Arithmetic often loads on a third factor variously interpreted as Freedom from Distractibility or Working Memory.

Comprehension

The Comprehension subtest is an eclectic collection of items that require explanation rather than mere factual knowledge. The easy questions stress common sense, whereas the more difficult questions require an understanding of social and cultural conventions. On the WAIS-III, the two most difficult questions require the examinee to interpret proverbs.

An easy item on Comprehension is of the form "Why do people wear clothes?" Difficult items resemble the following:

"What does this saying mean: A bird in the hand is worth two in the bush; "

"Why are Supreme Court Judges appointed for life?"

Comprehension would appear to be, in part, a measure of "social intelligence" in that many items tap the examinee's understanding of social and cultural conventions. Sipps, Berry, and Lynch (1987) found that Comprehension scores were moderately related to measures of social intelligence on the California Psychological Inventory. Of course, a high score signifies only that the examinee is knowledgeable about social and cultural conventions: choosing right action may or may not flow from this knowledge. However, recent studies by Campbell and McCord (1996) and Lipsitz, Dworkin, and Erlenmeyer-Kimling (1993) provide no support for the commonly accepted clinical lore that Comprehension scores are sensitive to social functioning.

Similarities

In this subtest, the examinee is asked questions the type, "In what way are shirts and socks alike The Similarities subtest evaluates the examinee ability to distinguish important from unimportant resemblances in objects, facts, and ideas. Indirectly these questions assess the assimilation of the concept of likeness. The examinee must also possess the ability to judge when a likeness is important rather than trivial. For example, "shirts" "socks" are alike in that both begin with the letter s, but this is not the essential similarity between these two items. The important similarity is shirts and socks are both exemplars of a concept namely, "clothes." As this example illustrates, Similarities can be thought of as a test of verbal concept formation.

We turn now to a description and analysis Wechsler performance subtests. With the exception of Matrix Reasoning on the WAIS-III, all of the performance subtests are timed, and for most the examinee earns bonus points for quick performance.

Letter-Number Sequencing

This is a new subtest found only on the WAIS-The examiner orally presents a series of letters numbers that are in random order. The examinee must reorder and repeat the list by saying the numbers in ascending order and then the letters in alphabetical order. For example, if the examiner says "R-3-B-5-Z-1-C," the examinee should respond "1-3-5-B-C-R-Z." This test measures attention concentration, and freedom from distractibility. Together with Arithmetic and Digit Span, this subtest contributes to the Working Memory Index score on the WAIS-ITI (see the following). Donders. Tulsl and Zhu (2001) found the Letter-Number Sequencing subtest to be highly sensitive to the effects of moderate and severe traumatic brain injury.

Picture Completion

For this subtest, the examiner asks the subject to identify the "important part" that is missing from picture. For example, a simple item might be of this type: a picture of a table with one leg missing.

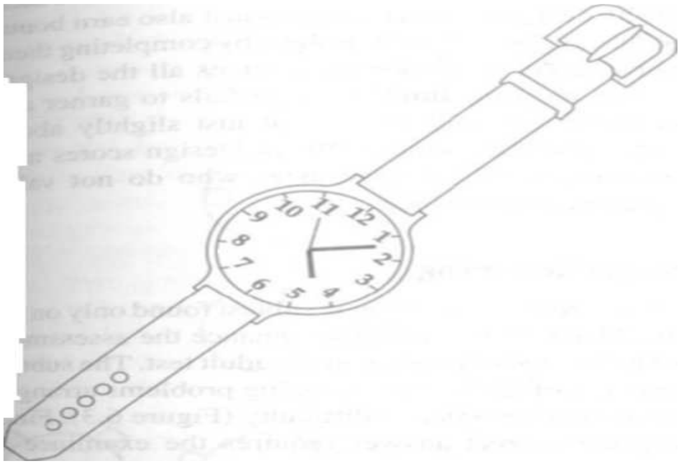


Figure 9.1 Picture Completion Item Similar to Those Found on the WAIS-III

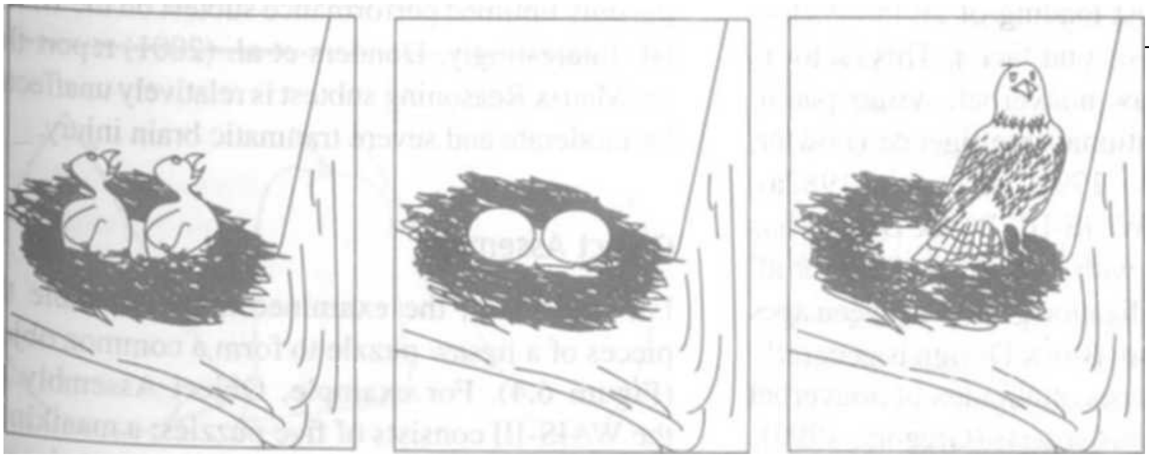
The items get harder and harder; testing continues until the examinee testing continues until the examinee misses several in a row. Figure 9.1 depicts an item similar to those found on the WAIS III.

Although Picture Completion is included on the performance half of each Wechsler test, the abilities required for this subtest overlap only modestly with the classic measures of performance intelligence (e.g., Block Design). For one thing, successful performance on Picture Completion largely involves access to long-term memory rather than perceptual-manipulative skill. True, the examinee must have good attention to visual detail. But high scores mainly reflect the ability to compare each drawing with similar items or situations stored in long-term memory. In sum Picture Completion really doesn't require a performance component. The examinee needs to verbalize the missing element or merely point to the section of the drawing that is anomalous. The Picture Completion subtest presupposes that the examinee has been exposed to the object or situation represented. For this reason, Picture Completion may be inappropriate for culturally disadvantaged persons.

Picture Arrangement

In this subtest, several panels of nonverbal cartoon strips are laid down out of order by the examiner. The examinee's task is to put the panels together in the correct order to tell a sensible story. Figure 9.2 depicts a picture arrangement task, such as might be found on the WAIS-III.

Figure 9.2 Picture Arrangement Item Similar to Those Found on the WAIS-III



Although Picture Arrangement is grouped with the Performance tasks, it loads about equally on the verbal and performance components revealed in factor-analytic studies of subtest intercorrelations (e.g., Silverstein, 1982a). The abilities tapped by Picture Arrangement are complex and multifaceted. Before sorting the pictures, the examinee must be able to decipher the gestalt of the entire story from its disarranged elements. This subtest also measures sequential thinking and the ability to see relationships between social events. On the WAIS-III, several of the Picture Arrangement stories have humorous themes. As a result, social sophistication and a sense of humor are required for successful performance.

Block Design

On the Block Design subtest, the examinee must reproduce two-dimensional geometric designs by proper rotation and placement of three-dimensional colored blocks. This subtest was depicted in Topic 2B, The Testing Process. For all of the Wechsler scales, the first few Block Design items can be solved through trial and error. However, the more difficult items require the analysis of spatial relations, visual-motor coordination, and the rigid application of logic. Block Design demands much more problem-solving and reasoning ability than most of the Performance subtests in which memory and prior experience are more heavily weighted. In factor analyses of the Wechsler scales, Block Design typically has the highest loading of all the Performance subtests on the second factor. This factor is variously identified as nonverbal, visuospatial, or perceptual-organizational intelligence (Fowler, Zillmer, & Macciocchio, 1990; Silverstein, 1982a). On the WISC-III and WAIS-III, Block Design has the highest correlation with Performance IQ for all but a few of the standardization groups between ages 6 and 89. For this reason, Block Design is generally recognized as the quintessential index of nonverbal intelligence on the Wechsler tests (Gregory, 1999).

Block Design is a strongly speeded test. Consider the WAIS-R version, which consists of 14 designs of increasing difficulty. To obtain a high score on this subtest, adults must not only reproduce of the designs correctly, they must also earn bon points on the last eight designs by completing the quickly. An examinee who solves all the designs within the time limit but who fails to garner bonus points will test out at just slightly above average on this subtest. Block Design scores be misleading for examinees who do not v speeded performance.

Matrix Reasoning

Matrix Reasoning is a new subtest found only on WAIS-III. It was added to enhance the assessment of nonverbal reasoning on the adult test. The sub' consists of 26 figural reasoning problems arranged in increasing order of difficulty (Figure 9.3). Finding the correct answer requires the examinee identify a recurring pattern or relationship between figural stimuli drawn along a straight line (simple items) or in a 3 x 3 grid (hard items) in which t last item is missing. Based upon nonverbal reasoning about the patterns and relationships, the examinee must infer the missing stimulus and select from five choices provided at the bottom of the c Matrix Reasoning was designed to be a measure of fluid intelligence, which is the capacity to perform mental operations such as manipulation of abstract symbols. The items tap pattern completion, reasoning by analogy, and serial reasoning. Overall, the subtest is an excellent measure of inductive reasoning based on figural stimuli. Matrix Reasoning is the only untimed performance subtest on the WAIS-III. Interestingly, Donders et al. (2001) report that the Matrix Reasoning subtest is relatively unaffected by moderate and severe traumatic brain injury.

Object Assembly

For each item, the examinee must assemble the pieces of a jigsaw puzzle to form a common object (Figure 9.4). For example, Object Assembly on the WAIS-III consists of five puzzles: a manikin (6 pieces), a profile (7 pieces), an elephant (6 pieces), a house (9 pieces), and a butterfly (7 pieces). The examiner does not identify the items, so the examinee must first discern the identity of each item from its disarranged parts. Success on this subtest requires high levels of perceptual organization; that is, the examinee must grasp a larger pattern or gestalt based on perception of the relationships among the individual parts. Object Assembly is one of the least reliable of theWechsler subtests. For example, on the WAIS-III this subtest has an average split-half reliability of only .70 (Tulsky, Zhu, & Ledbetter,1997). Among the WAIS-III subtests, only Picture Arrangement with a value of .74 approaches the unreliability of Object Assembly. These two subtests stand apart from the other, more reliable, Wechsler subtests. The modest reliability of Object Assembly may reflect, in part, the small number of items as well as the role of chance factors in solving jigsaw puzzles.

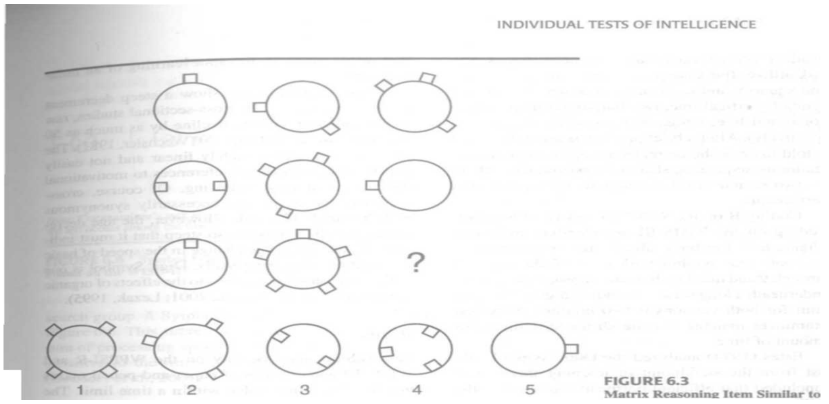

INDIVIDUAL TESTS OF INTELLIGENCE
?
1
2
3
4
5
FIGURE 6.3
Matrix Reasoning Item Similar to

Figure 9.3 Matrix Reasoning Item Similar to Those Found on the WAIS-III

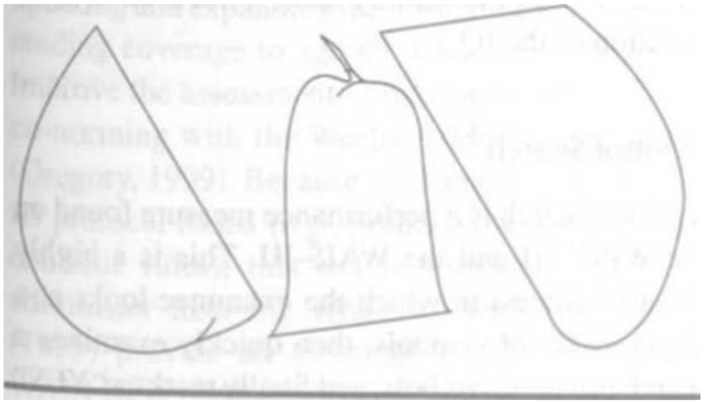


Figure 9.4 Object Assembly Item Similar to
Those Found on the WAIS-III

Coding/Digit Symbol

Although the tasks are nearly identical, this subtest is called Coding on the WISC-III and Digit Symbol-Coding on the WAIS-III. The WISC-III version consists of two separate and distinct parts, one for examinees under age 8 (Coding A) and another for those 8 years of age and over (Coding B). In Coding A, the child must draw the correct symbol inside a series of randomly sequenced shapes. The task utilizes five shapes (star, circle, triangle, cross, and square), and each shape is assigned a unique symbol (vertical line, two horizontal lines, single horizontal line, circle, and two vertical lines, respectively). After a brief practice session, the child is told to draw the correct symbol inside 43 of the randomly sequenced shapes. However, since there is a two-minute time limit, high scores require rapid performance.

Coding B on the WISC-III and Digit Symbol-Coding on the WAIS-III are identical in format (Figure 6.5). For both subtests, the examinee must associate one symbol with each of the digits 0 through 9 and quickly draw the appropriate symbol underneath a long series of random digits. The time limit for both versions is two minutes. Very few examinees manage to code all the stimuli in this amount of time.

Estes (1974) analyzed the Digit Symbol subtest from the standpoint of learning theory and concluded that efficient performance requires the ability to quickly produce distinctive verbal codes to represent each of the symbols in memory. For example, in Figure 6.5, the examinee might code the symbol underneath the number 2 as an "inverted T." Verbal coding mediates quick performance by simplifying a difficult task. Efficient performance also demands immediate learning of the digit symbol pairings so that the examinee need not look from each digit to the reference table to determine the correct response. In this regard, Digit Symbol is unique: It is the only Wechsler subtest that necessitates on-the-spot learning of unfamiliar task.

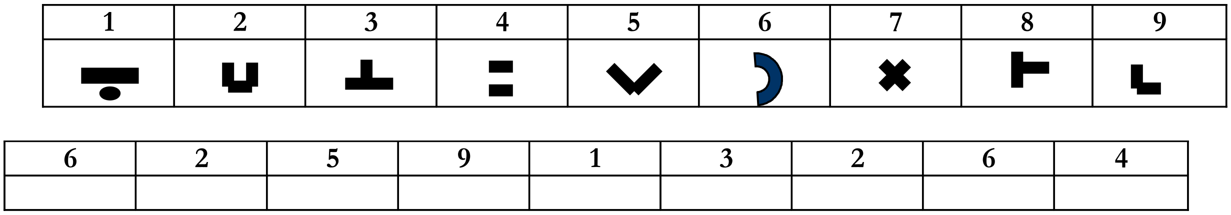


Figure 9.5 Digital Symbol Items Similar to Those Found on the WAIS-III

Digit Symbol scores show a steep decrement with advancing age. In cross-sectional studies, raw scores on Digit Symbol decline by as much as percent from age 20 to age 70 (Wechsler, 1981). The decrement is approximately linear and not easily explained by superficial references to motivational differences or motor slowing. Of course, cross sectional results are not necessarily synonymous with longitudinal trends. However, the age decrement on Digit Symbol is so steep that it must indicate, in part, a real age change in the speed of bas:-information-processing skills. Digit Symbol is of the most sensitive subtests to the effects of organic impairment (Donders et al., 2001; Lezak, 1995).

Mazes

This subtest appears only on the WPPSI-R WISC-III and consists of paper-and-pencil mazes that the child must solve within a time limit. The examinee is told not to lift the pencil and is counseled "try not to enter any blind alleys." Full credit for each maze is given if the child solves it wit the time limit (30 seconds to 150 seconds, depending upon difficulty) without entering any blind al leys. One raw score point is deducted for each blind alley entered.

Mazes taps perceptual-motor skills, motor speed, visual planning, and the ability to inhibit impulsive responding. This subtest is a poor measure of general intelligence, but measures perceptual organization reasonably well. On the WISC-III, Mazes is a supplementary subtest not used in computation of the IQ.

Symbol Search

Symbol Search is a performance measure found on the WISC-III and the WAIS-III. This is a highly speeded subtest in which the examinee looks at a target group of symbols, then quickly examines a search group of symbols, and finally marks a "YES" or "NO" box to indicate whether one or more of the symbols in the target group occurred within the search group. Symbol Search is highly sensitive to the impact of traumatic brain injury (Donders et al., 2001).

STANFORD-BINET INTELLIGENCE SCALES: FIFTH EDITION

With a lineage that goes back to the Binet-Simon scale of 1905, the Stanford-Binet: Fifth Edition (SB5) has the oldest and perhaps the most prestigious pedigree of any individual intelligence test. In Table 9.4, we outline some important milestones in the development of the SB5 and its predecessors. Released in 2003, the SB5 is a very new test (Roid. 2002, 2003). For this reason, evaluation of this instrument is based, in part, upon its resemblance in content and subtests to the SB4, about which a large body of in? dependent research literature has been amassed.

The SB5 Model of Intelligence

In early editions of the Stanford-Binet, the examiner obtained only a composite IQ. Although the pattern of right and wrong answers could be analyzed qualitatively, the earlier Stanford-Binet tests (prior to the fourth edition) did not provide a basis for quantitative analysis of the subcomponents of the entire scale. The fourth and fifth editions corrected this shortcoming.

Table 9.4 Milestones in the Development of the Stanford-Binet and Predecessor

Year	Test/Author	Comment
1905	Binet and Simon	Simple 30-item test
1908	Binet and Simon	Introduced the mental age concept
1911	Binet and Simon	Expanded to include adults
1916	Stanford-Binet Terman and Merrill	Introduced the IQ concept
1937	Stanford-Binet-2 Terman and Merrill	First use of parallel forms (L and M)
1960	Stanford-Binet-3 Terman and Merrill	Modern item-analysis methods used
1972	Stanford-Binet-3 Thorndike	SB-3 restandardized on 2,100 persons
1986	Stanford-Binet-4 Thorndike, Hagen and Sattler	Complete restructuring into 15 subtests
2003	Stanford-Binet-5 Roid	Five factors of intelligence

The organization of the SB5 was guided by the principle that each of five factors of intelligence can be assessed in two distinct domains—nonverbal and verbal. The five factors—derived from modern cognitive theories such as Carroll (1993) and Baddeley (1986)—are fluid reasoning, knowledge, quantitative reasoning, visual-spatial processing, and working memory. When these five factors of intelligence are "crossed" with the two domains (nonverbal and verbal), the result is an instrument with ten subtests (Figure 6.7). Thus, the SB5 provides a number of different perspectives on the cognitive functioning of an examinee: ten subtest scores (mean of 10, SD of 3), three IQ scores (the familiar Full Scale IQ, Verbal IQ, and Nonverbal IQ), as well as five factor scores (Fluid Reasoning, Knowledge, Quantitative Reasoning, Visual-Spatial Processing, and Working Memory). The IQ and factor scores are normed to a mean of 100 and SD of 15.

Routing Procedure and Tailored Testing

The SB5 maintains the historical tradition of this instrument by using a routing procedure to estimate the general cognitive ability of the examinee before proceeding to the remainder of the test. The purpose of the routing procedure is to identify the appropriate starting points for subsequent subtests. The routing items are both nonverbal (object series and matrices) and verbal (vocabulary). These items also provide the Abbreviated IQ, sometimes used for screening purposes. Roid (2002) describes the advantages of using a routing procedure:

> *This tailored approach to assessment provides greater richness of factor measurement within a shorter, efficient test administration. The use of modern item response theory in the design of SB5 allows for greater precision of measurement due to the adaption of the test to the functional level of the examinee in an efficient time frame.*

Thus the purpose of the routing procedure is not just to reduce the number of items administered (and therefore save time), but to do so without loss of measurement precision. This is possible because the SB5 was constructed according to the principles of item response theory (Embretson, 1996). When a test is constructed within the framework of item response theory, item difficulty levels and other parameters are precisely calibrated during the development phase.

Domains

		Nonverbal	Verbal
Factors	Fluid reasoning	Nonverbal fluid reasoning	Verbal fluid reasoning
	Knowledge	Nonverbal knowledge	Verbal knowledge
	Quantitative reasoning	Nonverbal quantitative reasoning	Verbal quantitative reasoning
	Visual-Spatial reasoning	Nonverbal visual-spatial processing	Verbal visual-spatial processing
	Working memory	Nonverbal working memory	Verbal working memory
	Full Scale IQ	Nonverbal IQ	Verbal IQ

Figure 9.6 Structure of the Stanford-Binet: Fifth Edition

Special Features of the SB5

In addition to providing a more familiar partition of intelligence into Full Scale IQ, Verbal IQ. and Nonverbal IQ, the SB5 also features a number of other improvements over its predecessor, the SB4. The test now includes extensive high-end items, designed to assess the highest level of gifted performance. Many of these items are updates from very early editions of the Stanford-Binet, when the instrument was renowned for its very high ceiling. At the other extreme, improved low-end items provide better assessment for very young children (as young as age 2) and adults with mental retardation. In addition, the items and subtests that contribute to the Nonverbal IQ do not require expressive language, which makes this part of the test ideal for assessing individuals with limited English, deafness, or communication disorders. The developers of the SB5 also screened test items for fairness based on religious as well as traditional concerns. Expert panels examined the entire test on fairness issues related to the standard variables (gender, race, ethnicity, and disability) and religious tradition (Christian, Jewish, Muslim, Hindu, and Buddhist backgrounds). This is the first time in the history of intelligence testing that religious tradition has been considered in test development. Finally, the Working Memory factor, consisting of both verbal and nonverbal subtests, shows promise in helping to assess and understand children with attention-deficit/hyperactivity disorder.

Standardization and Psychometric Properties of the SB5

The SB5 is suitable for children age 2 through adults age 85 and older, and the standardization sample consists of 4,800 individuals stratified by gender, ethnic, regional, and educational levels in the United States, based on the year 2000 census. In part because item selection was determined by modern item response theory, the reliability of subtests, indices, and IQ scores is very strong and comparable to other mainstream individual intelligence tests. For example, the Verbal IQ, Nonverbal IQ. and Full Scale IQ each have reliabilities in the .90s, and the individual subtests are in the range of .70 to .85 (Roid, 2002).

As is typical in the release of a new test, the manual for the SB5 (Roid, 2003) reports on numerous affirming correlational studies (e.g., with the Wechsler scales, the SB4. the UNIT) that provide strong support for criterion-related validity. The validity of the test as a measure of general intelligence is also supported by its resemblance to the SB4, about which a large body of research can be cited. For example Lamp and Krohn (2001) studied the longitudinal predictive validity of the SB4 in a sample of 89 Head Start children (39 African American and 50 white) from impoverished backgrounds who ranged in age from about 4 to 6V2. These children were retested several times over an 8-year period on both the SB4 and the Metropolitan Achievement Test. The correlations between the initial SB4 score and the subsequent achievement scores were very strong (mainly in the .50s), and the test was equally good at predicting outcome for African American and white children. In another study (Atkinson, Bevc, Dickens, & Blackwell, 1992), the concurrent validity of the SB4 was tested against the Leiter International Performance Scale and the Vineland Adaptive Behavior Scales in a sample of 24 children with developmental delays. The correlations were very robust (.78 and .70, respectively). These and many other studies strongly support the validity of the SB4 as a measure of general intelligence. As new research is reported on the SB5, it is likely that this recent edition also will prove to be highly valid and even more useful than its predecessor as a measure of intelligence.

In summary, the SB5 is a very promising new test that is especially useful at both ends of the cognitive spectrum—the very young or those with developmental delays, and very gifted persons. Based upon me care with which the instrument was constructed, me test is likely to become a mainstay of individual intelligence testing in a wide variety of settings.

DETROIT TESTS OF LEARNING APTITUDE-4

The Detroit Tests of Learning Aptitude-4 (DTLA-4; Hammill, 1999) is a recent revision of instrument first published in 1935. The test is individually administered and designed for schoolchildren from 6 through 17 years of age. The DTLA-4 consists of 10 subtests that form the basis for computing 16 composites, including general intelligence, optimal level, and 14 ability areas. The subtests are largely within the Binet-Wechsler tradition, although there are a few surprises such as the inclusion of Story Construction, a measure of storytelling ability (Table 9.5).

The General Mental Ability composite is formed by combining standard scores for all 10 subtests in the battery. The Optimal Level composite is based upon the highest 4 standard scores earned by the subject and is thought to represent how well the examinee might perform under optimal circumstances. Each of the remaining 14 composite scores is derived from a combination of several subtests thought to measure a common attribute. For example, subtests that involve knowledge of words and their use are combined to form the Verbal Composite, whereas subtests that do not involve reading, writing, or speech comprise the Nonverbal Composite. Several of the composite scores are designed to represent major constructs within contemporary theories of intelligence. In addition to the General Mental Ability composite and the Optimal Level composite, the remaining 14 DTLA-3 composite scores are as follows:

Table 9.5 Brief Description of the DTLA-4 Subtests

Subtest	Task
Word Opposites	Provide antonyms—word opposites.
Design Sequences	Discriminate and remember nonsensical graphic material.
Sentence Imitation	Repeat orally presented sentences.
Reversed Letters	Short-term visual memory and attention.
Story Construction	Create a logical story from several pictures.
Design Reproduction	Copy designs from memory.
Basic Information	Knowledge of everyday facts and information.
Symbolic Relations	Select from a series of designs the part that was missing from a previous design
Word Sequences	Repeat a series of unrelated words.
Story Sequences	Organize pictorial material into meaningful sequences.

Verbal	Nonverbal	(Linguistic)
Attention-enhanced	Attention-reduced	(Attentional)
Motor-enhanced	Motor-reduced	(Motoric)
Fluid	Crystallized	(Horn & Cattell)
Simultaneous	Successive	(Das)

Associative	Cognitive	(Jensen)
Verbal	Performance	(Wechsler)

The 16 composite scores are based upon the familiar mean of 100 and standard deviation of 15. The 10 subtests are normed for a mean of 10 and standard deviation of 3.

The composites were designed to offer contrasting assessments such that a difference between scores may be of diagnostic significance. For example, an examinee who scored well on Attention-Reduced aptitude but poorly on Attention-Enhanced aptitude (in the Attentional domain) presumably experiences difficulty with immediate recall, short-term memory, or focused concentration.

The DTLA-4 was standardized on 1,350 students whose backgrounds closely matched census data for sex, race, urban/rural residence, family income, educational attainment of parents, and geographic area. The reliability of this instrument is similar to other individual tests of intelligence, with internal consistency coefficients generally exceeding .80 for the subtests and .90 for the composites, and test-retest coefficients for the subtests and the composites in the .80s and .90s. Criterion-related validity is well established through correlational studies with other mainstream instruments such as the WISC-III, K-ABC, and Woodcock-Johnson.

A concern with the DTLA-4 is that the conceptual breakdown into composites is not sufficiently supported by empirical evidence. For example, while it may be true that the Simultaneous composite does measure the simultaneous cognitive processes proposed by Das, Kirby, and Jar-man (1979), there is scant empirical support to buttress this claim. Another problem with this instrument is that there are more composites than there are subtests! Inevitably, the composites win be highly intercorrelated, because each subtest occurs in several composites. In sum, DTI ,A-4 may be a good measure of general intelligence, but the use of composite scores for purposes of psycho-educational planning requires additional empirical study. Smith (2001) and Traub (2001) provide thorough reviews of the DTLA-4.

KAUFMAN BRIEF INTELLIGENCE TEST (K-BIT)

The individual intelligence tests previously discussed in this and the preceding topic are excellent measures of intellectual ability, but they are not without their drawbacks. One problem is the time required to administer them. Testing sessions with the Wechsler scales, Kaufman Assessment Battery for Children, and the Stanford-Binet easily can last one hour, and two hours is not unusual if the examinee is bright and highly verbal. A second disadvantage to these mainstream tests is the amount of training required to administer them. Proper administration of most individual intelligence tests is based upon the assumption that the examiner has an advanced degree in psychology or a related field and has received extensive supervised experience with the instruments in question.

Alan Kaufman responded to the need for a brief, easily administered screening measure of intelligence by developing the Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990; Kaufman & Wang, 1992). The K-BIT consists of a Vocabulary section and a Matrices section. The Vocabulary test contains two parts: Expressive Vocabulary (naming pictures) and Definitions (providing a word based upon a brief phrase and ft partial spelling). The Matrices test requires solving 2x2 and 3x3 analogies using figural stimuli.

The K-BIT is normed for subjects ages 4 to 90 years and can be administered in 15 to 30 minutes. The test yields standard scores with mean of 100 and SD of 15 for Vocabulary, Matrices, and the combination of the two, called the IQ Composite, to spite of the comparability of these scoring dimensions with well-known intelligence tests, the K BIT authors make it clear that their instrument is not intended as a substitute for traditional approaches (e.g.. WPPSI-R. K-ABC. WISC-III or SB: FE). The K-BIT is mainly a screening test useful in signaling the need for more extensive assessment. The brevity of this instrument also makes it a natural choice for research on intelligence.

Reliability findings for the K-BIT are exceptionally strong. Split-half reliability and test-retest coefficients for a variety of samples were in the .90s for Vocabulary, the .80s and .90s for Matrices, and .90s for IQ Composite. The normative sample of 2,022 individuals was within 1 to 3 percentage points of the 1990 U.S. Census figures for sex, geographic region, race or ethnic group, and educational attainment of the parents (for subjects 4 to 19 years of age) or examinees themselves (20 years of age and above).

The K-BIT manual reports highly supportive validity data from 20 correlational studies. These results are similar to a recent concurrent validity study that compared K-BIT results and WAIS-R scores for 200 referrals to a neuropsychological assessment center (Naugle, Chelune, & Tucker, 1993). The patient sample included persons with seizure disorders, head injuries, substance abuse, psychiatric disturbance, stroke, dementia, and other neurological conditions. The heterogeneity of the referral sample guaranteed a wide range of functional ability, a desirable feature in a validation study. Although the K-BIT scores tended to be about 5 points higher than their WAIS-R counterparts, the correlations between these two instruments were extremely high and theory-confirming. Vocabulary IQ (K-BIT) and Verbal IQ (WAIS-R) correlated .83; Matrices IQ (K-BIT) and Performance IQ (WAIS-R) correlated .77; and overall IQs from the two instruments correlated an amazing .88. In a study comparing the K-BIT and the WISC-III scores for 50 referred students, Prewett (1995) also reported strong correlations (r = .78 for overall scores) and also discovered that the K-BIT scores tended to be about 5 points higher than their WISC-III counterparts. In a sample of 65 children with reading disability, Chin, Ledesma, Cirino, and others (2001) also found that the K-BIT overestimated WISC-III IQs by 1.2 to 5.0 points, on average. However, their study also showed that, in individual cases, K-BIT scores can underestimate or overestimate WISC-III scores by as much as 25 points, reaffirming that the K-BIT is not appropriate for placement and diagnostic purposes. Canivez (1995) found comparable scores between the K-BIT and the WISC-III for 137 elementary- and middle school children and also reported very strong correlations between the two tests, especially for overall scores (r = .87). Eisen-stein and Engelhart (1997) found that the K-BIT performed well in estimating IQs in adult neuropsychology referrals, but Donders (1995) recommends caution when using the test with brain-injured children. The reason for caution is that K-BIT scores show a negligible relationship with length of coma; that is, the test is not a good index of neuropsychological status in children. Even so, the K-BIT is an outstanding screening measure of general intelligence for use in research or when time constraints preclude use of a longer measure.

Group Tests of Intelligence

A group intelligence test allows for the quick and efficient testing of dozens or hundreds of examinees at the same time. In this topic we introduce the reader to a sampling of prominent group tests. For better or for worse, the number of group tests currently marketed is simply astonishing scores of them are available. Several dozen entries are reviewed in recent issues of the Mental Measurements Yearbook (Mitchell. 1985: Conoley & Kramer, 1989, 1992) and the Test Critiques series (Keyser & Sweetland. 1984-1988) and new instruments are published every year. Comprehensive coverage of this burgeoning field is simply not feasible. Consequently, we focus here on issues raised by group tests and then review an eclectic assortment of these diverse instruments.

ORIGINS AND CHARACTERISTICS OF GROUP TESTS

ORIGINS OF GROUP TESTS

The first useful group intelligence tests were developed early in the twentieth century in the United States. Nonetheless, the origins of these instructs can be traced to the efforts of nineteenth century European psychologists. The modern group intelligence test owes a debt especially to the completion technique developed in the 1890s by Ebbinghaus (1896). His test consisted of several passages of text with words or parts of words omitted, as in the following brief example:

Little Red Riding Hood

__________there was a sweet young ______________, beloved by every __________who eyes on her. Her __________ mother gave her a little cap of ____________ silk, which she wore ______________ the time. The __________was known as __________ Red Riding Hood.
One ____________ her mother told her, “your __________ is ill and weak. __________ take this cake and wine to her. Do not stray from the ____________ and do not __________ to strangers."

The student's task was to fill in as many blanks as possible (for several selections) in a five-minute time limit. The completion test was commonly administered to an entire class by one person. The task was highly speeded: Only four times in several thousand cases did a student fill in all of the blanks. Ebbinghaus used the total number correct as a basis for comparing individuals as to their intellectual ability (DuBois, 1970).
A few years later, the practical success of the Binet scales inspired psychologists to develop intelligence tests that could be administered simultaneously to large numbers of examinees. We have noted in a previous chapter that the need to quickly test thousands of Army recruits for WWI inspired psychologists in the United States, led by Robert M. Yerkes, to make rapid advances in psychomet-rics and test development.

Parallel developments occurred in school systems where administrators desired an efficient means for testing and placing students. However, fill-in-the-blank and open-ended questions severely limited the efficiency of assessment. Group testing quickly evolved into its modern design: the multiple-choice format.

Differences between Group and Individual Tests

Group tests differ from individual tests in five ways:

- Multiple-choice versus open-ended format
- Objective machine scoring versus examiner scoring
- Group versus individualized administration
- Applications in screening versus remedial planning
- Huge versus merely large standardization samples

We discuss each of these points in turn.

The most obvious difference is that group tests generally employ a multiple-choice format. Although early group tests did use open-ended questions, this feature was quickly dropped because of the excessive amounts of time required for scoring. As a result of the multiple-choice format, group tests can be quickly and objectively scored by an optical scanning device hooked up to a computer. Computer scoring eliminates examiner errors and halo effects that may occur in the scoring of individual tests. In addition, psychometricians gain nearly instant access to item analyses and test data banks, so computer scoring promotes the quick development and revision of group tests.

Group tests also differ from individual tests in the mode of administration. In a group test, the examiner plays a minimal role that is restricted largely to reading instructions and enforcing time limits. There is negligible opportunity for one-on-one interaction between the test giver and the test taker. For most examinees, this will not matter, but for a few—the shy, the confused, the unmotivated—the absence of examiner rapport can have disastrous results.

Traditional intelligence tests excel as aids in the diagnosis and remediation of individual learning difficulties, whereas group intelligence tests more commonly used for mass screening in the furtherance of institutional decision making. Thus group tests might be used in school systems to "flag" children in need of academic remediation or enrichment; in industrial settings to identify good candidates for specific jobs; or in military settings to help cull out mentally impaired recruits.

Group tests are generally standardized on ultra-large samples—hundreds of thousands of subjects instead of just the few thousand carefully selected cases used with individual tests. Of course, the suitability of a standardization sample must never be taken for granted. Whether using huge standardization samples for group testing, or smaller standardization samples for individual testing, it is still important to determine the degree to which the sample is representative of the population at large.

SHIPLEY INSTITUTE OF LIVING SCALE (SILS)

The Shipley Institute of Living Scale (SILS) is also known as the Shipley-Hartford because of its inception in Hartford, Connecticut, decades ago (Shipley, 1940, 1983). The SILS was originally proposed as an index of intellectual deterioration, in an attempt to gauge the effects of dementia, brain damage, and other organic conditions. However, the test has been used primarily as a short screening test of intelligence, particularly within the mental health system of the Veterans Administration.

Background and Description

The SILS consists of two subtests, vocabulary and abstractions. The original intention of the test was to detect organic intellectual deterioration by contrasting performance on the vocabulary and abstractions sections. Vocabulary was thought to be relatively unaffected by organic deterioration, whereas it was believed that abstraction ability would show significant decline. A large discrepancy favoring vocabulary over abstractions therefore would appear to signify the presence of organic impairment. However, numerous studies and reviews concluded that the SILS performs poorly as an index of brain damage (e.g., Yates, 1954; Johnson, 1987), and the instrument is seldom used for t purpose.

The SILS consists of 40 multiple-choice vocabulary items and 20 abstract-thinking items-Each item is scored right or wrong. The abstract items count double, so the maximum score on each "^lf of the test is 40 points. A composite score is also reported. The test is self-administered with a *q minute limit for each of the two sections. Some users favor an untimed use of the test, and separate norms have been developed for this approach (Heinemann, Harper. Friedman. & Whitney, 1985). Few persons require more than 10 minutes per section; most examiners consider the SILS to be entirely a power measure. A microcomputer version of the test is also available. The computer administers and scores the test and produces a narrative report and graphic depiction of scores.

The examinee's task on the vocabulary section is to select the synonym of a word from four alternatives. The 40 items resemble the following:

• SHIP	house	tree	fork	boat
• INANE	fat	timely	silly	dry

The vocabulary score is the number correct plus one point for every four items omitted. Adding points for items omitted provides a correction for the refusal to guess. As a result of this correction factor, the minimum score is about 10 out of the 40 points.

The intention of the abstractions items is that they should require the examinee to infer a principle common to a given series of components and then to demonstrate this understanding of the principle by finishing the series. Each item is a series of letters or numbers followed by blanks to indicate the number of characters in the answer. The 20 items resemble the following:

- A B D G K ___________
- Bog hob mars tram 268 ___________ ___________ ___________
- 135 341 52 12 __________

The examinee must complete each series and place me appropriate answer in the blanks. (Answers to the preceding items are P, 962. and 3). Of course, 0 derive the correct answer the examinee must infer the rule that governs the progression of stim-1 »n each item and then use that rule to determine e continuation. (In item 1 the distance between letters increases arithmetically: in item 2 the pairs e mirror images of each other, except for last and first letters which increment by one ___ g to h, s to t; in item 3, each group of numbers sums to one less than the previous group ________9,8,7,….).

A MULTILEVEL BATTERY: THE COGNITIVE ABILITIES TEST (CogAT)

One important function of psychological testing is to assess students' abilities that are prerequisite to traditional classroom-based learning. In designing tests for this purpose, the psychometrician must contend with the obvious and nettlesome problem that school-aged children differ hugely in their intellectual abilities. For example, a test appropriate for a sixth grader will be much too easy for a tenth grader, yet impossibly difficult for a third grader.

The answer to this dilemma is a multilevel battery, a series of overlapping tests. In a multilevel battery, each group test is designed for a specific age or grade level, but adjacent tests possess some common content. Because of the overlapping content with adjacent age or grade levels, each test possesses a suitably low floor and high ceiling for proper assessment of students at both extremes of ability. In addition, multilevel batteries usually provide a much desired continuity in the abilities measured. Furthermore, multilevel batteries generally employ highly comparable normative samples at the successive levels. For all of these reasons, multilevel batteries are considered ideal for gauging student readiness for school learning. Virtually every school system in the United States uses at least one nationally normed multilevel battery.

The Cognitive Abilities Test (CogAT) is one of the best school-based test batteries in current use (Lohman & Hagen, 2001). A recent revision of the test is the CogAT Multilevel Edition, Form 6, released in 2001. We discuss this instrument in sonne detail and then provide a brief summary of peting tests.

Background and Description

The CogAT evolved from the Lorge-Thorndike In-Higence Tests, one of the first group tests of intelligence intended for widespread use within school systems. The CogAT is primarily a measure f scholastic ability, but also incorporates a nonverbal reasoning battery with items that bear no direct relation to formal school instruction. The two primary batteries, suitable for students in kindergarten through third grade, are briefly discussed at the end of this section. Here we review the multilevel edition intended for students in third through twelfth grade.

The nine subtests of the multilevel CogAT are grouped into three batteries as follows:

Verbal Battery	Quantitative Battery	Nonverbal Battery

Verbal classification
Sentence completion
Verbal analogies

Quantitative relations
Number series
Equation building

Figure classification
Figure analogies
Figure analysis

For each CogAT subtest, items are ordered by difficulty level in a single test booklet. However, entry and exit points differ for each of eight overlapping levels (A through H). In this manner, grade appropriate items are provided for all examinees. All subtests except one use a multiple-choice format. The exception is Figure Analysis, in which the examinee responds yes or no to a series of alternatives.

The subtests are strictly timed, with limits that vary from eight to twelve minutes. Each of the three batteries can be administered in less than an hour. However, the manual recommends three successive testing days for younger children. For older children, two batteries should be administered the first day, with a single testing period the next.

Many subtests of the CogAT bear a striking resemblance to portions of the Stanford-Binet: Fourth Edition. For example, both tests include paper-folding items. Common parentage is the explanation: Both tests were developed by Elizabeth Hagen; both tests were published by Revised Publishing Company. We see once again the hybrid character of modern intelligence tests, in which new tests incorporate the best features of their predecessors.

Raw scores for each battery can be transformed into an age-based normalized standard score with mean of 100 and standard deviation of 15. In addition, percentile ranks and stanines for age groups and grade level are also available. Interpolation was used to determine fall, winter, and spring grade level norms.

CULTURE FAIR INTELLIGENCE TEST (CFIT)

The Culture Fair Intelligence Test (Cattell 1940, IPAT, 1973) is a nonverbal measure of fluid intelligence first conceived in the 1920s by the prom-measurement psychologist Raymond B. CatteJ The goal of the CFIT is to measure fluid intelligence—analytical and reasoning ability in abstract and novel situations—in a manner that is as "free" of cultural bias as possible. This test was originally called the Culture Free Intelligence Test. The name was changed when it became evident that cultural influences cannot be completely extirpated from tests of intelligence.

Background and Description

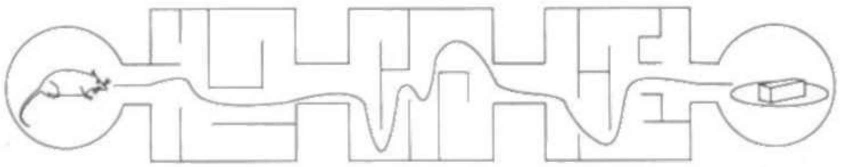
The CFIT has undergone several revisions, emerging in its current form in 1961. The test consists of three versions: Scale 1 is for use with mentally defective adults and children ages tour to eight: Scale 2 is for adults in the average range of intelligence and children ages eight to thirteen; Scale 3 is for high ability adults and for high school and college students. Scale 1 involves considerable interaction between tester and examinee —four of the subtests must be administered individually. Thus, in some respects Scale 1 is more of an individual intelligence test than a group test. We discuss only Scales 2 and 3 here, because they are truly group tests of intelligence. These two tests differ mainly in difficulty level.

Two equivalent forms, called Form A and Form B, are available for each scale. The test developers recommend administering both forms to each subject to obtain what is called the full test. Each form by itself is referred to as a short test. In spite of the recommendation to use both forms as a combined test, it is very common for CFIT users to rely upon a single, brief form for purposes of screening.

Each form consists of four subtests: Series, Classification, Matrices and Conditions. Sample items are shown in figure 6.9. Of course each subtest is preceded by several practice items. The entire test is neatly packaged in an eight page booklet.

The CFIT is a highly speeded test. Each form f Scales 2 and 3 takes about 30 minutes to administer but only 12.5 minutes is devoted to actual test taking. Results can therefore be misleading for persons who place no premium on speed of performance in problem solving. Fortunately, Scale 2 can be used as an untimed power test. However, the norms for this manner of administration are limited (IPAT, 1973).

Mazes



Classification

Pick out the two odd items in each row of figures.

Technical Features

Standardization samples for Scales 2 and 3 were respectably large, but not described in sufficient detail to determine the extent to which they mirror the general population. The standardization samples were characterized as follows:

> *The standardization group for Scale 2 consists of 4,328 males and females sampled from varied regions of the United States and Britain. Scale 3 norms are based on 3,140 cases, consisting of American high school students equally divided among freshmen1: sophomores, juniors, and seniors, and young adults in a stratified job sample. (IPAT, 1973)*

Raw scores are converted to normalized standard score IQs with mean of 100 and standard deviation of 16.

Test-retest, alternate-forms, and internal consistency reliabilities are generally in the .70s for individual forms of Scales 2 and 3. The reliabilities of the full test are higher, generally in the mid .80s. These results are based on dozens of studies with thousands of subjects and indicate a respectable degree of reliability for such a short instrument (IPAT, 1973).

The validity of the CFIT as a measure of general intelligence is established beyond any reasonable skepticism. CFIT scores correlate in the mid-.80s with the general factor of intelligence and show consistently robust relationships—largely in the .70s and .80s—with other mainstream measures of intelligence (WAIS, WISC, Raven Progressive Matrices, Stanford-Binet, Otis, and General Aptitude Test Battery; see IPAT, 1973, p. 11). There is no doubt that the CFIT is a well-designed, useful, and valid test of intelligence.

But is the CFIT a culture-fair test, as its title proclaims? One professed goal of this instrument was to "minimize irrelevant influences of cultural learning and social climate" and thereby produce a "cleaner separation of natural ability from specific learning" (IPAT, 1973). Unfortunately, the available evidence indicates that the CFIT is no more successful than traditional measures in the pursuit of a culturally fair method for measuring intelligence (Koch, 1984). For example, Willard (1968) f0Un that 83 culturally disadvantaged African American children scored about the same on the Stanford! Binet (M = 68.1) as on the CFIT (M = 70.0). Mo: over, 14 of the children hit the CFIT "floor received the lowest possible CFIT IQ score of 57 whereas Stanford-Binet IQs scores were dispersed in a pattern more like a bell-shaped curve.

Comment on the CFIT

The CFIT is an excellent brief, nonverbal measure of general intelligence. Even when Form A and Form B are both used to obtain what is referred to as the full test, the CFIT can be administered to large groups in less than an hour. An important caution to test users is that the laudable goal of producing a culture-fair test has not been accomplished by the CFIT. Moreover, the goal itself may be chimerical:

Cultures differ with respect to the importance they place on competition with peers in performing tasks or solving problems, on speed or quality of performance, and on a variety of other test-related behaviors. Some cultures emphasize concrete rather than abstract problem solving, often to the extent that a problem has no meaning except in a concrete setting. The very notion of taking some artificially contrived test is nonsensical in such situations. (Koch, 1984)

It is doubtful that a truly culture-fair test is eve possible. In future editions, the CFIT developers would be well advised to rename their test so unsophisticated users do not invest this instrument with imaginary properties.

Even though the CFIT is a worthy test, it is badly in need of revision and renorming. The test is rather old-fashioned in appearance. Some of the test item drawings are so small that only persons with perfect vision can infer the figural relations depict in the item components. Previous standardization samples have been poorly specified and would appear to be convenience samples rather than carefully selected stratified representations of the population at large.

RAVEN'S PROGRESSIVE MATRICES (RPM)

First introduced in 1938, Raven's Progressive Matrices (RPM) is a nonverbal test of inductive reasoning based on figural stimuli (Raven, Court, & Raven, 1986, 1992). This test has been very popular in basic research and is also used in some institutional settings for purposes of intellectual screening.

Background and Description

RPM was originally designed as a measure of Spearman's g factor (Raven, 1938). For this reason, Raven chose a special format for the test that presumably required the exercise of g. The reader is reminded that Spearman defined g as the "education of correlates." The term education refers to the process of figuring out relationships based on the perceived fundamental similarities between stimuli. In particular, to correctly answer items on the RPM, examinees must identify a recurring pattern or relationship between figural stimuli organized in a 3 x 3 matrix. The items are arranged in order of increasing difficulty, hence the reference to progressive matrices.

Raven's test is actually a series of three different instruments. Much of the confusion about validity, factorial structure, and the like stems from the unexamined assumption that all three forms should produce equivalent findings. The reader is encouraged to abandon this unwarranted hypothesis. Even though the three forms of the RPM resemble one another, there may be subtle differences in the problem-solving strategies required by each.

The Colored Progressive Matrices is a 36-item test designed for children from 5 to 11 years of age. Raven incorporated colors into this version of the test to help hold the attention of the young children.

The Standard Progressive Matrices is normed for examinees from 6 years and up, although most of the items are so difficult that the test is best suited for adults. This test consists of 60 items grouped into 5 sets of 12 progressions. The Advanced Progressive Matrices is similar to the Standard version, but has a higher ceiling. The Advanced version consists of 12 problems in Set I and 36 problems in Set II. This form is especially suitable for persons of superior intellect.

Technical Features

Large sample U.S. norms for the Colored and Standard Progressive Matrices are reported in Raven and Summers (1986). Separate norms for Mexican American and African American children are included. Although there was no attempt to use a stratified random-sampling procedure, the selection of school districts was so widely varied that the American norms for children appear to be reasonably sound. Sattler (1988) summarizes the relevant norms for all versions of the RPM. Recently, Raven, Court, and Raven (1992) produced new norms for the Standard Progressive Matrices, but Gudjonsson (1995) has raised a concern that these data are compromised because the testing was not monitored.

For the Colored Progressive Matrices, split-half reliabilities in the range of .65 to .94 are reported, with younger children producing lower values (Raven, Court, & Raven, 1986). For the Standard Progressive Matrices, a typical split-half reliability is .86, although lower values are found with younger subjects (Raven, Court, & Raven, 1983). Test-retest reliabilities for all three forms vary considerably from one sample to the next (Burke, 1958; Raven, 1965; Raven et al., 1986). For normal adults in their late teens or older, reliability coefficients of .80 to .93 are typical. However, for preteen children, reliability coefficients as low as .71 are reported. Thus, for younger subjects, RPM may not possess sufficient reliability to warrant its use for individual decision making.

Factor-analytic studies of the RPM provide little, if any, support for the original intention of the test to measure a unitary construct (Spearman's g truly comparable alternate form of the 60 item Stanford Progressive Matrices. For each of the original 60 items, they developed a similar item that was comparable in terms of difficulty level and underlying cognitive strategy required for solution. An alternate forms reliability analysis on a diverse group of 449 children who took both tests in counterbalanced order revealed a reliability coefficient of .90, which is on a par with immediate test retest data In this same sample, the distribution of scores showed no differences for standard deviation, skewness, and rank order of item difficulties. The mean number correct was 36.1 on the SPM and 35.5 on the new test. In sum the two versions of the test are nearly identical in overall psychometric characteristics and also in difficulty level. The new test promises to serve an important role in research studies that require retesting.

Comment on the RPM

Even though the RPM has not lived up to its original intentions of measuring Spearman's g factor, the test is nonetheless a useful index of nonverbal, figural reasoning. The recent updating of norms was a much-welcomed development for this well-known test, in that many American users were leary of the outdated and limited British norms. Nonetheless, adult norms for the Standard and Advanced Progressive Matrices are still quite limited.

The RPM is particularly valuable for the supplemental testing of children and adults with hearing, language, or physical disabilities. Often, these examinees are difficult to assess with traditional measures that require auditory attention, verbal expression, or physical manipulation. In contrast, the RPM can be explained through pantomime, if necessary. Moreover, the only output required of the examinee is a pencil mark or gesture denoting the chosen alternative. For these reasons, the RPM is ideally suited for testing persons with limited command of the English language. In fact, the RPM is about as culturally reduced as possible: the test protocol does not contain a single word in any language. Mills and Tiissot found that the Advanced Progressive Matrices identified a higher proportion of minority children as gifted than did a more traditional measure of academic aptitude (the School and College Ability Test).

A final note of caution: Some very bright and high-functioning persons perform abysmally on the RPM. Gregory and Gernert (1990) tested nearly 100 university faculty members with a variant of the RPM. One participant, an accomplished researcher who had risen to a vice presidential level, hadn't the slightest clue how to solve the RPM problems and scored at a chance level. Some persons of above-average intelligence simply do not perform well on figural-reasoning tasks. Examiners would be well advised to question the validity of a low score obtained by an otherwise accomplished individual.

PERSPECTIVE ON CULTURE-FAIR TESTS

Cattell's Culture-Fair Intelligence Test (CFIT) and Raven's Progressive Matrices (RPM) are often cited as examples of culture-fair tests, a concept with a long and confused history. We will attempt to clarify terms and issues here.

The first point to make is that intelligence tests are merely samples of what people know and can do. We must not reify intelligence and overvalue intelligence tests. Tests are never samples of innate intelligence or culture-free knowledge. All knowledge is based in culture and acquired over time. As Scarr (1994) notes, there is no such thing as a culture-free test.

But what about a culture-fair test, one that poses problems that are equally familiar (or unfamiliar) to all cultures? This would appear to be a more realistic possibility than a culture-free test, but even here the skeptic can raise objections. Consider the question of what a test means, which differs from culture to culture. In theory, a test of matrices would appear to be equally fair to most cultures. But in practice, issues of equity arise. Persons reared in Western cultures are trained in linear, convergent thinking. We know that the purpose of a test is to find the single, best answer and to do so quickly. We examine the 3 x 3 matrix from left to right and top to bottom, looking for the logical principles invoked in the succession of forms. Can we assume that persons reared in Nepal or New Guinea or even the remote, rural stretches of Idaho will do the same? The test may mean something different to them. Perhaps they will approach it as a measure of aesthetic progression rather than logical succession. Perhaps they will regard it as so much silliness not worthy of intense intellectual effort. To assume that a test is equally fair to all cultural groups merely because^ stimuli are equally familiar (or unfamiliar) is appropriate. We can talk about degrees of cult' fairness (or unfairness), but the notion that any is absolutely culture-fair surely is mistaken.

Individual Tests:

"A test of intelligence, personality or any kind of psychological attribute designed to administer to one respondent at a time."

(By ANDREW . M. COLMAN)

Advantages of Individual Tests:

- Individual tests can provide a wealth of information about a subject beyond a test score. Because, in these tests the methods of administration are as identical as possible, the situation in which the subjects take an individual test is typically the same. Therefore, the differences observed in attitudes most likely reflect differences in the individual taking the test.

- After examiners have gained experience with an individual test and know how to use it properly, they can observe different reaction from individual placed in the same situation.
- Examiners have an opportunity to observe behavior in a standard situation that can be helpful in understanding the unique behavior of a person and interpreting the meaning of a test score.
- Examiner flexibility can elicit maximum performance if permitted by standardization.

Disadvantages of Individual Test:

- Practical difficulty encountered with separate tests is that the less experienced careful examiner may make timings wrong. Such errors are more likely to occur and are more serious with several short time limits than with a single long time limit for the whole test.
- It consumes a lot of time.
- It will not elicit the same response if conducted repeatedly.
- The results of the individual test depend on the mood of the examiner.
- From the beginning up to the end, the examiner's mood variations might effect the result.

Examples:

- ➢ Standford-Binet intelligence test
- ➢ The Wechsler Scales
- ➢ The Kaufman Scales

Group Tests:

"A test of intelligence, personality or any other land of psychological attribute that is present in multiple choice format and" can be administered to groups of respondents, simultaneously.

[BY ANDRE W.M. COLMAN]

Advantages of Group Tests:

- Group test is designed primarily as instruments for mass testing
- Group tests can be administered simultaneously to as many persons as can be fitted comfortably into the available space and reached through a microphone.
- In group test printed items are utilized and simple responses that can be recorded on a test booklet
- The need for one to one relationship between the examiner and the examinee was eliminated.
- Group tests facilitate mass testing by greatly simplifying the examiner's role.
- Scoring is typically more objective in; group testing.
- Group tests can be stored on computers.
- Group tests provide better-established norms than do individual test.
- Many subjects are tested at a time
-

  Subjects record own responses
- Scoring is straight forward and objective
- Group test are cost efficient

- Require less examiner skill and training

- Have more objective and more reliable scoring procedures

- It has a very broad application

Disadvantages of Group Tests:

- In group test have less opportunity to establish a rapport, obtain concentration and maintain the interest of the examinees

- Any temporary condition of the examinee such as illness etc may effect the performance of the examinee

- It lacks flexibility

Examples

- Alpha and beta army test
- Multilevel batteries

Lesson 10

TEST BIAS AND TESTING SPECIAL POPULATIONS

The individual and group intelligence tests reviewed in previous chapters are suitable for persons with normal or near-normal capacities in speech, hearing, vision, movement, and general intellectual ability. However, not every examinee falls within the ordinary spectrum of physical and mental abilities. By reason of youthful age, physical disability, diminished intellect, or language disadvantage, a large proportion of the population falls outside the reach of traditional tests and procedures. According to the U.S. Census Bureau, about 25 million Americans (one in ten) have a severe disability that prevents them from performing one or more activities or roles (www.census.gov, 1998). This estimate does not include persons living in institutions. In these special cases, novel tests are needed for valid assessment. In Topic 7A, Testing Special Populations, we discuss instruments designed for exceptional and difficult consultations, such as persons with sensory/motor impairment, recent immigrants from non-English-speaking countries, and individuals with significant intellectual deficiencies. In Topic 7B, Test Bias and Other Controversies, we continue a circumspect theme by raising a number of concerns about the use and meaning of intelligence test scores.

ORIGINS OF TESTS FOR SPECIAL POPULATIONS

Beginning in the 1950s, a renewed commitment to the needs and rights of physically and mentally disabled persons arose in the United States (Maloney & Ward, 1979; Patton, Payne, & Beirne-Smith, 1986). Societal attitudes toward those with special needs shifted from outright disdain to a more supportive stance that favored new programs and initiatives on behalf of the disabled. Progress has been slow, but we are no longer surprised to see bathroom facilities with wheelchair access for persons with physical disability, large-print books for persons with visual impairments, or closed-captioned television programs for persons with hearing disabilities. Furthermore, the special needs of citizens with mental retardation are increasingly served by small community care facilities instead of massive, impersonal institutions.

In the early 1970s, the renewed concern for the needs of disabled persons was translated into federal legislation. In 1973, Public Law 93-112 was passed, serving as a "Bill of Rights" for disabled individuals. This legislation outlawed discrimination on the basis of disability. Two years later, the landmark Education for All Handicapped Children Act (Public Law 94-142) was enacted. This legislation mandated that disabled schoolchildren receive appropriate assessment and educational opportunities. In particular, psychologists were directed to assess children in all areas of possible disability—mental, behavioral, and physical—and to use instruments validated for those express purposes.

In this topic, we examine tests that can be used for the assessment of persons with sensory, motor, or mental disabilities. However, before discussing specific tests, we review certain distinctions between the types of tests that are available for exceptional assessments. The reader also will appreciate a brief summary of the legal mandates that have shaped assessment practices with disabled individuals.

Approaches to Assessment of Special Populations

Special tests were first devised in the early 1900s to test non-English-speaking immigrants, people who are deaf, and persons with speech defects (DuBois, 1970). These early special instruments were largely performance or nonlanguage tests that could be administered by pantomime. The examinee manipulated objects or used paper and pencil to complete easy-to-understand tasks such as tracing a path through a maze?

Special instruments also have been devised for nonreading examinees who possess some ability to understand spoken English. These nonreading tests are intended for young children and other illiterate persons who nonetheless can comprehend and follow oral instructions. Many nonreading tests involve the manipulation of objects. However, a nonreading test also can assess language comprehension skills by using a picture vocabulary format: The examiner says a word and the examinee points to the one picture from an array of pictures that depicts the word. Several picture vocabulary tests are discussed subsequently.

A motor-reduced test requires the barest minimum of motor output for a response. In a motor-reduced test, the examinee merely points or gestures to the correct answer from among several alternatives. For example, an examinee with cerebral palsy might respond to picture vocabulary items by placing a hand over the chosen alternative. Some non-reading tests—particularly those that use a picture vocabulary format—are also motor-reduced tests.

Finally, we should mention that several important assessment devices are not really tests at all. A developmental schedule is a standardized device for observing and evaluating the behavioral development of infants and young children. These instruments usually inquire into major developmental milestones such as sitting alone, standing unaided, and so forth. It is characteristic of such tools that the "examinee" doesn't take a test per se or, for that matter, do anything out of the ordinary. A developmental schedule is really just a structured form of observation. Likewise, a behavior scale is an instrument for determining the profile of behavioral skills (and perhaps excesses) exhibited by a child or adult with mental retardation. Behavior scales are usually filled out by a knowledgeable adult (parent, teacher, or psychologist).

THE LEGAL MANDATE FOR ASSESSING PERSONS WITH DISABILITIES

Many practices in the assessment of persons with disabilities are the direct result of legislation and court cases. As background to the discussion of specific tests and procedures, we offer a quick review of public laws relevant to the assessment of persons with disabilities. The coverage is purposefully brief. Readers can find lengthier discussions in Bruyere and O'Keeffe (1994), Salvia and Ys-seldyke (2001), and Stefan (2001).

Public Law 94-142

In 1975, the U.S. Congress passed a compulsory special education law, Public Law 94-142, known as the Education for All Handicapped Children Act.[1] According to Ballard and Zettel (1977) this law was designed to meet four major goals:

- To ensure that special education services are available to children who need them
- To guarantee that decisions about services to disabled students are fair and appropriate
- To establish specific management and auditing requirements for special education
- To provide federal funds to help the states educate disabled students

Many practices in the assessment of disabled persons stem directly from the provisions of Public Law 94-142. For example, the law specifies that each disabled student must receive an individualized education plan (IEP) based on a comprehensive assessment by a multidisciplinary team. The IEP must outline long-term and short-term objectives and specify plans for achieving them. In addition, the IEP must indicate how progress toward these objectives will be evaluated. The parents are intimately involved in this process and must approve the particulars of the IEP. Pertinent to testing practices, PL 94-142 includes a number of provisions designed to ensure that assessment procedures and activities are fair, equitable, and nondiscriminatory. Salvia and Ysseldyke (1988) summarize these provisions as follows:

1. Tests are to be selected and administered in such a way as to be racially and culturally nondiscriminatory.
2. To the extent feasible, students are to be assessed in their native language or primary mode of communication.
3. Tests must have been validated for the specific purpose for which they are used.
4. Tests must be administered by trained personnel in conformance with the instructions provided by the test producer.

[1]

5. Tests used with students must include those designed to provide information about specific educational needs, and not just a general intelligence quotient.
6. Decisions about students are to be based on more than performance on a single test.
7. Evaluations are to be made by a multidisciplinary team that includes at least one teacher or other specialist with knowledge in the area of suspected disability.
8. Children must be assessed in all areas related to a specific disability, including—when appropriate—health, vision, hearing, social and emotional status, general intelligence, academic performance, communicative skills, and motor skills.

PL 94-142 also contains a provision that disabled students should be placed in the least restrictive environment—one that allows the maximum possible opportunity to interact with nonimpaired students. Separate schooling is to occur only when the nature or the severity of the disability is such that instructional goals cannot be achieved in the regular classroom. Finally, the law contains a due process clause that guarantees an impartial hearing to resolve conflicts between the parents of disabled children and the school system.

In general, the provisions of PL 94-142 have provided strong impetus to the development of specialized tests that are designed, normed, and validated for children with specific disabilities. For example, in the assessment of a child with visual impairment, the provisions of PL 94-142 virtually dictate that the examiner must use a well-normed test devised just for this population rather than relying upon traditional instruments.

Public Law 99-457

In 1986, Congress passed several amendments to the Education for All Handicapped Children Act, expanding the provisions of PL 94-142 to include disabled preschool children. Public Law 99-457 requires states to provide free appropriate public education to disabled children ages 3 through 5. The law also mandates financial grants to states that offer interdisciplinary educational services to disabled infants, toddlers, and their families, thus establishing a huge incentive for states to serve children with disabilities from birth through age 2. Public Law 99-457 also provides a major impetus to the development and validation of infant tests and developmental schedules. After all, the early and accurate identification of at-risk children would appear to be the crucial first step in effective interdisciplinary intervention.

Americans with Disabilities Act

The 1990 Americans with Disabilities Act (ADA) forbids discrimination against qualified individuals with disabilities in both the public sector (e.g., government agencies and entities receiving federal grants) and the private sector (e.g., corporations and other for-profit employers). Under the ADA, disability is defined as a physical or mental impairment that substantially limits one or more of the major life activities (Parry, 1997). Examples of ADA-recognized disabilities include sensory and physical impairments (e.g., blindness, paralysis), many mental illnesses (e.g., major depression, schizophrenia), learning disabilities, and attention-deficit/hyperactivity disorder.

Under the ADA, the process of qualifying an individual for work or educational accommodations requires current, detailed, and professional documentation. For example, a graduate student who was seeking a special arrangement for taking tests (such as a quiet room) because of attentional problems might need to submit a comprehensive endorsement from a licensed psychologist, detailing the history, current functioning, clinical diagnosis of attention-deficit/hyperactivity disorder, and necessity for accommodations (Gordon & Keiser, 1998). In other words, the ADA is a civil rights act, not a program of entitlement:

> *The ADA does not guarantee equal outcomes, establish quotas, or require preferences favoring individuals with disabilities. Rather, the ADA is intended to ensure access to equal employment opportunities based on merit. The ADA is designed to "level the playing field" by removing the barriers that prevent qualified individuals with disabilities from having access to the same employment opportunities that are available to individuals without disabilities. (Klimoski & Palmer, 1994, p. 45)*

In sum, the purpose is to ensure that individuals who are otherwise qualified for jobs or educational programs are not denied access or put at improper disadvantage simply because of a disability.

In regard to psychological testing, an important provision of the ADA is that agencies and institutions must make reasonable testing accommodations for persons with disabilities. With appropriate documentation (discussed earlier), the relevant accommodations might include any of the following:

- Assistance in completing answer sheets
- Audiotape or oral presentation of written tests
- Special seating for tests
- Large-print examinations
- Retaking exams
- Dictating rather than writing test answers
- Printed version of verbal instructions
- Extended time limit

In general, changes in the testing medium (e.g., from written to oral) are consistent with the intention of ADA, if such a change is needed to accommodate a disability. For example, an appropriate accommodation in the testing medium would be the audiotaped presentation of test items for persons who are visually impaired. On the other hand, changing a test from a printed version into a sign language version for persons with hearing impairment would be considered translation into another language, not a simple change of medium.

In most testing accommodations mandated by the ADA, it is necessary to change the time limits, usually by providing extra time. This raises problems of test interpretation, especially when a strict time limit is essential to the validity of a test. For example, Willingham, Ragosta, Bennett, and others (1988) found that extended time limits on the SAT significantly reduced the validity of the test as a predictor of first-year college grades. This was especially true for examinees with learning disabilities, whose first-year grades were subsequently overpredicted by their SAT scores. Thus, although it seems fair to provide extra time on a test when the testing medium has been changed (e.g., audiotaped questions replacing the printed versions), from a psychometric standpoint, the challenge is to determine how much extra time should be provided so that the modified test is comparable to the original version. Nester (1994) and Phillips (1994) provide thoughtful perspectives on the range of reasonable accommodations required by the ADA.

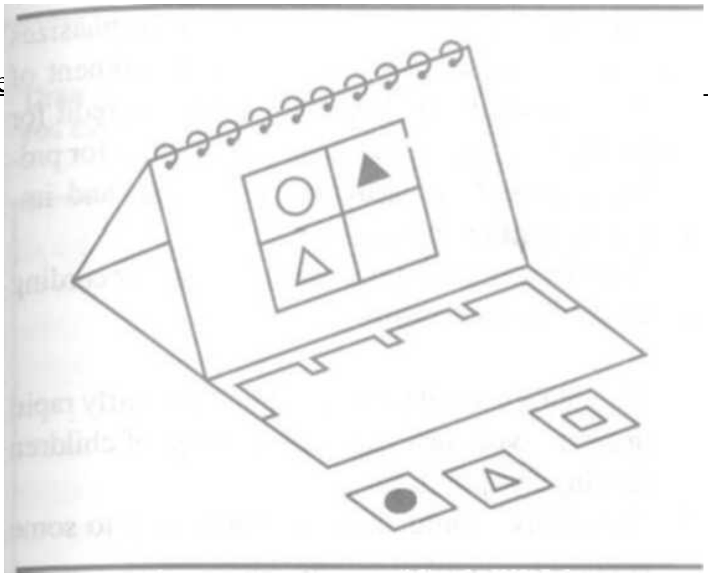
Now that we have summarized the legal background to the assessment of persons with special needs, we turn to a review of typical instruments used for the testing of individuals with disabilities. We organize the review around the following topics: nonlanguage tests, nonreading and motor-reduced tests, tests for persons with visual impairment, and the assessment of adaptive behavior in those with mental retardation.

NONLANGUAGE TESTS

As the reader will recall, nonlanguage tests require little or no written or spoken language from examiner or examinee. Thus, they are particularly suited for assessment of non-English-speaking

Figure 10.1 A Characteristic Item from the Leiter International Performance Scale-Revised

persons, referrals with speech impairments, and examinees with weak language skills. These instruments can also be used as supplementary tests for examinees who have no disabilities.



Leiter International Performance Scale-Revised

The Leiter International Performance Scale-Revised (LIPS-R, Roid & Miller. 1997) is a recent revision of a classic and highly praised test of nonverbal intelligence and cognitive abilities (Leiter, 1948, 1979). Leiter devised an experimental edition of the test in 1929 to assess the intelligence of those with hearing or speech impairment, those who were bilingual or non-English-speaking examinees. The scale was field-tested with several ethnic groups in Hawaii, including children of Japanese and Chinese descent. The first edition was based upon test results for American children, high-school students, and WWII Army recruits. Although highly praised and widely used after its initial release, this test received strong criticism in recent years because of poor illustrations and outdated norms. The revised Leiter answers all criticisms handily, and the LIPS-R deserves wide use as a culture-reduced measure of nonverbal intelligence.

A remarkable feature of the Leiter is the complete elimination of verbal instructions. The Leiter-R does not require a single spoken word from the examiner or the examinee. With an age range of 2 years to 20 years and 11 months, the Leiter-R is particularly suitable for children and adolescents whose English language skills are weak. This includes children with any of these features: nonEnglish-speaking, autism, traumatic brain injury, speech impairment, hearing problems, or an impoverished environment. The test is also useful in the assessment of attentional problems, as described in the following.

Testing is performed by the child or adolescent matching small laminated cards underneath corresponding illustrations on an easel display (Figure 10.1). The test is untimed because the initial items

are transparently obvious, most examinees catch on quickly without need of pantomime demonstration. The Leiter-R contains 20 subtests organized into four domains: Reasoning, Visualization, Memory, and Attention. Not all subtests are administered to every child. For example, the figure rotation subtest is too difficult for 2-year-olds and the immediate recognition subtest is too easy for adolescent examinees. The four Reasoning subtests include classification and design analogies. The six Visualization subtests include matching, figure-ground, paper folding, and figure rotation. The eight Memory subtests include memory span, spatial memory, associative memory, and delayed recognition memory. The two Attention subtests consist of an underlining test (e.g., marking all squares printed on a page full of geometric shapes) and a measure of divided attention (e.g., observing a moving display and simultaneously sorting cards correctly).

The Leiter-R yields a composite 1Q with the familiar mean of 100 and standard deviation of 15. The test also produces subtest scaled scores with a mean of 10 and standard deviation of 3, as well as a variety of composite scores useful in clinical diagnosis. The test was normed on over 2,000 children and adolescents, from 2 to 21 years of age. Using 1993 census statistics, these subjects were carefully stratified according to race, age, gender, social class and geographic region. Internal consistency reliability for subtests, domain scores, and IQ scores is excellent. Typical coefficient alphas are in the high .80s for subtests and the low .90s for domain scores and IQ scores. Extensive studies of item bias reveal that the items appear to function similarly in separate racial groups (white, African American, and Hispanic samples); that is, there is no evidence of bias (defined as differential item functioning). Coupled with the fact that the test is completely nonverbal, the absence of test bias indicates that the Leiter-R is a good choice for culture-reduced testing of minority children.

Empirical research with the Leiter-R is scant at this time. The test has been shown to have utility in the assessment of medically fragile children (Hooper, Hatton, Baranek, Roberts, & Bailey. 2000) and the evaluation of children classified as language impaired (Farrell & Phelps, 2000). In this latter study, the Leiter-R also demonstrated a validity-confirming correlation of r = .80 with another nonverbal measure of intelligence. Studies with the first edition indicate strong relationships with other intelligence test scores. For example, the Leiter and the WISC Performance IQ correlated near the .80s; correlations with the WISC Verbal IQ are more typically in the .60s (Arthur, 1950; Matey, 1984). Reeve, French, and Hunter (1983) compared the Leiter and the Staniford-Binet: Form L-M as predictors of Metropolitan Achievement Test scores for 60 kindergartners. Correlations were .77 between Stanford-Binet and MAT total, and .61 between Leiter and MAT total. The authors note that although the Stanford-Binet proved to be a marginally better predictor of standard achievement, children with hearing and/or speech problems may require the Leiter or other nonverbal instruments.

The Leiter-R is a welcome revision of an obsolete test. In the hands of a careful clinician, the test is helpful in the intellectual assessment of children with weak skills in English. Other uses for the revised test include the assessment of attention-deficit/hyperactivity disorder (comparisons of the Attention subtests with the other domains are crucial here) and the evaluation of giftedness in young children (the extremely/ high ceiling of the test proves invaluable for this application). Whereas reviewers warned against using the original Leiter for placement or decision-making purposes (Sattler, 1988; Salvia & Ysseldyke, 1991), the revised Leiter is a huge improvement in regards to psychometric quality and standardization excellence. Thorough reviews of the Leiter-R and other nonverbal assessment instruments are provided by Athanasiou (2000) and McCallum, Bracken, and Wasserman (2001).

Human Figure Drawing Tests

Most children enjoy drawing human figures and do so routinely and spontaneously. Since the early 1900s, psychologists have tried to tap into this almost instinctive behavior as a basis for measuring intellectual development. The first person to use human figure drawing (HFD) as a standardized intelligence test was Florence Goodenough (1926). Her test, known as the Draw-A-Man test, was revised by Harris (1963) and renamed the Good-enough-Harris Drawing Test. More recently, the HFD technique has been adapted by Naglieri (1988). An additional approach by Gonzales (1986) is not reviewed here. We should also mention that human figure drawings are widely used as measures of emotional adjustment, but we do not discuss that application here.

The Goodenough-Harris Drawing Test is a brief, nonverbal test of intelligence that can be administered individually or in a group. Goodenough (1926) published the first edition of this test, while Harris (1963) provided important refinements in scoring and standardization, including the use of a deviation IQ. Strictly speaking, the Goodenough-Harris test doesn't fit the criteria for nonlanguage tests insofar as the examiner must convey certain instructions in English or through a translator. However, the instructions are brief and basic ("I want you to draw a picture of a man [or woman]; make the very best picture you can"). The Good-enough-Harris test is, for all practical purposes, a nonlanguage test.
The purpose of the Goodenough-Harris Drawing Test is to measure intellectual maturity, not artistic skill. Thus, the scoring guide emphasize accuracy of observation and the development of conceptual thinking. The child receives credit for including body parts and details, as well as for providing perspective, realistic proportion, and implied freedom of movement.

The 73 scorable items were selected according to the following criteria:

1. The items should show a regular and fairly rapid increase with age, in the percentage of children passing the point.
2. The items should show a relationship to some general measure of intelligence.
3. The items should differentiate between children scoring high on the scale as a whole and those scoring low on the scale as a whole (Harris, 1963).

In addition to the Man scale, the Harris (1963) revision also includes two additional forms: the Woman scale and the Self scale. For these last two scales, examinees are instructed to draw a picture of a woman and of themselves. Scores on the Man and Woman scales are very highly correlated for examinees of either sex (r = .91 to .98). These two versions can be considered equivalent forms. The Self scale was intended as a projective test of self-concept. However, self-concept is a fuzzy construct that is difficult to objectify. The Self scale has largely fallen by the wayside, although some psychologists use it purely as an unscored extension of the clinical interview.

The standardization sample for the Good-enough-Harris Drawing Test was large (N = 2,975 children), geographically varied (from urban and rural areas throughout the United States), and carefully selected to match U.S. population values for parental occupational status. The test covers ages 3 to 16, but the norms are best for ages 5 to 12. Beyond age 12, examinees begin to approach an asymptote of performance and age differences are reduced. The Man scale yields a deviation IQ-like standard score with mean of 100 and standard deviation of 15. One concern is simply that Drawing Test norms are now quite dated. Abell, Horkheimer, and Nguyen (1998) found that the scoring system for this test consistently underestimated IQ scores on the W1SC-R.

The reliability of the test has been assessed by split-half procedures, test-retest studies, and interscorer comparisons (Anastasi, 1975; Frederickson, 1985; Harris, 1963). Split-half reliabilities near .90 are common. However, stability coefficients seldom exceed the .70s, even when the test-retest interval is only a few weeks. This suggests that scores on the Goodenough-Harris Drawing Test possess a sizeable band of measurement error. On the other hand, scoring is quite objective: Interscorer correlations are typically in the .90s.

Examiners who have mastered the elaborate point scoring system may then use a simpler global method called the Quality Scale. The Quality Scale consists of 24 drawings (12 for the Man scale and 12 for the Woman scale) used as standardized reference points. The examiner matches the examinee's drawing to one of the 12 reference drawings, and then consults a table to determine the corresponding standard score. The Quality score is quicker, but slightly cruder: Interscorer reliabilities are typically in the low .80s.

The Goodenough-Harris test is often used as a nonverbal measure of cognitive ability with children who have language disabilities and minority or bilingual children. Oakland and Dowling (1983) view the Drawing Test as a culturally reduced test that is appropriate for initial screening of minority children. The test works best with younger children, particularly those with lower intellectual ability (Scott, 1981). For samples of 5-year-old children at a day care center for lower socioeconomic families, Frederickson (1985) reported correlations between Goodenough-Harris Drawing Test scores and WPPSI Full Scale IQ in the range of .72 to .80. In several other studies, correlations with individual IQ tests are more variable, but the majority are over .50 (Abell, Briesen, & Watz, 1996; Anastasi, 1975).

In response to criticisms of the Goodenough-Harris Drawing Test, Naglieri (1988) developed a quantitative scoring system and renormed the human figure drawing procedure. His scoring system, The Draw A Person: A Quantitative Scoring System (DAP), was normed on a sample of 2,622 individuals ages 5 through 17 years who were representative of the 1980 U.S. Census data on age, sex, race, geographic region, ethnic group, social class, and community, size. The DAP yields standard scores with the familiar mean of 100 and standard deviation of 15. In a study of 61 subjects ages 6 to 16 years, the DAP correlated .51 with WISC-R IQ and produced similar overall scores, with a mean IQ of 100 versus mean DAP score of 95 (Wisniewski

& Naglieri, 1989). Lassiter and Bardos (1995) found that the DAP score underestimated IQ scores obtained from the WPPSI-R and the K-BIT in a sample of 50 kindergartners and first graders.

Reviewers praise the DAP for its clear scoring system, strong reliability, and careful standardization (Cosden, 1992). However, results of validity studies are more cautionary. Harrison and Schock (1994) note that the accumulated evidence with HFD tests indicates low to moderate predictive validity. In spite of their popularity and appeal, HFD tests do not effectively identify children with learning difficulties or developmental disabilities, and they may not be valid for use even as screening measures.

Hiskey-Nebraska Test of Learning Aptitude

The Hiskey-Nebraska Test of Learning Aptitude (H-NTLA) is a nonlanguage performance scale for use with children ages 3 to 17 years (Hiskey, 1966). This test can be administered entirely through pantomime and requires no verbal response from the examinee. However, verbal instructions can be used with children with normal and mild hearing impairment. The H-NTLA consists of 12 subtests:

Bead Patterns
Block Patterns
Memory for Color
Completion of Drawings
Picture Identification
Memory for Digits
Picture Association
Puzzle Blocks
Paper Folding
Picture Analogies
Visual Attention Span
Spatial Reasoning

Raw scores on the subtests are converted into a Deviation Learning Quotient (LQ) with mean of 100 and standard deviation of 16. H-NTLA scores correlate quite robustly with achievement scales for grades 2 through 12 (median r = .49) and also with WISC-R Performance IQ (r = .85). Although the LQ yields average scores that are remarkably close to WISC-R Performance IQ for samples of children with hearing impairment and those who are deaf the H-NTLA scores are substantially more variable (Watson & Goldgar, 1985; Phelps & Ensor, 1986). Thus, use of the H-NTLA may increase the risk of false positive misclassification—labeling children as gifted when they are only bright, or as having mental retardation when they are merely borderline.

The H-NTLA is useful with children who are deaf, have speech or language impairments or mental retardation, or those who are bilingual. An interesting feature of this test is the development of parallel norms: The H-NTLA was standardized on 1,079 children who were deaf and 1,074 normal hearing children ages 2Vz to 17!/2. However, the chief weakness of the instrument is the inadequacy of these norms. For example, the representativeness of the sample of those who were deaf—picked on an opportunistic basis from schools for those who are deaf—is largely unknown. Standardization of the normal-hearing sample was based on occupational level of parents according to the 1960 U.S. Census. A contemporary and more detailed re-standardization of the test would be quite helpful.

Test of Nonverbal lntelligence-3

The Test of Nonverbal Intelligence-3 (TONI-3) is a language-free measure of cognitive ability designed for disabled or minority populations (Brown, Sherbenou, & Johnsen, 1998). In particular, the authors recommend the test for assessing persons with aphasia, non-English speakers, those with hearing impairments, and persons who have experienced a variety of severe neurological traumas. The test instructions are pantomimed by the examiner and the examinee answers by pointing to one of six possible responses. The test consists of two equivalent forms of 50 abstract/figural problem solving items. These items were carefully selected from an initial pool of items according to item total correlations, appropriate difficulty level, and acceptability to potential users and technical experts. The TONI-3 items fall into several categories, including the following:

Simple matching
Analogies
Classification
Intersection
Progressions

Except for the simple-matching items, the TONI-3 items require the examinee to solve problems by identifying relationships among abstract figures. Many of the items are similar in format to those found on Raven's Progressive Matrices. The test yields two kinds of scores: percentile ranks and TONI-3 quotients (mean of 100 and standard deviation of 15).

The TONI-3 was carefully standardized on over 3,000 subjects ranging in age from 6 through 89. Sample characteristics paralleled census data for sex, race, ethnicity, urban-suburban-rural residence, grade, parental education/occupation, and geographic region. Reliability data are quite satisfactory, with internal consistency coefficients typically exceeding .90 and alternate-forms reliability in the range of .80 to .95.

Validity studies of the TONI-3 are scant, but investigation of prior editions (which are highly similar in content) are supportive of this test as a culture-reduced index of general intelligence. Nonetheless, research does not support the view that the TONI-3 is a nonverbal test, except in the trivial sense that verbal responses are not required. For example, the TONI-2 manual reports correlation coefficients in the .70s between TONI-2 scores the Language Arts subtest of the SRA Achievement Series. In general, research studies with precursors to the TONI-3 indicate that it is a good Measure of general intelligence, but they do not Support the view that it is mainly a measure of non-verbal intelligence (Murphy, 1992). Overall, the TONI-3 is highly regarded as a brief nonlanguage screening device for subjects with impaired language abilities (e.g., for those who are aphasic, deaf, or non-English-speaking or who have mental retardation). The test is more carefully standardized than most and possesses excellent reliability. A useful feature of the TONI-3 is that the untimed administration seldom exceeds 20 minutes.

Two instruments discussed earlier in the text also qualify as nonlanguage tests. Raven's Progressive Matrices and the Cattell Culture Fair Intelligence Test utilize nonverbal items and require essentially no language-based interactions between examiner and examinee. A new and promising language-free test is the Universal Nonverbal Intelligence Test (UNIT), a comprehensive and multidimensional measure of nonverbal intelligence (McCallum & Bracken, 1997; Reed & McCallum, 1995). This test is designed for children with hearing impairment or limited English proficiency. Sophisticated item analyses indicate that the UNIT is an unbiased measure of nonverbal intelligence in children who are profoundly deaf (Mailer, 2000). The UNIT provides a good measure of g and several subscores, including clear, factor-based scores on memory and reasoning.

NONREADING AND MOTOR-REDUCED TESTS

As the reader will recall, nonreading tests are designed for illiterate examinees who can, nonetheless, understand spoken English well enough to follow oral instructions. Nonreading tests of intelligence are well suited to young children, illiterate examinees, and persons with speech or expressive language impairments. These tests need not be specialized or esoteric: The performance subtests of most mainstream instruments qualify as non-reading tests. For example, examiners may use the WISC-III performance subtests to estimate the intelligence of examinees with language disabilities.

However, clients with cerebral palsy or other orthopedically impairing conditions will score very poorly on nonreading tests that require manipulatory responses. Obtaining valid test results from such persons can present an enormous challenge. The motor deficits, increased tendency to fatigue, and inexactness of purposive movements common to persons with cerebral palsy will negatively affect their performance on cognitive assessment tools. Orthopedically impaired clients need tests that are both nonreading and motor-reduced. In particular, tests that permit a simple pointing response are well suited to the assessment of children and adults with cerebral palsy or other motor-impairing conditions.

Case Exhibit the challenge of assessment in cerebral palsy

The challenges inherent to special consultations are well typified by a client with cerebral palsy recently tested by a consulting psychologist. The young examinee was totally confined to a battery-powered wheelchair, except when a live-in attendant would transfer him to a bed or chair. Even a dispassionate ob-server would have to agree that the client didn't look very capable, sitting hunched over in his chair, unable to control his drooling, one arm arched out at an awkward angle. Yet, in spite of his disability, he had achieved a fair degree of personal independence. Using a simple joystick control device, he could guide his wheelchair to the grocery store, library, and community center where he would complete simple transactions by pointing to appropriate words and phrases in a plastic-bound spiral notebook. Because of his poor motor control, interactions with this client took quite a long time. Nonetheless, he was very efficient with short communications. Here is a typical exchange, j with the client's notebook-designated responses shown in capital letters:

"I understand you have a new synthesized-voice communication box, how do you like it?" YOU ASKED TWO QUESTIONS. "You're right. I'll bet that happens a lot. Do you have a communication box?" YES. "What do you think of it?" IT'S NOT EASY. "Now that we are done testing, should I find your driver?" NO, I'LL WAIT. HE IS COMING BACK.

How intelligent is this client? What is his level of verbal comprehension? How well does he understand abstract concepts? For example, is he capable of understanding the essentials of microcomputer usage such as data entry, file storage, and directory commands? Could he learn to program a microcomputer? These are precisely the referral questions asked by a vocational rehabilitation counselor who was contemplating huge expenditures—thousands of dollars—to purchase a computer system for this disabled client.

Certainly it would be easy to underestimate the potential of this young man with severe motor and language disabilities because —in a quite literal sense— his intelligence was hidden away, trapped inside his incapacitated body. The task of the examiner was to find the able mind inside the disabled body, a formidable challenge indeed. Using the Test of Nonverbal Intelligence-2 and the Peabody Picture Vocabulary Test-Revised, the examiner determined that the young client possessed at least average intelligence and could likely learn the fundamentals of data processing with microcomputers.

Peabody Picture Vocabulary Test-Ill

Peabody Picture Vocabulary Test-Ill (PPVT-III) is the best known and most widely used of the nonreading, motor-reduced tests (Dunn & Dunn. 1998) The PPVT-1II is used to obtain a rapid measure of listening vocabulary with persons who are (leaf or who have neurological or speech impairments. Although the PPVT-III is useful with any examinee who cannot verbalize well, the test is especially useful with examinees who also manifest motor-impairing conditions such as cerebral palsy or stroke.

The PPVT-III conies in two parallel versions, each consisting of 4 practice plates and 204 testing plates. Each plate contains four line drawings of objects or everyday scenes. The examiner presents a plate, states the stimulus word orally, and asks the examinee to point to the one picture that best depicts e stated word. The test items are precisely ordered cording to difficulty level, arranged in 17 sets of 2 items each for efficient identification of basal and ceiling levels. The entry level is determined by age and examinees continue until they reach their ceiling level. Although the test is untimed, administration seldom exceeds 15 minutes. Raw scores are converted to age equivalents or standard scores (mean of 100, standard deviation of 15).

The PPVT-III was standardized on a representative national sample of 2.725 individuals ranging from 2'/2 to 90 or more years of age. Reliability data for the new edition are exceptionally strong, with typical internal consistency coefficients of .94, alternate-forms reliabilities of .94, and test-retest correlations of .92. Concurrent validity studies are also highly supportive, demonstrating robust correlations with verbal intelligence measures. For example, the test developers report correlations of .91 with WISC-III Verbal IQ and .82 with K-BIT Vocabulary scores (Dunn & Dunn, 1998).

The test developers of the PPVT-III took great care to minimize and balance cultural influences in the test items. Independent consultants representing the perspectives of African Americans, Asians, Hispanics, Native Americans, and women reviewed the content and artwork of the FPVT-III during development, and adjustments were made following these reviews. The test items denonstrate attractive art work that is balanced for racial and gender differences, including persons with physical disabilities. However, the evidence is mixed as to whether the PPVT-III is a culturally fair instrument that serves as a valid measure with minority children. For example, Washington and Craig (1999) found that 59 African American preschoolers at risk for academic failure averaged 91 on the test (SD of 11), which was seen as commensurate with their environmental disadvantages. These authors laud the test as "culturally fair." However, Campbell. Bell, and Keith (2001) reported an average score of 82 (SD of 12) for 416 African American children of low socioeconomic status, which was 8 points lower than their overall score on the K-ABC. These researchers concluded: "Despite the attempts to reduce racial differences, the PPVT-III appears to perform similarly to prior editions of the Peabody scales. On average, the PPVT-III tends to underestimate both intellectual ability and scholastic achievement, as measured by the K-ABC, in low SES, African American children". Further research will be needed to clarify the utility of this test with minority children.

Several lines of evidence support the validity of the Peabody test, but only as a narrow measure of vocabulary, not as a general measure of intelligence (Altepeter. 1989; Altepeter& Johnson, 1989). Dunn and Dunn (1981) sought to ensure content validity by searching Webster's New Collegiate Dictionary for all words whose meanings could be represented by a picture. Thus, the authors had a specific content universe in mind, and the items from the Peabody appear to be a fair sampling from this domain. In addition, the authors used sophisticated item-selection techniques based on the Rasch-Wright latent-trait model to help build construct validity into the test. This model enables researchers to construct a growth curve for the latent trait being measured (hearing vocabulary) and to select items that best fit the curve. Using tryout and calibration data, the curve was drawn repeatedly on a computer. If an item did not tit the Rasch-Wright latent-trait model (too flat or too steep an item-characteristic curve) it was discarded from consideration.

Using a sophisticated structural equation model, Miller and Lee (1993) demonstrated that an earlier edition, the PPVT-R, can be assumed to reflect true developmental level of vocabulary. These researchers were able to predict rank order of the PPVT-R stimulus words reasonably well based upon complex word characteristics (date of entry into the English language, word length, number of separate meanings, and frequency of occurrence). The predictor variables provided a reasonable theoretical account of the word ordering in the PPVT-R; that is, they confirmed the construct validity of the test.

Concurrent and predictive validity data for the Peabody are somewhat limited, but promising. Several investigators have correlated the PPVT-R with achievement measures, where modest relationships (r's from .30 to .60) are common (Naglieri, 1981; Naglieri & Pfeiffer, 1983). Correlations with reading achievement tend to be higher than with spelling and arithmetic achievement, suggesting that the PPVT-R has appropriate discriminant validity (Vance, Kitson, & Singer, 1985).

Several investigators have correlated earlier versions of the Peabody with intelligence measures, particularly the WISC-R and WAIS-R, and healthy correlations (near .70) are the rule (e.g., Haddad, 1986; Naglieri & Yazzie, 1983). As might be expected, correlations tend to be higher with Verbal IQ than Performance IQ.

In a very important and ingenious study, Maxwell and Wise (1984) investigated the vocabulary loading of the Peabody in a sample of 84 inpatients from psychiatry and psychology wards. Their study utilized the

PPVT, but this earlier edition is similar to the PPVT-III, so that the conclusions are pertinent here. The researchers investigated the hypothesis that the PPVT assesses more than vocabulary in adults. In addition to the PPVT, the researchers collected data on the following: WAISR, Wechsler Memory Scale, name-writing speed, and years of education. Name-writing speed is simply the number of seconds required for the examinee to write his or her full name. Even though all variables h significant correlations with PPVT IQ, Waisr Vocabulary had by far the strongest correlation (.88). More important, when the variance accounted for by Vocabulary was removed, none of the remaining variables had any predictive relationship with the PPVT. In short, the Peabody is a g00J measure of vocabulary (hearing vocabulary, in particular) but could be misleading if used as a global measure of intellect.

The PPVT-III is a recent revision, so independent research with the test is limited. One caution with the previous edition, the PPVT-R. is that standard scores may be substantially lower than Wechsler IQs, particularly with persons with mental retardation and minority examinees. In a sample of 21 adults with mild mental retardation. Prout and Schwartz (1984) found the PPVT-R standard scores (mean of 56) to be an average of 9 points lower than the WAIS-R IQ (mean of 65). Naglieri and Yazzie (1983) found a huge 26-point difference with a sample of Navajo Indian children, who averaged a standard score of 61 on the PPVT-R in contrast to WISC-R IQ of 87. On a similar note, with the PPVT-III, Bell, Lassiter, Matthews, and Hutchinson (2001) found that the instrument tended to underestimate WAIS-III IQ scores of bright college students by about 10 points.

Overall, we may conclude that the Peabody is a well-normed measure of hearing vocabulary that is useful with nonreading and motor-impaired examinees. However, the instrument is not a substitute for a general intelligence test and PPVT-III scores may underestimate intellectual functioning in some groups (e.g., minority children, high-functioning adults).

Testing Persons with Visual Impairments

Many millions of American adults have some degree of visual impairment, including more than 1 million individuals who are legally blind—a term used in determining eligibility for government benefits. This term applies to individuals with central visual acuity of 20/200 or less in the better eye (with correction) or to those with significant reduction in their visual field to a diameter of degrees or less (Bradley-Johnson & Ekstrom, 1998). The number of children with visual impairment is substantially smaller, with only 0.4 percent of students between the ages of 6 and 21 years receiving special education services because of a vision problem (U S. Department of Education, 1992). In addition to special arrangements in testing, individuals with visual impairment may require unique instruments for valid assessment.

In assessing the intellectual functioning of the visually impaired, examiners have historically relied upon adaptations of the Stanford-Binet. The Hayes-Binet revision for testing those with visual impairment was based on the 1916 Stanford-Binet; this instrument has since undergone several revisions. The most recent adaptation is the Perkins-Binet (Davis, 1980). The Perkins-Binet retains most of the verbal items from the Stanford-Binet, but also adapts other items to a tactual mode. The Perkins-Binet possesses acceptable split-half reliability and shows high correlations with verbal scales of the WISC-R (Coveny, 1972; Teare & Thompson, 1982). The developers of the Perkins-Binet have acknowledged that visual problems exist on a continuum by developing separate norms for children with usable vision (Form U) and no usable vision (Form N).

Test developers have also succeeded in modifying the Wechsler Performance scales for use with individuals with visual impairments. The Haptic Intelligence Scale for the Adult Blind (HISAB) consists of six subtests, four of which resemble the Digit Symbol, Block Design, Object Assembly, and Picture Completion tests of the WAIS Performance scale (Shurrager, 1961; Shurrager & Shurrager, 1964). The remaining two subtests consist of Bead Arithmetic, which involves the use of an abacus to solve arithmetic problems, and a Pattern Board, which requires the examinee to reproduce the pattern felt on a board that has rows of holes with pegs in them. The reliability of the HISAB is excellent and the authors provide normative data on a sample of adults with visual impairment. Most encouraging of all, HISAB scores correlate .65 with the WAIS

Verbal IQ (Shurrager & Shurrager, 1964). Although the HISAB is still manufactured and sold by Stoelting Company, unfortunately, the test has never been investigated empirically. A search of PsychlNFO for research with this instrument did not locate a single article.

Another interesting instrument is the Blind Learning Aptitude Test (BLAT), a tactile test for children from 6 to 16 years of age who are blind (Newland, 1971). The BLAT items are in bas-relief form, consisting of dots and lines similar to Braille. The items consist of six different types: recognition of differences, recognition of similarities, identification of progressions, identification of the missing element in a 2 x 2 matrix, completion of a figure, and identification of the missing element in a 3 x 3 matrix. Most of the items were adapted from Raven's Progressive Matrices and the Cattell Culture Fair Intelligence Test. The BLAT is standardized on 760 children with visual impairment, but the norms are outdated and the test manual is incomplete and somewhat slipshod (Herman, 1988). Nonetheless, the test possesses exceptional reliability and correlates very well with the Hayes-Binet (r = .74) and the WISC Verbal scale (r = .71). The BLAT also shows strong correlations with Braille oral reading speed and comprehension (Baker, Koenig, & Sowell, 1995). In conjunction with a verbal test, the BLAT is a promising instrument for testing the intelligence of children with visual disabilities. However, the test would profit substantially from minor revisions, updated norms, and a more thorough test manual.

TESTING INDIVIDUALS WHO ARE DEAF OR HARD OF HEARING

Upward of 1 million Americans are deaf or sufficiently hard of hearing that they rely upon American Sign Language (ASL) as their primary means of communication (Brauer, Braden, Pollard, & Hardy Braz, 1998). Given the typical limited mastery of the English language of persons who are deaf, and, vice versa, the typical psychologist's limited (or nonexistent) skill in ASL, the proper and valid assessment of individuals who are deaf poses a profound cross-cultural challenge.

More is involved than just picking a test developed for, and normed upon, individuals who are deaf or hard of hearing and who use sign language. One problem is that sign language "'can now be characterized on a multidimensional continuum encompassing numerous styles, lexical variants, syntactic structures, dialects, and approximations to or departures from English word ordering" (Brauer et al., 1998, p. 299). Thus, a test developed in standard ASL is not equally fair to all persons who are deaf. In general, the proper and valid assessment of persons who are deaf requires that interested psychologists immerse themselves in the Deaf culture and also seek relevant educational and training experiences:

> *One especially needs a thorough understanding of the implications of deafness and the use of sign language for making diagnoses for people who are deaf. Few hearing psychologists have these skills. The push is for specialized training programs in deafness and psychology, a need that has been recognized for decades. (Brauer et al., 1998, p. 303)*

If a consulting psychologist does not possess these skills, then the assessment of persons who are deaf should be referred to a person or agency with the requisite talents and expertise.

The use of a sign language interpreter in the testing of persons who are deaf is a complicated and controversial matter. One concern is that the interpreter may inadvertently alter the content of the test, therefore affecting the validity of the findings. Certainly, it is unwise for parents or teachers to serve as interpreters. However, it is also true that persons who are deaf and who use sign language achieve higher IQs when the directions are signed than when they are delivered in the traditional manner (Braden, 1992). The preferred resolution is for the examiner to be fluent in sign language, so that any necessary translations stay within the bounds of standardized procedure.

For the intellectual assessment of persons who are deaf or hard of hearing, the Wechsler Performance subtests remain the tools of choice (Braden & Hannah, 1998). The impact of English language facility is minimized on these subtests, so it is thought that they provide a more accurate measure of cognitive skill than the Verbal subtests. Others tests sometimes used with persons who are deaf include Raven's Progressive Matrices (Raven, Court, & Raven, 1992) and the Hiskey Nebraska Test of Learning Aptitude, discussed previously. The WAIS-III is now available in a formal ASL translation (demonstrated on videotape), endorsed and disseminated by the test publisher (Kostrubala & Braden, 1998).

ASSESSMENT OF ADAPTIVE BEHAVIOR IN MENTAL RETARDATION

The assessment of mental retardation is a complex and multifaceted concern that rightfully deserves a chapter or book on its own. Owing to space limitations, our coverage is necessarily abridged; interested readers are referred to American Association on Mental Retardation (2002), Nihira (1985), and Sattler (1988, chaps. 15 and 21). Here, we briefly summarize the diagnostic criteria for mental retardation, then review two contrasting assessment instruments in modest detail. We close with a tabular summary of several prominent measures of adaptive behavior.

Definition of Mental Retardation

The most authoritative source for the definition of mental retardation is the manual of terminology and classification of the American Association on Mental Retardation (AAMR, 2002). This manual defines mental retardation as follows:

> *Mental retardation refers to substantial limitations in present functioning. It is characterized by significantly subaverage intellectual functioning, existing concurrently with related limitations in two or more of the following applicable adaptive skill areas: communication, self-care, home living, social skills, community use, self-direction, health and safety, functional academics, leisure, and work.*
> *Mental retardation manifests before age 18. (AAMR, 2002)*

The manual further specifies that significantly subaverage intellectual functioning is an IQ of 70 to 75 or below on scales with a mean of 100 and a standard deviation of 15. On tests such as the Stanford-Binet: Fourth Edition that possess a standard deviation of 16, the approximate range for retarded intellectual functioning would be an IQ of 68 to 73 or below. The manual also explicitly affirms the importance of professional judgment in individual cases.

A low IQ by itself is an insufficient foundation for the diagnosis of mental retardation. The AAMR definition also specifies a second criterion, that of limitations in two or more of the relevant adaptive skill areas. A diagnosis of mental retardation is warranted only when an individual displays a sufficiently low IQ and limitations in adaptive skill. Further, these deficits in intellect and adaptive functioning must have arisen during the developmental period—defined as between birth and the eighteenth birthday.

This most recent AAMR manual represents a departure from previous terminology, which recognized four levels of retardation: mild, moderate, severe, and profound. Instead of focusing upon the shortcomings of the person, the manual introduces a hierarchy of "Intensities of Needed Supports," which redirects attention to the rehabilitation needs of the client. The four levels of needed supports are intermittent, limited, extensive, and pervasive. However, the previous terminology referring to levels of retardation will likely prevail for quite some time, so we have chosen to blend the old and the new approach in Table 10.1. The reader will notice a zone of uncertainty between levels of retardation, which signifies that clinical judgment about all sources of information is required in diagnosis. Furthermore, even though these levels are calibrated by IQ ranges, we remind the reader that the examinee must also show corresponding deficit in two or more areas of adaptive skill. Under no circumstances is an IQ test a sufficient basis for diagnosing mental retardation.

Limitations in adaptive skill are more difficult to confirm than a low IQ. The AAMR manual lists 10 different areas of adaptive skill and specifies that

Table 10.1 Four Levels of Mental Retardation

Mild Mental Retardation: IQ of 50-55 to 70-75+, Intermittent Support required. Reasonable social and communication skills; with special education, attain 6th grade level by late teens; achieve social and vocational adequacy with special training and supervision; partial independence in living arrangements.

Moderate Mental Retardation: IQ of 35-40 to 50-55, Limited Support required. Fair social and communication skills but little self-awareness; with extended special education, attain 4th grade level; function in a sheltered workshop but need supervision in living arrangements.

Severe Mental Retardation: IQ of 20-25 to 35-^0, Extensive Support required. Little or no communication skills; sensory and motor impairments; do not profit from academic training; trainable in basic health habits.

Profound Mental Retardation: IQ below 20-25, Pervasive Support required. Minimal functioning; incapable of self-maintenance; need constant nursing care and supervision.

Source: Based on AAMR (2002) and Patton. Payne, and Beirne-Smith (1986).

The client must show substantial limitations in two or more of them:

- Communication
- Self-care
- Home living
- Social skills
- Community use
- Self-direction
- Health and safety
- Functional academics
- Leisure

As to how these limitations are to be assessed, the manual proposes that well normed measured of adaptive skills are desirable, but the final determination is always a matter of clinical judgment.

A test developer faces major problems in calibrating limitations in adaptive skill. About the only hard fact we have in this domain is that environmental expectations for adaptive behavior increase sharply from birth through young adulthood. In addition, the expression of adaptive behavior changes character throughout. In childhood, adaptive behaviors may be reflected in sensory-motor skills and facility with language. In adulthood, vocational attainment and social responsibility become important. Just as with intellectual assessment, tools for appraising adaptive behavior must be carefully age-graded.

The first standardized instrument for assessing adaptive behavior was the Vineland Social Maturity Scale (Doll, 1935, 1936). Somewhat simplistic and coarse-grained by modern standards, the original Vineland scale consisted of 117 discrete items arranged in a year-scale format. An information familiar with the examinee would check off applicable items. From these results the examiner would calculate and equivalent social age, helpful in the diagnosis of mental retardation. Still a respected instrument, the Vineland has undergone several revisions and is now known as the Vineland Adaptive Behavior Scales (Sparrow, Balla, & Cicchetti, 1984).

Since the release of the original Vineland scale, over 100 scales of adaptive behavior have been published (Nihira, 1985; Reschly, 1990; Walls, Werner, Bacon, & Zane, 1977). These instruments vary greatly in structure, intended purpose, and targeted population. Broadly speaking, we can distinguish two types of instruments designed for two different purposes. One group of mainly norm-referenced scales is used largely to assist in diagnosis and classification. Another group of mainly criterion-referenced scales is used largely to assist in training and rehabilitation. We have chosen one representative instrument from each group for more detailed analysis.

Scales of Independent Behavior-Revised

The Scales of Independent Behavior-Revised (SIB-R; Bruininks, Woodcock, Weatherman, & Hill, 1996) is an ambitious, multidimensional measure of adaptive behavior that is highly useful in the assessment of mental retardation. The instrument consists of 259 adaptive behavior items organized into 14 subscales. The scale is completed with the help of a parent, caregiver, or teacher well acquainted with the examinee's daily behaviors. B each subscale, the examiner reads a series of items and for each item records a score from 0 (never or rarely does task) to 3 (does task very well). A useful feature of the SIB-R is that examiners need a minimum of training and experience. Of course, a much higher level of competence is required to evaluate results and make decisions about placement or treatment.

The 14 subscales of the SIB are arranged into 4 clusters, as outlined in Table 10.2. In turn, these 4 clusters constitute the Broad Independence Scale. Each subscale consists of a small number of discrete, developmentally ordered items. For example, the subscale on Eating and Meal Preparation has 19 graded items, including spearing food with a fork, eating soup with a spoon, taking appropriate-sized portions, and preparing snacks that do not require cooking. For each subscale, items are administered until a predetermined ceiling is reached (e.g., 3 of 5 consecutive items scored 0).

Raw scores for a subtest are added to obtain a part score. The part scores for each cluster are then added to obtain the cluster score. The score for the Broad Independence Scale is derived from the four cluster scores. The subtest scores, cluster scores, and the Broad Independence score can then be converted to a variety of normative scores to permit comparison of the examinee's performance with the performance of the national norming sample. The normative scales include age scores, percentile ranks, standard scores, stanines, and normal curve equivalents.

A separate, unique part of the SIB-R also assesses maladaptive behavior by measuring the frequency and severity of problem behaviors. The Problem Behaviors Scale includes eight major categories of personal and social maladjustment that could affect adaptive behavior: Hurtful to Self, Hurtful to Others, Destructive to Property, Disruptive Behavior, Unusual or Repetitive Habits, Socially Offensive Behavior, Withdrawal or Inattentive Behavior, and Uncooperative Behavior. Examples of problem behaviors are listed, and the respondent must indicate the behaviors displayed by the examinee. In addition, the respondent

Table 10.2 The Subscales and Clusters of the Scales of Independent Behavior-Revised

1. **Motor Skills**

Gross Motor—19 large muscle skills such as sitting without support or taking part in strenuous physical activities.
Fine Motor—19 small muscle skills such as picking up small objects or assembling small objects.

2. **Social and Communication Skills**

Social Interaction—18 skills requiring interaction with other people such as handing toys to others or making plans with friends to attend social activities.
Language Comprehension—18 skills involving the understanding of spoken and written language such as looking toward a speaker or reading.
Language Expression—20 tasks involving talking such as making sounds to get attention or explaining a written contract.

3. **Personal Living Skills**

Eating and Meal Preparation—19 skills related to eating and meal preparation, ranging from drinking from a glass to planning a meal.
Toileting—17 skills necessary to bathroom and toilet use.
Dressing—18 skills related to dressing, ranging from holding out arms and legs while being dressed to arranging for clothing alterations.

Personal Self-Care—16 tasks involved in basic grooming and health maintenance, for example, washing hands and making a medical appointment.
Domestic Skills—18 tasks needed to maintain a home, ranging from putting empty dishes in the sink to selecting appropriate housing.

4. **Community Living Skills**

Time and Punctuality—19 tasks involving time concepts and time management such as keeping appointments.
Money and Value—20 skills related to money concepts, such as saving money and using credit.
Work Skills—20 skills related to prevocational and work habits, for example, indicating that an assigned task is completed.
Home-Community Orientation—18 skills involved in getting around the home and neighborhood and traveling in the community, for example, locating a dentist.

describes the one most serious behavior in each category and rates it according to frequency of occurrence, severity, and typical management.

The standardization of the SIB-R was well conceived and executed. The norm group consisted of 2,182 persons sampled to reflect the 1990 census characteristics. The normative data cover persons from age 3 months to adults over age 80. An additional sample of persons with mental retardation, learning or hearing disabilities, and behavior disorders was also tested. The value of the SIB-R was further strengthened by anchoring it to the norms for the Woodcock-Johnson Psycho-Educational Battery-Revised. The SIB-R is one component of this larger test battery, but can be used on its own.

The reliability of the SIB-R is generally respectable, but somewhat variable from subscale to subscale and from one age group to another. The individual subscales tend to show split-half reliabilities in the vicinity of 0.80; the four clusters have median composite reliabilities around 0.90; the Broad Independence Scale has a very robust reliability in the high .90s (Bruininks, Woodcock, Weatherman, & Hill, 1996).

Initial validity data for the SIB-R are very promising. For example, the mean scores of various samples of disabled and nondisabled subjects show confirmatory relationships: SIB-R scores are lowest among those persons known to be most severely impaired in learning and adjustment. For disabled examinees, SIB-R scores correlate very strongly with intelligence scores (in the .80s), whereas with nondisabled examinees, the relationship is minimal (Bruininks et al., 1996).

In sum, the SIB-R is an excellent tool for providing insights into an examinee's current level of functioning in real-life situations in the home, school, and community settings. Although this instrument does not have a one-to-one correspondence with the 10 areas of adaptive skill listed in the definition of mental retardation, there is substantial similarity. For example, the following areas of AAMR-listed adaptive skills are well covered by subscales or clusters of the SIB-R: communication, self-care, home living, social skills, community use health and safety, and work. The SIB-R or a similar instrument ranks as a mandatory supplement to individual intelligence testing in the diagnosis and assessment of mental retardation.

Independent Living Behavior Checklist (ILBC)

The Independent Living Behavior Checklist (ILBC) is an extensive list of 343 independent living skills classified and presented in six categories: mobility, self-care, home maintenance and safety, food, social and communication, and functional academic (Walls, Zane, & Thvedt, 1979). Unlike most of the instruments discussed so far in this text, the ILBC is completely nonnormative. The sole purpose of the ILBC is to facilitate the training of the individual examinee in the skills required for independent living. For this purpose, a collection of carefully selected criterion-referenced skills works better than a group of norm-based scores. The ILBC focuses on what the examinee can do, not on how the examinee compares to other persons. An exact age range is not specified, but the instrument appears to be suitable for persons 16 years of age through adulthood.

For each skill, the ILBC specifies a condition, a behavior, and a standard. Table 10.3 lists a sample of ILBC items. The reader will notice that all three components (condition, behavior, and standard) are defined with enough precision that reasonable observers would likely agree when a skill has been mastered. In fact, test-retest and interobserver agreement for ILBC skills range from .96 to a perfect 1.00.

The items within each ILBC category were carefully selected to encompass the important and relevant skills for independent living. Apparently, the authors succeeded in identifying essential skills; insofar as their instrument has a 100 percent overlap with another—initially unknown—checklist for independent living (Schwab, 1979). In addition, the ILBC items were carefully ordered from easiest to hardest. When used on a continuing basis over a several-year training period, the ILBC thus provides a checklist of skills mastered and also furnishes guidance for further rehabilitation.

Additional Measures of Adaptive Behavior

We remind the reader that measures of adaptive behavior vary greatly. Some scales are designed mainly for diagnosis, others for remediation. Some scales are useful with persons with severe and profound mental retardation who will never be employed, others with individuals with mild mental retardation seeking vocational training. Some scales are useful exclusively with children, others with adults. These instruments are not interchangeable, and the potential user must study their strengths and limitations carefully.

The Vineland Adaptive Behavior Scales (VABS; Sparrow, Balla. & Cicchetti, 1984) is the most widely used measure of adaptive behavior in existence. The instrument is the outcome of a major revision and restandardization of the Vineland Social Maturity Scale, originally published in 1935 by Edgar A. Doll. Based upon a semistructured interview with a caregiver or parent, the VABS provides

Table 10.3 A Sampling of ILBC Items

Rubber Scraper 35

Condition: Given a bowl containing ingredients, a pan, and a rubber scraper
Behavior: Client pours the ingredients into the pan and scrapes the sides of the bowl
Standard: Behavior within 2 minutes. No ingredients must be spilled. All ingredients must be removed from the bowl

Compliments 30

Condition: Given a role play or natural situation in which the client is complimented
Behavior: Client accepts the compliment(s) (e.g., says "Thank you.")
Standard: In the role play or natural situation, all persons interviewed must independently state that the client accepted the compliment(s) politely and was not overly gracious or vain

Address 38

Condition: Given a piece of paper with an address of place located within 3 blocks of the client
Behavior: Client finds the appropriate location with or without assistance
Standard: Behavior within one hour. The appropriate location must be found. The location may be found by the client alone or by the client with assistance (e.g., asking directions from others such as a policeman)

an evaluation in the following domains and subdomains: Communication (receptive, expressive, written), Daily Living Skills (personal, domestic, community), Socialization (interpersonal relationships, play and leisure time, coping skills), Motor Skills (gross, fine).

The VABS is a widely respected instrument with good concurrent validity, including correlations in the range of .50 to .80 with the WISC-R and Stanford-Binet. However, some of the interview items require knowledge that the informants may not possess (e.g., whether a child says 100 recognizable words). Silverstein (1986) faults the normative data, noting discontinuous jumps in standard scores from one age group to another. Even so, the Vineland continues to be a highly popular test in clinical practice and research.

The American Association on Mental Retardation (AAMR) has developed several scales useful in the assessment of persons with cognitive limitations. We mention here just one of its products, the AAMR Adaptive Behavior Scales: Second Edition (Nihira, Leland, & Lambert, 1993). The residential and community version of this test, suitable for persons 18 to 80 years of age, is a psychometric tour de force that borders on overkill. The normative sample includes more than 4,000 persons with developmental disabilities from 43 states, residing in the community or in residential settings. In addition to assessing the appropriate behavioral domains (e.g., independent functioning, domestic activity, self-direction, and responsibility), a noteworthy feature of the instrument is the careful attention to maladaptive behaviors, which are evaluated in eight domains:

- Violent and antisocial behavior
- Rebellious behavior
- Eccentric and self-abusive behavior
- Untrustworthy behavior
- Withdrawal
- Stereotyped and hyperactive behavior
- Inappropriate body exposure
- Disturbed behavior

This scale has been extensively validated clearly distinguishes persons independently classified at different adaptive behavior levels.

TEST BIAS AND OTHER CONTROVERSIES

THE QUESTION OF TEST BIAS

Beyond a doubt, no practice in modern psychology has been more assailed than psychological test Commentators reserve a special and often vehement condemnation for ability testing in particular. In his wide-ranging response to the hundreds criticisms aimed at mental testing, Jensen (1980) concluded that test bias is the most common rallying point for the critics. In proclaiming test bias the skeptics assert in various ways that tests are culturally and sexually biased so as to discriminate unfairly against racial and ethnic minorities, wornen and the poor. We cite here a sampling of verbatim criticisms (Jensen, 1980):

- Intelligence tests are sadly misnamed because they were never intended to measure intelligence and might have been more aptly called CB (cultural background) tests.
- Persons from backgrounds other than the culture in which the test was developed will always be penalized.
- There are enormous social class differences in a child's access to the experiences necessary to acquire the valid intellectual skills.
- IQ scores reported for African Americans and low socioeconomic groups in the United States reflect characteristics of the test rather than of the test takers.

- The poor performance of African American children on conventional tests is due to the biased content of the tests; that is, the test material is drawn from outside the African American culture.
- Women are not so good as men at mathematics only because women have not taken as much math in high school and college.

Are these criticisms valid? The investigation of this question turns out to be considerably more complicated than the reader might suppose. A most important point is that appearances can be deceiving. As we will explain subsequently, the fact that test items "look" or "feel" preferential to one race, sex, or social class does not constitute proof of test bias. Test bias is an objective, empirical question, not a matter of personal judgment.

Although critics may be loath to admit it, dispassionate and objective methods for investigating test bias do exist. One purpose of this section is to present these methods to the reader. However, an aseptic discussion of regression equations and statistical definitions of test bias would be incomplete, only half of the story. Conceptions of test bias are irretrievably intermingled with notions of test fairness. A full explanation of the story surrounding the test-bias controversy requires that we investigate the related issue of test fairness, too.

Differences in terminology abound in this area, so it is important to set forth certain fundamental distinctions before proceeding. Test bias is a technical concept amenable to impartial analysis. The most salient methods for the objective assessment of test bias are discussed in the following. In contrast, test fairness reflects social values and philosophies of test use, particularly when test use extends to selection for privilege or employment. Much of the passion that surrounds the test-bias controversy stems from a failure to distinguish test bias from test fairness. To avoid confusion, it is crucial to draw a sharp distinction between these two concepts. We include separate discussions of test bias and test fairness, beginning with an analysis of why test bias is such a controversial topic.

The Test-Bias Controversy

The test-bias controversy has its origins in the observed differences in average IQ among various racial and ethnic groups. For example, African Americans score, on average, about 15 points lower than white Americans on standardized IQ tests. This difference reduces to 7 to 12 IQ points when socioeconomic disparities are taken into account. The existence of marked racial/ethnic differences in ability test scores has fanned the fires of controversy over test bias. After all, employment opportunities, admission to college, completion of a high-school diploma, and assignment to special education classes are all governed, in part, by test results. Biased tests could perpetuate a legacy of racial discrimination. Test bias is deservedly a topic of intense scrutiny by both the public and the testing professions.

One possibility is that the observed IQ disparities indicate test bias rather than meaningful group differences. In fact, most laypersons and even some psychologists would regard the magnitude of race differences in IQ as prima facie evidence that intelligence tests are culturally biased. This is an appealing argument, but a large difference between defined subpopulations is not a sufficient basis for proving test bias. The proof of test bias must rest upon other criteria outlined in the following section.

Racial and ethnic differences are not the only foundation for the test-bias controversy. Significant gender differences also exist on some ability measures, most particularly in the area of spatial thinking (Maccoby & Jacklin, 1974: Halpern, 1986). In one study (Gregory, Alley, & Morris, 1980), males outscored females on the spatial-reasoning component of the Differential Aptitude Test by a full standard deviation. Such findings raise the possibility that spatial-reasoning tests may be biased in favor of males. But how can we know? When do test score differences between groups signify test bias? We begin by reviewing the criteria that should be used to investigate test bias of any kind, whether for race, gender, or any other defining characteristic.

Criteria of Test Bias and Test Fairness

The topic of test bias has received wide attention from measurement psychologists, test developers, journalists, test critics, legislators, and the courts. Cole and Moss (1998) underscore an unsettling consequence of the proliferation of views held on this topic, namely, concepts of test bias have become increasingly intricate and complex. Furthermore, the understanding of test bias is made difficult by the implicit and often emotional assumptions—held even by scholars—that may lead honest persons to view the same information in different ways.

In part, disagreements about test bias are perpetuated because adversaries in this debate fail to clarify essential terminology. Too often, terms such as test bias and test fairness are considered interchangeable and thrown about loosely, without definition. We propose that test bias and test fairness commonly refer to markedly different aspects of the test-bias debate. Careful examination of both concepts will provide a basis for a more reasoned discussion of this controversial topic.

As interpreted by most authorities in this field, test bias refers to objective statistical indices that examine the patterning of test scores for relevant subpopulations. Although experts might disagree about nuances, on the whole there is a consensus about the statistical criteria that indicate when a test is biased. We will expand this point later, but we can provide the reader with a brief preview here: In general, a test is deemed biased if it is differentially valid for different subgroups. For example, a test would be considered biased if the scores from appropriate subpopulations did not fall upon the same regression line for a relevant criterion.

In contrast to the narrow concept of test bias test fairness is a broad concept that recognizes importance of social values in test usage. Even test that is unbiased according to the tradition technical criteria of homogeneous regression might still be deemed unfair because of the social consequences of using it for selection decisions. The crux of the debate is this: Test bias (a statistical concept is not necessarily the same thing as test fairness (values concept). Ultimately, test fairness is based on social conceptions such as one's image of a jus' society. In the assessment of test fairness, subjective values are of overarching importance; the statistical criteria of test bias are merely ancillary. We will return to this point later when we analyze the link between social values and test fairness. But le* us begin with a traditional presentation of technical criteria for test bias.

The Technical Meaning of Test Bias: A Definition

One useful way to examine test bias is from the technical perspective of test validation. The reader will recall from an earlier chapter that a test is valid when a variety of evidence supports its utility and when inferences derived from it are appropriate, meaningful, and useful. One implication of this viewpoint is that test bias can be equated with differential validity for different groups:

> *Bias is present when a test score has meanings or implications for a relevant, definable subgroup of test takers that are different from the meanings or implications for the remainder of the test takers. Thus, bias is differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers. (Cole & Moss, 1998)*

Perhaps a concrete example will help clarify this definition. Suppose a simple word problem arithmetic test were used to measure youngsters' addition skills. The problems might be of the form "II you have two six-packs of pop, how many cans do you have altogether?" Suppose, however, the test is used in a group of primarily Spanish-speaking seventh graders. With these children, low scores might indicate a language barrier, not a problem with arithmetic skills. In contrast, for English-speaking children low scores would most likely indicate a deficit in arithmetic skills. In this example, the test has differential validity, predicting arithmetic deficits quite well for English-speaking children, but very poorly for Spanish-speaking children. According to the technical perspective of test validation, we would conclude that the test is biased.

Although the general definition of test bias refers to differential validity, in practice the particular criteria of test bias fall under three main headings: content validity, criterion-related validity, and construct validity. We will review each of these categories, discussing relevant findings along the way. The coverage is illustrative, not exhaustive. Interested readers should consult Jensen (1980), Cole and Moss (1998), and Reynolds and Brown (1984b).

Lesson 11

TESTING IN BUSINESS AND INDUSTRY

Chapter Goals and Objectives

After completing this chapter you should be able to:

- List and describe the ways tests are used in business and industry.
- Identify the types of reliability and validity important for different testing; scenarios.
- Contrast the use of tests with the use of interviews in employee selection and describe the advantages and limitations of each technique.
- Describe the use of ability tests, work sample tests, and integrity tests in the selection of employees and the issues raised by each type of test.
- Discuss the problems associated with validating selection tests.
- Describe the use of productivity measures, personnel measures, and evaluation scales in the assessment of job performance and the issues surrounding the use of each measure.
- Describe federal regulation of selection and performance testing and the impact of these regulations.
- Describe the use of human factors studies, organizational studies, and marketing studies in the evaluation of business activities.

The final scenario to consider in our discussion of test use is the world of business and industry. Many businesses and corporations employ psychologists to screen job applicants, to evaluate employees, or to assess the effectiveness of company operations. In terms of training, these psychologists are most likely to have either a master's or doctoral degree in industrial/organizational (I/O) psychology, an applied field focused on the application of psychological research to the world of work. Industrial/organizational psychologists comprise about 7% of the membership of the American Psychological Association (Slapp & Fuichcr, 1983). This chapter will focus on the ways psychologists and others use tests in the workplace and the issues surrounding the use of tests.

USES OF TESTS IN BUSINESS AND INDUSTRY

When psychological tests are used in the workplace, our primary concern is the selection or construction of tests that are reliable and valid for the task at hand. This seemingly simple statement has wide-ranging implications. For example, if a single test will be used repeatedly to evaluate employees, the test must possess adequate test-retest reliability. On the other hand, if a test with two forms will be used to screen applicants, such that some will receive one form and some will receive the other, the two forms must have alternate-form reliability. If a multidimensional test will be used, generating scores on several different subscales, the scales must be internally consistent.

In terms of validity, a primary concern is criterion validity. In many business applications, tests are used to predict the future performance of test takers. The prediction may relate to a job applicant's likelihood of success in the job (a hiring decision) or an employee's likelihood of success in a new position (a promotion decision). Construct and content validity are important when tests are used to measure variables such as employee attitudes about the workplace or consumer attitudes about products and programs. In this section, we will examine the specific ways test^ are used and the factors to consider in each scenario.

Selection of New Employees

At some point, nearly every business is faced with the task of selecting new employees. In a small business, the selection process is likely to be conducted by the owner. In larger businesses and corporations, the process typically is coordinated by the personnel officer, who may be an I/O psychologist or a psychologist hired on a contract basis. The ideal scenario is a multistage process in which the business (1) identifies the specific tasks to be performed by the new employees using the process of job analysis (see p. 49), (2) develops a description of the job based on the job analysis, (3) recruits applicants, (4) evaluates the applicants, and (5) selects the new employees based on the evaluation data (Wise, 1989).

Job applicants routinely are required to complete application forms, submit letters of reference, and be interviewed. The use of psychological tests, such as ability tests, is more variable. Research indicates a recent decline in the use of testing for applicant screening. In a survey of members of the American Society of Personnel Administrators (Tenopyr, 1981), approximately 75% of the respondents indicated that they did less employee testing than they had done 5 years earlier. Only 60% of companies with 25,000 or more employees used at least one psychological test, and only 39% of companies with fewer than 100 employees used a psychological test. The decline has been attributed to the increased regulation of test use by professional societies, lawmakers, and the courts and to business' general distrust of paper-and-pencil tests.

Evaluation of Current Employees

"Evaluation of job performance by current employees serves several important functions. Assessment of job performance is a necessary component of decisions about raises and promotions. Il also can provide feedback lo employees, communicating information about employer expectations and employee performance relative lo these expectations. In addition, job performance evaluations may provide valuable input lo the process of job analysis (Wise, 1989). For example, the results of an employee evaluation may lead a business to revise its definition of the tasks comprising a given job. An evaluation might indicate that the job is poorly defined or includes too many tasks, making it difficult for employees to be successful.

Performance evaluation may be a formalized, ongoing process that occurs at periodic intervals or a more informal, occasional event. Many businesses evaluate their employees on a regular basis and compile the results of these performance evaluations in each employee's personnel file. Regardless of the procedure and mechanism used, all individuals can be expected to be evaluated at some point during their employment. In its ideal form, performance evaluation centers around the tasks identified in lire job analysis written when applicants were recruited. Measures are selected or constructed for each job task and are considered together for decisions about the adequacy of job performance. Performance evaluations typically focus on rating scales.

Evaluation of Programs and Products

Psychological tests are also used lo evaluate aspects of the workplace itself or lo evaluate the goods and services provided by a business. Evaluation of the workplace involves two separate issues. Human [actors research assesses the impact of the work environment on employee behavior and seeks to identify ways to improve employee performance by modifying that environment (Mc-Cormick & Sanders, 1982). For example, psychological tests could be used to determine how the redesign of a console affects an individual's ability to operate a machine or how workers are affected by variables such as lighting, noise, and temperature. Because we know that behavior, such as a worker's level of productivity, is influenced by altitudes, human factors research also examines how changes lo the environment affect employee satisfaction. Human factors research, therefore, could investigate how the addition of an employee lounge on each floor affects employee's morale. Human factors studies may use observational techniques, experimental manipulations incorporating skills tests, and altitude scales.

Evaluation of the workplace also involves organizational research. The employees of a business or industry constitute a group of people working within a hierarchical structure. Their attitudes and behavior are affected by both the characteristics of that structure and the dynamics that emerge as employees interact with each other (Sicgcl & Lane. 1982). In organizational research, we examine issues such as supervisor-worker relationships, opportunities for workers to be involved in business decisions, and conflict resolution processes. The goal is to identify aspects of organizational structure and dynamics that contribute to productivity and satisfaction and organizational elements that need to be changed. Organizational studies may involve observational research, interviews, and the administration of attitude scales.

In contrast to research evaluating the characteristics of the workplace, marketing research is used to determine the success or likely future success of a product or service. Rather than focusing on workers, marketing research targets consumers, usually through attitude scales and product field testing (Schiffman & Kanuck, 1983). Data generated through marketing studies may be used to revise the design of a product or the delivery of a service or as the basis for later advertising campaigns.

PROCEDURES FOR SELECTING NEW EMPLOYEES

Although the use of testing in employee selection has declined, it is estimated that about two-thirds of companies in the United States use some form of written testing the selection process (Friedman & Williams, 1982). The administration of a test, however, does not necessarily indicate its importance in the decision process. In fact, when tests are used, their role in the process is extremely variable (Tenopyr. 1981). Written tests are more common and more emphasized_ for office positions and positions involving specific skills than for production and sales jobs. Applicants for clerical positions represent the most heavily tested group, followed by applicants for skilled positions such as electrician or mechanic. In contrast, screening for management positions relies heavily on credentials, interviews, and letters of reference. The use of testing also varies as a function of employer. Public sector jobs, such as civil service jobs, jobs with state, county, and local government, and jobs within the military, routinely require psychological testing (Friedman & Williams, 1982).

Selection Testing versus Employment Interviews

Psychological testing and employment interviews are common but very different approaches to the selection of new employees. It is useful, therefore, to reflect on the relative merits of each technique. The popularity of interview reflects its simplicity, its flexibility, and its ability to provide information not available through testing. It is unfortunate, however, that the current trend is for companies to rely more on interview and less on testing. Testing is an objective assessment procedure using standardized content, standardized administration and. standardized scoring. Tests are an efficient and effective way to obtain the same information, under the same conditions, about all applicants for a job. Good published tests are available for employment screening with reliability coefficients at or above .8 and validity coefficients at or above .6. On the other hand, interview is a subjective technique in which both content and administration may vary. In fact, this flexibility was just cited as one of the strengths of the interview process. It is however, a two-edged sword. Research on employment interviews identifies some serious reliability and validity problems that occur most often when interviewers do vary the content and administration of questions. As a result, most psychologists recommend the use of structured employment interviews.

In structured interviews, a predetermined set of questions is posed to all applicants in a specific order. Unstructured interviews are "more like clinical interviews using broad, open ended questions whose content and sequence is determined largely by the applicant’s earlier answers. One obvious difference Is the nature of the information gathered. Structured interviews lead to the collection of comparable data on all applicants that facilitates the process of evaluating individual differences. Structured interviews are also more likely to lead to potential agreement among different interviewers, a measure of interview reliability, and accurate predictions about future performance, a measure of interview validity (Wiesner & Cronshaw, 1988). Research indicates that both the reliability and validity of structured interviews are twice that of unstructured interviews (Harris, 1989).

Even when a structured format is used, interviews are open to a variety of sources of potential bias that threaten their reliability and validity. People tend lo form first impressions rapidly in interpersonal encounters, and first impressions exert a powerful influence over our processing of later information (e.g., Lyman, Hatlclid, & MacCundy, 1981). We tend to focus more on information that confirms our impression and lo discount information that runs counter to it. Interviewers often form a quick first impression based on an obvious or outstanding characteristic of an applicant and may compound the problem by making inappropriate inferences from those features. For example, an applicant's personal appearance may be used as the basis for inferences about level of intelligence (Gilmore, Bechr, & Love, 1986). In a similar way, interviewers may make quick judgments about a person's characteristics on the basis of gender, ethnic group, social class, or cultural heritage (Sattler, 1993). Finally, interviews are particularly susceptible to the "halo effect," in which a single, general impression leads an interviewer to form a favorable or unfavorable attitude early in the interview process (Cooper, 1981).

Although interview is prone to the classic problems of subjective assessment techniques, it can be a valuable source of information during the screening process. When interviewers are trained to focus on evaluation of specific applicant characteristics, the reliability and validity of judgments improve (Dougherty. Ebert. & Callender. 1986). Furthermore, when interviewers use a structured format, validity coefficients often are close to .7 (Harris, 1989).

On the other hand, research comparing the use of standardized tests, interviews, biographical information, letters of reference, and work samples clearly identifies standardized tests as a superior technique (e.g., Reilly & Qiao, 1982). According to a report by the National. Academy of Sciences (Wigdor & Garner 1982), although each technique has merits, none of the alternatives to standardized tests is as informative, fair, and psychometrically.

Types of Tests

The tests used to select employees may be written or performance based. The two types represent a distinction between the use of signs and the use of samples (Wernimont & Campbell, 1968). An applicant's score on a clerical test is viewed as a sign or indicator of potential success in the job. The inference is based on theoretical and research support for a relationship between the abilities tested and the tasks comprising the job. However, an applicant's performance on a series of clerical tasks provides an actual sample of job-relevant behaviors. As the old maxim in psychology states, "Nothing predicts behavior like behavior." Research on performance-based or work .sample tests suggests that they are often better predictors than written ability tests (e.g., Reilly & Chao, 1982).

Integrity tests address a separate issue in employee selection. Integrity tests are designed to predict a candidate's likely level of honesty and trustworthiness. Over the past few decades, employers have become increasingly concerned with these issues. In some cases, employees may be placed in positions where they have access to sensitive or confidential information. Employers naturally want to ensure that the individuals selected for these jobs can be trusted. In other cases, the concern is based more on economics. Businesses have become increasingly concerned about protecting themselves from all types of employee theft, ranging from pilfering of office supplies to embezzlement. As the use of integrity tests has increased, questions have been raised about the validity of these tests as predictors of employee behavior patterns.

Ability Tests. The ability tests used in employee selection fall into several categories: general ability or intelligence tests, aptitude test batteries, and tests of specific aptitudes. Research on popular tests in each category indicates that employers are most likely to use tests of specific aptitudes and that these tests are most likely to make accurate and useful predictions about an applicant's potential job performance.

General Ability Tests. The general ability or intelligence tests used tend to be paper-and-pencil, relatively short, group tests such as the Wonderlic Personnel Test. The Wonderlic is a 50-item multiple-choice test of mental ability normed on a sample of 50,000 individuals 20 to 65 years old. The Wonderlic includes verbal, mathematical, pictorial, and analytic items and is available in five different forms. Alternate form and split-half reliabilities typically exceed .9, and scores correlate with performance in a variety of jobs (Dodrill & Warner, 1988; Murphy, 1984a).

Multiple-aptitude Batteries. Three popular multiple-aptitude batteries used in employment testing are the Armed Services Vocational Aptitude Battery (ASVAB), the General Aptitude Test Battery (GATB), and the Differential Aptitude Tests (DAT). Designed for the Department of Defense, the ASVAB is used in both educational and military settings. The 10 subtests are grouped into three academic composites—academic ability, verbal, and math—and four occupational composites: mechanical and crafts, business and clerical, electronics and electrical, and health, social, and technology. Data from the 1980 revision indicate that the internal consistency coefficients for composite scores average close to .9 and that the test is a valid predictor of performance during job training.

The problem with the ASVAB lies in the relationship between the composite scores. Since each composite score is designed to measure a different type of aptitude, we would expect lo find very small correlations between the composite scores. In fact, the average correlation between composite scores is .86. The high correlations reflect the fact that the same subtest may be part of several different composite scores. For example, the score on the Arithmetic Reasoning subtest is used in computation of five of (he seven composite scores. This parallels the problem discussed in Chapter 10 (sec p. 348): The scoring of items on more than one subscale leads lo the creation of scales that are not independent. The high correlation between composite scores makes it difficult to view these as measures of distinct aptitudes. The ASVAB composites are more like multiple measures of overall level of aptitude (Murphy, 1984b).

It is possible, however, to develop batteries that do .measure multiple aptitudes. The General Aptitude Test Battery or GATB includes 12 separate tests that yield a score on general mental ability and scores on eight different factors: verbal, numerical, and spatial aptitudes, form perception, clerical perception, motor coordination, manual dexterity, and finger dexterity. In contrast lo (he ASVAB, the average intercorrelalion between these nine scores is only .24. Furthermore, the correlations between the motor tests, like finger dexterity, and the intellectual tests, like verbal aptitude, are close to 0. We can explain this in part by noting the use of the word factors to describe the subscales at the beginning of the paragraph. The GATB scales were developed through a factor analytic procedure. As discussed in Chapter 4 (see p. 104), factor analysis is designed to generate a set of subscales that are as independent (uncorrelated) as possible.

The GATB has been used routinely by the Department of Labor and state employment services. However, on July 10, 1990, the Department of Labor (DOL) announced that it would discontinue using the GATB because it appeared to be a discriminatory test. African-American and Hispanic-American test takers typically score lower on the test than nonminority individuals, a possible indication of differential validity. Because of the group score differences, the DOL had been using separate scoring criteria for minority and nonminority groups, a procedure endorsed in a 1989 report by the National Academy of Sciences reviewing research on the GATB (APA, 1990a). Norm-referenced scores for minority group members were generated by comparing the performance of minority applicants to the performance of others within their specific ethnic groups. However, the Academy report did raise questions about the merits of these score adjustments, and provisions within the pending Civil Rights Act, subsequently passed in 1991, prohibited the use of score adjustments. The Department of Labor felt that it had no alternative but to suspend use of the GATB until the test could be revised and thoroughly researched.

Both the ASVAB and GATB are hindered by psychometric problems—the ASVAB in the independence of composite scores, the GATB in the prediction of performance for minority group members. A third test, the Differential Aptitude Tests (DAT), is a possible alternative for applicant screening. As discussed in Chapter 11 (see p. 394), the DAT was designed for educational and vocational counseling with high school students, and includes eight tests designed to tap aptitudes that are relevant to academic or occupational choices. However, the DAT has been criticized as a tool for employee selection because few studies are available on the relationship between DAT scores and measures of actual job performance (Bennett, Seashore, & Wesman, 1982). Until additional criterion validity studies are conducted, it is unlikely that the DAT will fill the gap created by reduced use of the other two tests.

Tests of Specific Aptitudes. Unlike the multiple-aptitude batteries, tests for specific aptitudes are generally valid and useful predictors of future job performance. This should not be surprising since each test focuses on specific skill areas whose direct relevance to a job can easily be determined. Since we identified clerical and skilled technical jobs as the two cases in which written tests are likely to be used, we will illustrate this category using the Minnesota Clerical Test and the Bennett Mechanical Comprehension Test.

The Minnesota Clerical Test (MCT) is designed to measure perceptual speed and accuracy. It falls into the category of speed tests described in Chapter 2 (see p. 34). All items are short and straightforward, and individual differences are reflected in the number of items answered correctly within a strict time limit. Its two subtests, number comparison and name comparison, are composed of pairs of numbers or names that must be identified as the same or different. In both subtests, the "different" items vary in only a minor detail such as a single digit or a single letter in the spelling of a name. Test-retest reliability usually is at least .7, and scores correlate well with later ratings of job performance by supervisors. However, the source of these criterion validity relationships is uncertain. The test may be a good predictor because it is tapping a cognitive difference in speed of mental operations, not because it presents tasks similar lo what clerical workers actually do (Murphy & Davidshofer, 1988).

The Bennett Mechanical Comprehension Test is designed lo measure mechanical knowledge and mechanical reasoning. Using pictorial multiple-choice items varying in level of difficulty, test takers are required to answer questions about the operation of machines, tools, and vehicles and lo solve problems by applying the principles of physics and mechanics. In addition to generating internal consistency coefficients in the .8 lo .9 range, the test is an excellent predictor both of performance in job training courses and later job performance (Ghiselli, 1966).

Work Sample Tests

The use of work samples in employee selection is a long-standing tradition. For example, an advertising company looking for a graphic designer invariably requires each candidate to bring a portfolio of previous work. Work sample tests represent an effort to standardize the collection of work samples. All applicants are required to perform the same tasks under the same conditions, facilitating the collection of comparable data and the assessment of individual differences in ability. Because the work samples are obtained during application for a job, it is reasonable to assume that candidates are motivated lo do their best. Work sample tests, therefore, should be viewed as tests of maximal performance.

Work sample tests range from performance of simple tasks to performance in complex scenarios. A candidate for a typing position might be required to type a 400-word letter, whereas a candidate for a pilot's job might be required to navigate a route via a flight simulator. The key to designing or selecting a valid work sample test is to select tasks that are actual elements of the job itself. This requires writing a detailed job analysis that identifies the components of the job and selecting a representative sample of these tasks for the work sample test. In essence, the criterion validity of these tests is a function of their content validity (Asher & Sciarrino. 1974). If the tasks measured in the work sample test are a representative sample of the job tasks (content validity), performance on the work sample test is likely to predict future job performance (criterion validity).

It is easy to sec how we could develop a work sample test for a skill-based job. It may be more difficult to imagine designing a work sample job for a managerial position. In fad, samples of management behaviors can be obtained through a variety of techniques and are useful predictors of performance as a manager (Cascio. 1982). A popular technique is the in-basket test (Frederickson, 1961), a simulation task in which each candidate is asked to respond to a collection of memos, letters, notes, and other materials in a manager's in-basket. Each applicant must actually do whatever is necessary to handle the tasks defined by the contents of the basket—write letters and memos, set up an agenda of meetings and the like. Responses can be scored in terms of the priorities each applicant sets, based on the order of handling tasks, responses to critical incidents or issues, and overall effectiveness in accomplishing the designated tasks.

Integrity Tests

Employers understandably are concerned about the honesty and trustworthiness of potential employees. Until recently, these concerns could only be addressed through on-the-job surveillance, subjective judgments, or use of a polygraph (lie detector) test. Video surveillance is an expensive procedure and can offend employees and produce a hostile work environment. Lesser forms of surveillance, such as requiring supervisor approval for purchases paid by check, are less effective and often annoying to customers.

Subjective' judgments based on application forms, letters of reference, or interviews typically $re neither reliable nor valid. The problems with these approaches have led many employers to turn to more '"scientific" techniques, such as polygraph or integrity tests.

A polygraph test involves comparison of physiological responses, such as heart rate and respiration, when applicants answer control questions and questions of integrity. For example, an applicant's physiological responses to questions about name, age, and address (control questions) could be compared to questions about stealing, lying, and cheating (integrity questions). Polygraph tests are extremely controversial and in fact are banned in several states.

The problem with polygraph tests is their level of accuracy in identifying instances of lying. Empirical studies of polygraph data indicate that the accuracy of judgments based on physiological responses frequently is at the chance level (e.g. Lykken, 1979). Furthermore, polygraph tests generate a relatively high proportion of false positives—people identified as lying when they are telling the truth (Ben-Shakar, Lieblich, & Bar-Hillel, 1982). Concern about the unreliability of polygraph tests led to passage of the Employee Polygraph Protection Act in 1988, which prohibits private employers from requiring or requesting polygraph exams. Security firms and firms that manufacture controlled substances are exempted from the prohibition. Although a subsequent polygraph law in Massachusetts included written exams in their definition of "lie detector tests," the federal statute does not ban the use of oral or written tests (Sackelt, Burris, & Callahan, 1989).
Contrary to what you may suspect, integrity tests appear to be both a reliable and valid approach to the issue of applicant honesty (e.g., APA, 1991a). Integrity tests can be grouped into two broad categories: overt integrity tests and personality-oriented measures (Sackelt & Harris. 1984). The overt integrity tests are designed .specifically to measure issues relevant to integrity. They typically have sections devoted to attitudes about theft and dishonesty and sections soliciting admissions about theft and dishonesty. Examples of overt integrity tests include the Reid Report (Brooks & Arnold, 1989), the Stanton Survey (Harris &. Gentry, 1992), and the London I louse Personnel Selection Inventory (London House, 1991).

The personality-oriented tests are not designed specifically to measure honesty. Instead, they focus on a variety of counterproductive tendencies, including such constructs as dependability, conscientiousness, and nonconformity. Examples of this type include the Hogan Personnel Selection Series (Hogan & Hogan, 1985), the London House Employment Productivity Index (Terris, 1986), and the PDI Employment Inventory (Paajanen, 1986). Although these tests may be used by employers who are concerned about theft, our discussion will focus on the more overt tests of integrity.

Internal consistency coefficients for overt integrity tests typically are .85 or higher, with test-retest coefficients ranging from the .60s to the .90s (e.g., Sackelt, Burris, & Callahan, 1989). The most compelling data in favor of integrity tests is their correlations with criterion measures such as on-the-job theft and disciplinary action. This is particularly important because of concerns about response bias in integrity testing, h is obviously in an applicants' best interests lo present themselves in a favorable light. In fact, scores on integrity tests do correlate with measures of social desirability, such as the MMPI validity scales and the Edwards Social Desirability Scale (Sackelt, Burris, & Callahan. 1989). However, the correlations between integrity test scores and criterion measures are based on studies of groups of people. Although they indicate that the tests are good predictors in general, their validity for predicting the behavior of an individual applicant is less clear.

Regulation of Selection Procedures

Employee selection procedures are regulated in a variety of ways: by legislation, by court decisions, and by the rules set down by professional societies. Chapter 1 described activities by the American Psychological Association, the American Education Research Association, and state licensing boards (see p. 24). In this section, we will focus on regulation at the federal level.

The Equal Employment Opportunity Commission (EEOC), created by Title VII of the 1964 Civil Rights Act, developed guidelines defining fair employee selection procedures in 1970 that were revised and published in 1978 (EEOC, 1978). These guidelines are used by all employers in the public sector and can also be imposed on private businesses that receive government funds. Furthermore, many private businesses voluntarily agree to follow these guidelines as a gesture of good faith in hiring practices.

The guidelines state explicitly that procedures used to screen potential employees must be valid. "Procedures" is broadly defined to include interviews, psychological tests, and even tests using physical standards such as height, weight, or strength (Hogan & Quigley, 1986). The "validity" to be demonstrated may be criterion or content validity, and the guidelines even define the procedures to be used to demonstrate validity. For example, demonstration of criterion validity requires an initial job analysis, the selection of a representative sample, the selection of criterion measures and the presence of a statistically significant (at the .05 level) predictor-criterion, relationship (EEOC, 1978). Recent reviews of federal court cases in which the validity of screening tests has been challenged indicate that the courts take these guidelines seriously. The screening tests successfully defended have been written tests focusing on specific job-related tasks for which there are several different validity studies using large samples of people (e.g., Thompson & Thompson, 1982).

As an outgrowth of the Civil Rights Act, the basic intent of the EEOC guidelines is to prohibit discrimination in hiring practices. A major focus is the identification of selection procedures that might have "adverse impact" on a particular racial, ethnic, religious, or gender group. The guidelines set a statistical criterion for defining adverse impact, known as the rule of four-fifths (EEOC, 1978). Stated simply, the rule identifies a procedure as discriminatory if the selection rate for any racial, ethnic, or gender group is less than four-fifths (80%) of the highest rate of selection for any other group. For example, assume that both white and African-American individuals apply for a job. Applicants in both groups are given a screening test with a specific cutoff score to be used for hiring decisions. According to EEOC guidelines, the selection rate using this cutoff score must be determined separately for each group, and the rates compared according to the four-fifths rule. The lower selection rate must be at least 80% of the higher rate for the procedure to be fair. If the selection rate for whites is 60%, the selection rate for African-Americans must be at least 80% of the white rate or 48% (80% of .60. or .80 X .60). If the selection rate for African-Americans is only 40%, the procedure has adverse impact and is discriminatory. If an African-American who was not hired challenged the decision, the employer would be required to demonstrate that there are extenuating circumstances preventing adherence to the four-fifths rule.

Although EEOC guidelines place the burden of proof on employers, a recent Supreme Court action (Wards Cove Packing Company v. Antonio, 1989) shifted that burden to employees. Minority employees of Wards Cove Packing, primarily nonskilled Eskimos and Filipinos, charged that the company's selection procedure was biased against them, preventing them from obtaining more desirable, higher-paying jobs. After losing in the lower courts, the employees brought their case to the Supreme Court. The Court refused to hear the case, noting that the employees had not demonstrated the procedure to be invalid. In response to this decision, Congress included a provision in the 1991 Civil Rights Acts that returned the burden of proof in selection testing to employers.

Although we have already discussed many of the legislative and court actions relative to employee testing, we have not addressed the issue of affirmative action in the workplace. Because tests are often used for hiring decisions, affirmative action decisions affect the way tests can be used in these areas. One aspect of affirmative action is aggressive recruiting of minority applicants. The four-fifths rule contained in the EEOC guidelines (EEOC, 1978) indirectly creates some problems for aggressive recruiting. A company that successfully recruits a large number of minority applicants may end up hiring a smaller percentage of minority workers than a company that begins with a small minority pool. The problem arises because the number of jobs available is independent of the size of the pool, but the percentage of people hired changes to reflect the size of the pool. Because EEOC guidelines require the percentage of minorities selected to reach a specific level relative to the hiring of other groups, the guidelines can discourage a company from working to develop a large minority pool. The EEOC recognizes this problem and can authorize exceptions to the four-fifths rule in specific cases.

A second aspect of affirmative action is the establishment of hiring goals or quotas for minority workers. In a 1985 case, U.S. v. City of Buffalo, the Supreme Court addressed the use of hiring goals for minority groups who were underrepresented at a workplace because of a previously used discriminatory selection procedure. Hiring goals were approved as a temporary measure while the employer developed a new, valid selection procedure.

A third aspect of affirmative action is the use of different selection criteria for different racial, ethnic, or gender groups. Although this issue is thought of most often in relationship to educational decisions, it is also an clement of hiring decisions. One reason for using different cutoff scores might be the presence of differential validity data (see p. 421) indicating that the predictor scores linked to successful job performance vary across groups. On the other hand, companies may decide to use different cutoff scores for minority and majority groups to increase the number of minority workers hired.

In the past, it has been permissible to use different cutoff scores to increase minority presence as a redress for past discrimination. Furthermore, the EEOC focus on valid selection procedures made it necessary to use different cutoff scores if differential validity in fact existed. This was the basis for the Department of Labor's use of separate scoring procedures for different minority groups on the GATB. Both of these actions, however, appear to be illegal under the 1991 Civil Rights Act. Although the law reaffirms a commitment to affirmative action. Section 9 on the use of test scores states that it is illegal to use different cutoff scores on the basis of race, ethnic group, religion, or gender. It further identifies adjusting or altering scores on the basis of group membership, as illegal. Although the provision is designed to prevent discrimination against minority group members, it may in some cases reduce the number of minority individuals hired by preventing score adjustments that favor minority groups. It may also reduce the use of tests in employment decisions, as seen in the Department of Labor's decision to discontinue using the GATB.

Conducting Validity Studies

To be both psychometrically sound and legal, the tests used in employee selection must be valid for the selection process. Since the focus is predicting likelihood of future success, the test's criterion validity is a key element. Most criterion validity studies for business and industry use a concurrent validity design (sec p. 225). In the concurrent design, we test a representative sample of current employees on the test under consideration and correlate their scores on the test with a separate measure of job performance. A predictive validity design would dictate that all applicants be tested and hired—a clearly impractical event—with data on job performance to be collected at a later time.

The use of concurrent designs, however, presents a different but no less important problem. For a validity study to be useful, we need to compare people with a variety of scores on both the predictor test and the criterion measure. Since the criterion measure is job performance, we may not generate a wide range of criterion scores by testing only people who are currently employed— people who perform poorly at the job probably either have quit or been fired. If our goal is to calculate a criterion validity coefficient, this procedure may lead to restriction of variability on the criterion measure, a factor that can statistically limit the size of the correlation and jeopardize our validity study. If our goal is to conduct a selection efficiency study, we may not be able to identify a group of people who (1) did well on the test but (2) did poorly on the job. The lack of these false positives similarly jeopardizes the results of our analysis.

The increasing use of work sample tests reflects concern over the concurrent validity design problem. In a work sample test, all applicants are evaluated based on a job-relevant sample of behavior. Then, we can compare their work samples to the work samples generated by current successful employees. In essence, we are creating a criterion-referenced procedure using the performance of current successful employees as the standard to which applicants are compared, if samples of current work correlate well with future job performance, the work sample test are a valid technique for candidate screening. Work samples are most likely to predict future performance accurately in skill-based jobs (Asher & Sciarrino, 1974; Reilly & Chao. 1982).

PROCEDURES FOR EVALUATING CURRENT EMPLOYEES

Evaluation of job performance is important in decisions about raises and promotions and can also provide feedback to employees and provide input to the process of job analysis (Wise, 1989). Performance evaluation may be a formal, ongoing process or a more informal, occasional event. Ideally, employees are evaluated on the tasks identified in the job analysis written during recruitment, using specific measures for each job task.

Evaluation Procedures

Theoretically, three types of information could be used to evaluate the performance of an employee (Murphy & Davidshofer, 1988). In reality, the usefulness of each measure varies according to the nature of the job. A productivity measure is a direct measure of worker accomplishment. In its simplest form, productivity can be defined as the total of the number of limes a particular task is performed. The postal service could count the number of letters sorted by each postal worker within a specific period of lime, whereas an automobile plant could count the number of parts installed by each worker on an assembly line. A personnel measure focuses on information recorded by supervisors or managers in employee personnel records. Personnel measures include absenteeism, lateness, and days off due to illness or accidents. A judgmental measure is a scale used by peers or a supervisor to evaluate employee performance. Evaluation scales can be designed to assess the performance of a single employee or a group of employees and use either a rating or ranking procedure.

Productivity Measures

Although productivity measures seem most appropriate for industry, they can be applied to a variety of businesses. For example, a telephone company supervisor could total the number of calls an operator fields in a given day. A real estate company could tally the total dollars worth of transactions completed by each salesperson in the preceding year. However, there are three problems with the use of productivity measures (e.g., Guion, 1965).

First, a productivity measure requires that we define a unit of behavior to be counted. Although we can count calls answered or dollars generated by sales, it is difficult to define countable units for managerial and professional positions. Second, the item we count must be a valid and useful measure of job performance. Imagine a hospital administrator who uses a count of the number of operations performed to evaluate the surgical staff! The number of units produced is not necessarily a good measure of an employee's effectiveness.

Third, a productivity measure requires that we select an item on which employees are likely to vary and an item on which variations are directly attributable to employee behavior. If the measure is to be useful, it must discriminate between employees doing well on the job and employees who are not. In other words, it must provide individual difference information. In some jobs, employees cannot vary much on count measures because their activities are constrained by other variables. The number of calls a telephone operator can answer is determined by the capacity of the machinery that feeds the calls to the operator's station. Other times, the variations between employees in productivity measures are due to factors other than employee competence. A real estate agent may sell fewer houses because of problems with traffic, road construction, or clients who want to see many houses and spend a lot of time at each property.

Personnel Measures

Personnel measures are used frequently in performance evaluations, but present serious problems to objective measurement of job performance. Personnel measures often show low levels of test-retest reliability (e.g., Hammer & Landau, 1981). The number of times an employee is late or absent is likely to vary considerably depending on the time when the measure is taken and the length of the interval over which data are collected. As we have said repeatedly, an unreliable measure cannot be a valid or useful way to assess individual differences in performance.

Part of the problem with personnel measures is (the operational definition of the characteristics to be measured. Absenteeism is a classic example. Recent reviews indicate as many as 40 different indexes of absenteeism, including number of absences, frequency of absences, liming of absences, voluntary absences, and involuntary (illness) absences (e.g., Landy & Farr, 1983). Furthermore, most employees differ little on dimensions such as absenteeism and lateness. When there is little variability on a measure, it is not surprising that it generates low reliability coefficients.

Evaluation Scales

Perhaps because of the problems characterizing the other two measures, evaluation scales have become the most popular technique for evaluation of job performance. Evaluation scales may require a ranking or rating process; rating is the most frequently used format. Regardless of scale format, the most useful data are generated when employees are evaluated using a separate scale for each task comprising the job and for each important employee characteristic (e.g., communication skills).

Ranking Scales. Ranking scales are comparative scales (see p. 122) that evaluate the members of a group by comparing each person's performance to the performance of the other group members. The simplest type is a full ranking scale in which the employees being evaluated are listed in order from "best" lo "worst." A series of full ranking scales could be used in which each scale represents a different job performance dimension. Employees could be ranked one time on productivity, another time on initiative, and a third lime on interpersonal skills. The process is illustrated in Table 11.1.
Full ranking may require the evaluator lo make very fine discriminations between people and therefore can result in arbitrary and unreliable classifications. Furthermore, full ranking can be difficult to use when evaluating a large group—imagine ranking a set of 20 shift workers in terms of job performance!

A simple alternative lo full ranking is w forced distribution scale. A forced distribution scale presents a series of categories and requires the evaluator lo identify a fixed number of people in each. For example, a forced distribution scale could require identification of employees in the top, middle, or bottom thirds on a performance dimension. By using a series of forced distribution scales, employers can identify people who are in the upper, middle, and lower groups on a variety of job-relevant dimensions. Table 11.2 presents a sample forced distribution ranking task.

Forced distribution scales do not require evaluators to make fine discriminations among people and are particularly useful with large groups. In testing in Business and Industry

Table 11.1 Sample Full Ranking Task

Instructions: Please list the 6 employees on your shift in order from best (l) to worst (6) on each dimension listed.

Productivity	Initiative	Dependability
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	6

Table 11.2 Sample Forced Distribution Ranking Task

Instructions: For each dimension listed, please identify the 2 employees on your shift who fall into each category.

Category l = top third of your shift group
Category 2 - middle third of your shift group
Category 3 = bottom third of your shift group

Employee	Productivity	Initiative	Dependability
Bill Murray	_________	_________	_________
Robin Williams	_________	_________	_________
Jerry Seinfeld	_________	_________	________
Roseanne Arnold	_________	_________	_________
John Belushi	_________	_________	_________
Gilda Radner	_________	_________	_________

addition, forced distribution scales also address the leniency and severity problems common in performance evaluation (see p. 123) by requiring the people being rated to be distributed across all rating categories.

A paired comparison scale presents the names of people to be evaluated as pairs, requiring the evaluator to identify the person performing best within each pair. A series of items can be generated so that the various pairs are evaluated separately for different elements of the job. The advantage of pair comparison scales is the simplicity of the judgment process. Each pair requires

Table 11.3 Sample Paired Comparison Task

Instructions: For each pair of employees on your shift, circle the name of the employee whose performance is better.

Which employee is more *productive*?

Murray or Williams
Murray or Seinfeld
Murray or Arnold
Murray or Belushi
Murray or Radner
Williams or Seinfeld
Williams or Arnold
Williams or Belushi
Williams or Radner
Seinfeld or Arnold
Seinfeld or Belushi
Seinfeld or Radner
Arnold or Belushi
Arnold or Radner
Belushi or Radner

Which employee shows more *initiative*?

Murray or Williams
Murray or Seinfeld
Murray or Arnold
Murray or Belushi
Williams or Radner
Seinfeld or Arnold
Seinfeld or Belushi
Seinfeld or Radner

Murray or Radner
Williams or Seinfeld
Williams or Arnold
Williams or Belushi
Arnold or Belushi
Arnold or Radner
Belushi or Radner

Which employee is more *dependable*?

Murray or Williams
Murray or Seinfeld
Murray or Arnold
Murray or Belushi
Murray or Radner
Williams or Seinfeld
Williams or Arnold
Williams or Belushi
Williams or Radner
Seinfeld or Arnold
Seinfeld or Belushi
Seinfeld or Radner
Arnold or Belushi
Arnold or Radner
Belushi or Radner

only comparison of the relative performance of two people, producing a highly reliable measure. A paired comparison procedure is illustrated in Table 11.3.

The problem, however, is that the total number of pairs to be compared is N(N - l)/2, in which N equals the number of people being evaluated. For a group of 5 employees, a relatively small group, the number of pairs to be evaluated is 5(5 - l)/2 or 10 pairs. But as the size of the group increases, the number of comparisons jumps radically. For a group of 10 people, the number of comparisons is 45!

The trade-off in choice of ranking scales is between precision of measurement and decision-making effort. Forced distribution scales are easy to use, but do not provide very precise information about individual differences. Paired comparison scales may require a lot of time and effort

Rating Scales. Unlike ranking scales, rating scales are standard scales that represent the performance of each employee independently (see p. 118). One reason for their popularity is that rating scales can be used with either a single employee or a group of employees. Continuous scales represent a performance dimension like "productivity" with a line, a set of boxes, or a series of numbers "'(e.g., 1 to 5). The end points are typically anchored with terms such as "excellent" and "poor," and sometimes other points along the dimension are identified as "average," "above average," or "below average." The evaluator marks the line, boxes, or numbers to indicate where an employee falls on that particular dimension. A set of continuous scales is presented in Table 11.4.

Continuous scales appear simple to use, but they often produce unreliable judgments or ratings with low validity. The problem is operationally defining the dimension being rated and the meaning of the anchor points. Because continuous scales traditionally use general terms like "initiative" (dimension) and "high" (anchor), ratings are influenced by the rater's interpretation of the terms used (Murphy & Davidshofer, 1988).

The alternative to continuous scales is behavioral scales, such as behaviorally anchored rating scales (BARS) and behavioral observation scales (BOS). In both cases, employee performance is compared to actual descriptions of job-relevant behaviors. Even characteristics like "communication skills" can be rated using behavioral statements such as "The supervisor explains each job requirement to new employees during their first day of work."

The difference between the BARS and BOS approach is the type of rating to be made. As illustrated in Table 11.5, each item in a BARS system presents the rater with a series of statements describing poor to superior performance. The rater must identify the particular statement from the set that

Table 11.4 Set of Continuous Rating Scales

Instructions: Please mark the box indicating the employee's performance on each dimension listed.

Employee: Bill Murray

Productivity: □□□□□□□

low high

Initiative: □□□□□□□

low high

Dependability: □□□□□□□

low high

best characterizes the employee's behavior. The sample BOS system in Table 11.6 presents a series of statements describing different levels of performance. The rater must indicate how often each behavior occurs. The scale may use numbers to represent patterns such as "never." "a few times," "half the time," "often," and "all of the time" or may require an actual count of the number of times the behavior has occurred.

Although behavioral scales are designed lo reduce the ambiguity of the rating task, research docs not indicate that they produce better or even different evaluations than continuous scales (e.g., Landy & Farr, I9K0). Given the time and effort needed to construct behavioral scales, most performance evaluation is conducting using simple graphic or numerical scales.

Issues in the Use of Scales. Inevitably, the ranking or rating of an employee's performance is a subjective process, and rater's judgments are prone to I he-same errors and biases regardless of scale format. Furthermore, it is difficult to determine the validity of these job performance measures. Although we can study their reliability by requiring raters to evaluate employees repeatedly, studies of the validity of these measures requires that we correlate ranking/rating scores with other measures of job performance. Because productivity measures and personnel data are likely lo focus on other aspects of job performance, we are left with the task of comparing one ranking/rating system to another.

Table 11.5 Hypothetical Behaviorally Anchored Rating Scale

Instructions: Please mark the statement that best describes your employee.

Employee: Bill Murray

Dependability	**Points**
Attends all meetings and completes assigned work on time.	5
Attends meetings and completes work although is occasionally late.	4
Attends meetings and completes work but is late about half the time.	3
Rarely attends meetings and frequently is late in completing work.	2
Rarely attends meetings or completes assigned work.	1

Table 11.6 Hypothetical Behavioral Observation Scale

Instructions: Rate how often each of the listed behaviors has occurred during the past month using the following scale:

5 = all the time
4 = most of the lime
3 = about half the lime
2 = a few times
1 = never

Employee: Bill Murray

Event	Frequency
Arrives on time for meetings.	________
Forgets to complete assigned work.	________
Arrives late for a meeting.	________
Turns assigned work in early.	________
Fails to attend meetings.	________
Turns assigned work in on lime.	________

Some researchers suggest that the validity of evaluation scales can be determined indirectly by comparing the rankings or ratings generated by a series of different raters (e.g., Bernardin, Alveres, & Cranny, 1976). A high degree of agreement among raters could be used to support the validity of the rating process. This is similar to using reliability data to make inferences about the content validity of a test (see p. 251) and confuses the true meaning of reliability and validity analyses. Although reliable ratings have the capacity to be valid, the presence of reliability alone is not sufficient to1 indicate validity. Despite these problems, evaluation scales, particularly those using rating scales, are a commonplace element of performance evaluation.

Regulation of Evaluation Procedures

EFOC guidelines also apply to the evaluation of workers for promotions and raises. In 1971, the Supreme Court affirmed these guidelines in Griggs v. Duke Power, a discrimination lawsuit filed by a group of African-American workers in North Carolina's Duke Steam Plant. The Court required Duke Power to provide specific evidence that the test used for promotion decisions was reliable and demonstrated a valid relationship with job activities.

The Court's emphasis on quantifiable validity data was underscored by the 1988 decision in Watson v. Fort Worth Bank and Trust. In this case, an African-American employee charged that a promotion procedure was discriminatory because the proportion of minorities selected was significantly less than what would be expected on the basis of (I) the proportion of African-Americans in the Fort Worth community and (2) the proportion of African-American's employed at the bank. In other words, the procedure used for promotion decisions selected significantly fewer minorities than would be expected by chance alone. Although this is not necessarily surprising—we expect a rigorous procedure to identify only the most competent people—the Court ruled that the selection ratios indicated adverse impact and that the procedure was discriminatory.

However, the courts do not always1 side with employees. In a 1979 case filed against Detroit Edison by the National Labor Relations Board, an employee denied promotion for a low test score was refused access to the test to check for a possible scoring error. The Supreme Court ruled that the company's desire to protect the test questions from possible distribution to other employees was reasonable.

PROCEDURES FOR EVALUATING PROGRAMS AND PRODUCTS

The preceding two sections described how psychological tests can be used to evaluate people—cither potential employees or current employees. In this section, we examine how psychological tests are used lo evaluate the workplace itself and the products provided by a business. Studies evaluating the workplace may focus on organizational issues or human factors issues. Studies of goods and services (business) products involve marketing research.

Organizational Studies

Except in cases of sell-employment, the employees ol a business or industry function as a group of people working within a hierarchical structure. Their attitudes and behavior are affected by both the characteristics of that structure and the dynamics that emerge as they interact (Sicgel & Lane, 1982). In other words, a workplace is a social system in which people have specific roles and are expected lo adhere to certain norms. Their job performance and their feelings about their jobs are affected by the way roles and norms are defined, the power structure created by the organization's hierarchy, and the general climate of the workplace (Muchinsky. 1987). Organizational studies may include observations, interviews, and the administration of attitude scales.

When assessing organizational issues, psychologists emphasize the importance of the fit between the person and the work environment (e.g., Pervin, 1968). When people work in jobs that match their interests, abilities, and goals and feel that they are treated fairly and with respect, they are likely to perform well, feel satisfied with their jobs, and experience low levels of stress. On the other hand, people who feel powerless, who dislike their jobs, or who feel that they are treated unfairly are likely to perform poorly, feel dissatisfied, and experience job stress.

When the fit between a person and a job is poor, we have two choices: replace the person or change something about the job. If our goal is to improve person-environment fit by modifying the workplace, it is important to develop techniques for assessing factors such as organizational climate, job satisfaction, and job-related stress. The results of these assessments can also be useful in evaluating how organizational characteristics contribute to job performance and work attitudes.

The Organizational Climate Questionnaire (Litwin & Stringer, 1968) is designed to assess employees' perceptions of organizational structure, standards, and reward policies. Each statement is rated on a 5-point scale from "strongly agree" to "strongly disagree." Statements address issues such as the extent to which jobs are clearly defined (structure), the extent to which people are accountable for their work (standards), and the extent to which people are given credit for their accomplishments (reward policies). The instrument is useful for exploring employee perceptions of the workplace itself and can be a point of departure for additional studies on the relationship of job performance and job satisfaction to organizational features.

The Maslach-Jackson Burnout Inventory (MJBI) is designed to measure the effects of workplace stress (Maslach & Jackson, 1981). Burnout is conceptualized as a consequence of severe and prolonged job stress in which employees become emotionally exhausted and develop negative, cynical attitudes about themselves and their jobs. The MJBI contains items such as "I feel emotionally drained from my work" and "I've become more callous toward people since I took this job." Measures of stress and burnout can be used along with measures of organizational climate to (1) identify individuals or groups of employees experiencing serious job-related problems and (2) explain how organizational features might contribute to these problems.

There are several popular measures of job satisfaction. The most frequently used and researched instrument is the Job Descriptive Index (JDI), designed to measure five aspects of the work environment: the work itself, supervision, pay, promotions, and co-workers (Smith, Kendall, & Hulin, 1969). The items on each scale are a series of words and phrases. Employees are instructed to mark each item "Y" (yes) if it describes their jobs, "N" (no) if it does not, and "?" (cannot say) if they cannot decide. Scores on the five sections are added to derive an overall job satisfaction score.

The inventory has adequate test-retest reliability (Schneider & Dachler, 1978), is equally valid for white, African American, and Hispanic-American workers (McCabe et al., 1980: Smith, Smith. & Rollo, 1974), and successfully measures several different facets of job satisfaction (Smith, Smith, & Rollo, 1974). In fact, the JDI may assess more aspects of job satisfaction than it was intended to measure, possibly because some scales include items lapping more than one element of a workplace din/iensipn (Yeager, 1981). For example, the supervision scale contains the items "hard to please" and "up-to-date," which appear to lap, respectively, supervisor interpersonal skills and the quality of supervision.

A second popular instrument is the Minnesota Satisfaction Questionnaire (MSQ), assessing 20 different aspects of work and the workplace (Weiss et al., 1967). Each item is rated on a 5-poinl scale from "very dissatisfied" (I) lo "very satisfied" (5). With 20 different scales, the MSQ permits assessment of several unique elements, such as creativity and independence as components of a job, and separate measurement of multidimensional elements such as job supervision.

One strategy in organizational research is to administer measures of job satisfaction, such as the JDI and the MSQ along with measures of organizational climate. The job satisfaction scores can be used to define groups of "satisfied" and "unsatisfied" employees, whose scores on climate measures can be compared to identify aspects of organizational structure and policies contributing lo different job satisfaction altitudes.

Human Factors Studies

Human factors2 studies ask whether the design of physical equipment, facilities, and the work environment is suitable for human use (McCormick & Sanders, 1982). The first extensive human factors research was conducted during World War II wllfo* problems arose in operating and maintaining new types of military equipment (McCormick & Ilgen, 1985). Today, human factors research explores the design of such diverse entities as industrial equipment, buildings, consumer products, and systems for transportation, communication, and the delivery of health care services.

The impact of a design feature can be assessed using physiological, performance, or subjective criteria (McCormick & Ilgen, 1985). Physiological criteria include heart rate, blood pressure, muscle activity, and energy expenditure and are particularly relevant to the design of industrial equipment. For example, it is possible that the design of a piece of industrial equipment places unnecessary physical stress on a worker's arms during the operation of the equipment. A human factors study could be used to test the impact of redesigning the machinery on muscle activity. Performance criteria include measures such as the time taken to complete a task, productivity, and the quality of job performance. Subjective criteria include attitudes, such as job satisfaction, and judgments, such as ratings of design features. Performance and subjective criteria are relevant to evaluation both of workplace design (e.g., a machine console) and the design of systems for delivering services (e.g., a transportation system).

Psychological tests may be useful when studies include performance criteria and/or subjective criteria. For example, a study of a new computer keyboard design might require a group of computer operators to:

1. Enter a data set within a^ specific time limit using an existing and a newly designed keyboard (a speed-based performance test), and
2. Rate the design features of each keyboard on dimensions such as key location, key tension, and finger cramps after 15 minutes of continuous use (a judgmental rating task).

Similarly, a study of the effect of adding an employee lounge on each floor might include a measure of job satisfaction (an attitude scale) and a survey assessing preferences among a set of possible lounge locations (a judgmental rating task).

The data generated by a human factors study will only be as good as the measures. A key concern, therefore, is the construction or selection of measures that are reliable and valid for each scenario. When psychological tests will be part of a human factors study, their use in the research project should be preceded by a pilot study of the measures themselves.

Marketing Studies

Rather than focusing on employees, marketing research examines the behavior and attitudes of consumers and others (e.g., legislators) who may affect the exchange of goods and services. The goal of a marketing study is to gather and analyze data relative to the current success of a product or service or predict the likely future success of a planned product or service (Schiffman & Kanuck, 1983).

Marketing research involves collection of data through observational studies, surveys, and experimentation, usually in the form of field testing a product or service (Stanton & Futrell. 1987). Two types of psychological tests—attitude and rating scales—can be used in the survey and experimentation techniques. For example, a company planning to market a new educational toy might administer attitude and rating scales to the parents of preschool children. A personal example can illustrate this procedure.

Several years ago, while approaching a discount store with my young son, I was approached by a researcher to participate in a marketing study. First, 1 was asked to complete a survey about the value of toys designed to help children learn basic concepts (e.g.. shape and color discrimination). The survey basically was an attitude scale, containing a series of items designed to identify the extent to which I was "pro" educational toys. When I completed the survey, the researcher look a moment lo glance over my answers and then invited me and my child to field test a new toy in an adjacent trailer. After 1 watched my son spend about 15 minutes playing with the toy, I was asked lo complete a questionnaire about the features of the toy and his interest in it. The questionnaire was a rating scale containing a variety of statements about the appropriateness of the toy for his age, its similarity to other currently marketed toys, its likely appeal to girls versus boys, and so on.

If we examine the interaction in detail, we can see how each component contributes to the marketing study. The initial attitude scale serves two important functions. First, the pattern of scores obtained would provide important data about the potential size of the market for educationally oriented toys. If many parents see toys as a valuable learning tool, then the potential market for these toys is large. Second, scores on the scale could be used lo identify the specific individuals with positive altitudes about educational toys. These individuals, who comprise the potential market for this toy, would be good candidates for a field test of the toy. The ratings scales administered after the field tests were designed lo provide information about the features of the toy and its merits relative lo other currently marketed items.

As we have said so many limes before, the data generated by psychological tests are only useful if the measures themselves are accurate. It would be important, therefore, to assess the reliability and validity of the altitude and rating scales for marketing research before deciding lo use them within a particular marketing study.

Lesson 12

ATTITUDES, INTERESTS, AND VALUES ASSESSMENT

In this chapter we examine approaches to the assessment of attitudes, interests, and values, broadly defined. Because they are formative in everything from work to worship, attitudes, interests, and values are fundamental to the identity of each individual. It is no accident that the adolescent who values aesthetic harmony later reveals an interest in literature and then pursues a vocation as English teacher. Nor is it surprising when a shy teenager with an analytic bent shows a passion for mathematics and becomes a computer scientist. The values held by persons shape their interests in life, which, in turn, shape career choices. Lives possess a coherency that is explained, in part, by the influence of interests and values.

Values not only link the individual to the world of work, they are intertwined in moral, spiritual, and religious matters as well. Whether we favor or oppose capital punishment, whether we find life meaningful or merely chaotic, whether we seek or avoid religious practice—these matters we resolve based upon personal values. In sum, the choices we make in matters of work, spiritual life, and personal conduct are not random, they are bound together by common threads that we call interests and values.

A problem faced by many young adults is that their values are unstated and their interests are unexplored. Furthermore, they lack knowledge about career options. In these cases, career selection can arouse anxiety, and perhaps it should. Lowman(1991) has noted that the process of finding a vocation can be as complex and as difficult as choosing a mate. The dilemma of career choice is not limited to young adults entering the job market, but also vexes older workers who are dissatisfied with their careers. Fortunately, a large array of tests and guidance approaches are available to help individuals identify values, interests, and potential career choices, as reviewed in this topic.

In Topic 12A, Interests and Values in Vocational Assessment, we survey the measurement of values and interests, especially as these concepts apply to vocational choice. We begin with a quick overview of historically relevant tests for the evaluation of general life values, and then turn to the application of specialized tests for career assessment and advising. In Topic 12B, Attitudes and the Assessment of Moral and Spiritual Concepts, we introduce the reader to methods and concerns in the measurement of attitudes, and then present assessment approaches pertinent to the moral, spiritual, and religious dimensions of the individual.

THE ASSESSMENT OF LIFE VALUES

In the popular media we find frequent reference to values and changes in values at the individual and national level. Politicians deplore the decline of family values, magazine editors denounce the absence of altruistic volunteerism, and columnists disparage the reemergence of materialism and careerism. Religious leaders enter the fray, too. As an antidote to global cynicism, they call for a return to spiritual values that affirm the meaning of life. Practically everyone has an opinion about values especially in regard to the presumed values of other persons or groups.

But what are values and how can they be measured? Although a huge amount of literature exists on the nature and definition of values, there is surprisingly little empirical research on their measurement. In general, psychologists define a value as a shared, enduring belief about ideal modes of behavior or end stales of existence (Rokeach, 1980). Values instill action, shape attitudes, and guide efforts to influence others. Values also arise in response to societal conditions and are therefore malleable to some degree (Ball-Rokeach, Rokeach, & Grube, 1984).

In this topic, we examine key issues and important tests that pertain to the assessment of personal values, broadly defined. We begin with a critique of wideband instruments that assess life values the social ends or goals considered desirable of achievement. The chapter then reviews assessment approaches in the moral, spiritual, and religious domains. This includes lengthy coverage of Kohlberg's (1981, 1984) classic method for the measurement of moral reasoning. We close with brief coverage of the overlooked literature on the measurement of spiritual and religious concepts.

Values are important because they provide a pervasive framework for personal actions and judgments. When we know the life values of an individual, we can predict typical behaviors and surmise likely attitudes. In a classic work on the topic, Rokeach (1968) underscores the importance of values:

> *To say that a person "has a value*' is to say that he has an enduring belief that a specific mode of conduct or end-state of existence is personally and socially preferable to alternative modes of conduct or end-states of existence. Once a value is internalized it becomes, consciously or unconsciously, a standard or criterion for guiding action, for developing and maintaining attitudes toward relevant objects and situations, for justifying one's own and others' actions and attitudes, for morally judging self and others, and for comparing self with others. Finally, a value is a standard employed to influence the values, attitudes, and actions of at least some others— our children's, for example, (pp. 159-160)*

This view that values are in some sense primary and formative also has been advanced by Kluckhohn (1951) and Smith (1963).

Values are more easily defined than measured. Few value scales have withstood the test of time. We survey three instruments here: the Study of Values is an interesting test mainly of historical importance; the Rokeach Value Survey is a highly respected research tool; the Values Inventory provides a cautionary illustration that bad tests occasionally do make their way into publication.

Study of Values

Psychologists have been interested in the assessment of personal values since early in the twentieth century. However, it is only in the last 30 years that psychometrically sound self-report measures of values have been developed. An early instrument in this vein was the Study of Values (SOV), an inventory designed to measure six basic evaluative attitudes: Theoretical (T), Economic (E), Aesthetic (A), Social (S), Political (P), and Religious (R) (Allport & Vernon, 1931; Allport, Vernon, & Lindzey, 1960). These six values were patterned directly after Spranger's (1928) Types of Men. In this influential book, the German intellectual Eduard Spranger argued that most people display one of the following as a dominant value that defines their personality:

- Theoretical (T): The dominant interest of the theoretical person is the discovery of truth.
- Economic (E): The economic person is primarily interested in what is useful.
- Aesthetic (A): The aesthetic person sees the highest value in form and harmony.
- Social (S): Love of people is the highest value for the social person.
- Political (P): The political person is interested primarily in power.
- Religious (R): The religious person places the highest value upon mystical unity with the cosmos.

The SOV scale consists of 30 questions which pit one value against another, and another 15 questions that require the rank ordering of values. Examples of the questions include the following:

- When you visit a church are you more impressed by a pervading sense of reverence and worship or by the architectural features and stained glass? [Religious versus Aesthetic]
- In your opinion, has general progress been advanced more by the freeing of slaves, with the enhancement of the value placed on individual life, or by the discovery of the steam engine, with the consequent industrialization and economic rivalry of European and American countries? [Social versus Economic]

From answers to the forced-choice questions and the rank ordering of values, a profile of values is plotted in ipsative manner, displaying the relative strength of the six values for each individual.

Lubinski, Schmidt, and Benbow (1996) demonstrated the merit of testing values with the SOV in a 20-year follow-up study of 203 intellectually gifted adolescents. Their gifted sample was first tested at age 13 and then again as adults at age 33. In general, the six themes revealed significant stability over this time period, with mean interindividual correlations of .37 for the various themes. This is remarkable, given that the teenage and young adult years are assumed to be a period of turmoil and change, especially in personal values, as young persons struggle to find an identity. Sex differences were notable: Males tended to shift toward a T-E-P profile as adults whereas females tended to shift toward an A-S-R profile. Even so, a common pattern was observed for all participants, with Aesthetic and Economic values taking on more saliency in young adulthood and Political and Social values revealing less dominance.

The Study of Values has provoked considerable discussion as a classroom demonstration tool in psychology courses, but otherwise has not been an influential test. A major problem with the instrument is that the six values are vaguely defined and too general to be of practical use. Nonetheless, the test did inspire others to develop more sophisticated and comprehensive approaches to values assessment. One of those who acknowledged a debt to Allport and the Study of Values was Milton Rokeach.

Rokeach Value Survey

Rokeach (1973) defined two kinds of values, instrumental and terminal. Instrumental values are desirable modes of conduct, whereas terminal values are desirable end states of existence. For example, ambition is an instrumental value, whereas family security is a terminal value. In devising the Rokeach Value Survey, a final list of 18 instrumental values was arrived at by condensing 555 "personality-trait" names into near-synonyms. The final list of 18 terminal values was derived from literature survey and other subjective, impressionistic approaches. The 36 values are listed in Table 12.1.

Table 12.1 The 36 Value Constructs from the Rokeach Value Survey, Form D

Terminal Values

A Comfortable Life
An Exciting Life
A Sense of Accomplishment
A World at Peace
A World of Beauty
Equality
Family Security
Freedom
Happiness
Inner Harmony
Mature Love
National Security
Pleasure
Salvation
Self-Respect
Social Recognition
True Friendship
Wisdom

Instrumental Values

Ambitious
Broadminded
Capable
Cheerful
Clean
Courageous
Forgiving
Helpful
Honest
Imaginative
Independent
Intellectual
Logical
Loving
Obedient
Polite
Responsible
Self-Controlled

Although the individual values are not defined in detail, each is accompanied by a short phrase or synonyms to clarify the item for respondents. For example, the first of the terminal values reads as follows: "A COMFORTABLE LIFE (a prosperous life)." Completing the survey is extremely simple. Respondents are asked to rank separately the 18 terminal and 18 instrumental values based on "their importance to you, as guiding principles in your life." The values are printed on gummed labels (for Form D). Subjects merely peel off the labels and arrange them in order of importance, removing and reattaching as needed. The rank for each item becomes the score for that value. Ties are not allowed, so value scores will range from 1 to 18, with lower scores indicating greater importance.

Reliability of the Value Survey can be approached in two ways. The first is the temporal stability of rank orderings for individual subjects. For this approach, the scale is administered twice and the two sets of rank orderings are correlated for each individual. Using this approach with four groups of college students (retest intervals of three weeks to four months), Rokeach (1973) reported median test-retest correlations ranging from .76 to .80 for terminal values, and .65 to .72 for instrumental values. The second way to examine reliability is to calculate the test-retest reliability of individual value scores separately, across all respondents. Using this approach, reliability of the individual scales is lower, about .65 for the terminal values and .56 for the instrumental values (Rokeach, 1973). These reliabilities are rather low in comparison to instruments with more items per scale—which is not surprising. After all, the "scales" on the Value Survey each consist of a single item. Nonetheless, with reliabilities this low, the Value Survey should be used only for research purposes such as description or comparison of group values. Individual interpretation for counseling purposes cannot be supported.

In an intriguing example of its application in research, Rokeach and his colleagues used the Value Survey to measure the effects of viewing a single 30-minute television program on values, attitudes, and behaviors (Ball-Rokeach, Rokeach, & Grube, 1984). The television program, hosted by EdAsner and known as "The Great American Values Test," was specially designed to influence viewers' ratings of the importance of the terminal values of freedom and equality. For example, over a fullscreen graphic display indicating that Americans had ranked freedom third and equality twelfth, on average, among 18 terminal values, Asner commented:

> *Americans feel that freedom is very important. They rank it third. But they also feel that equality is considerably less important. .. they rank it twelfth. Since most Americans value freedom far higher than they value equality, the question is: what does that mean? Does it suggest that Americans as a whole are much more interested in their own freedom than they are in freedom for other people" Is there a contradiction in the American people between their love of freedom and their lesser love for equality? By comparing your values with these results, you should be able to decide for yourself whether you agree with the average American's feelings about freedom and equality. (Ball-Rokeach et al.. 1984)*

A full discussion of this study would involve a lengthy detour away from the topic of psychological testing. However, the reader may appreciate a quick summary. The authors used a tightly controlled pretest-posttest design with experimental and control cities to determine the effects of viewing the program. For viewers who watched the show without interruption, mean rankings on equality went from 11.0 to 9.3, whereas for non-viewers the ratings on this value were quite stable. A number of other experimental checks (e.g., soliciting donations to provide cultural opportunities for African American children) also confirmed a real change in values. This study is a good example of the kind of social research for which the Value Survey is well suited.

Limitations of the Rokeach Value Survey

We have already mentioned that the individual scales of the Value Survey possess marginal reliability—which means that the instrument should not be used for individual guidance. Several additional limitations stem from the ipsative nature of the test. The reader will recall that an ipsative test is one in which the average of the scales is always the same for every examinee. In particular, the average rank for the 18 instrumental values will always be 9.5, and likewise for the terminal values. By definition, when an examinee gives some scales a high ranking, others must receive a low ranking. What is lost in this process is any absolute measure of the value for that individual. Suppose, for example, that we could measure the absolute strength of the 18 instrumental values on a scale from 1 to 100 (note: this is not possible with the Value Survey). Consider the case in which individual A has an absolute strength of 99 for ambitious and 98 for obedient with all other values below 90, whereas individual B has an absolute strength of 39 for ambitious and 19 for obedient with all other values below 10. Most likely, individual A would value ambition and obedience to a high degree, whereas individual B modestly values ambition and devalues obedience. In fact, individual B could be characterized as almost valueless. Yet, both persons would receive scores of 1 for ambitious and 2 for obedient. The Value Survey is not sensitive to magnitude differences within individual subjects, nor does it capture scaling differences between individuals.

Braithwaite and Law (1985) call attention to additional weaknesses of the Value Survey. They note that the inventory omits several important values, including physical well-being, individual rights, thriftiness, and carefreeness. Perhaps more significant, they criticize the Rokeach test for relying upon a single item for each value instead of using multi-item indices for the value constructs. They propose an alternative instrument (based on the Rokeach approach) that would presumably embody improved psychometric qualities in the measurement of personal values.

AN OVERVIEW OF INTEREST ASSESSMENT

In most applications of psychological testing, the goals of assessment are reasonably clear. For example, intelligence testing helps predict school performance: aptitude testing foretells potential for accomplishment: and personality testing provides information about social and emotional functioning. But what is the purpose of interest assessment? Why would a psychologist recommend it? What can a client expect to gain from a survey of his or her interests?

Interest assessment promotes two compatible goals: life satisfaction and vocational productivity. It is nearly self-evident that a good fit between individual interests and chosen vocation will help foster personal life satisfaction. After all when work is interesting we are more likely to experience personal fulfillment as well. In addition, persons who are satisfied with their work are more likely to be productive. Thus, employees and employers both stand to gain from the artful application of interest assessment. Several useful instruments exist for this purpose, and we will review the most widely used interest inventories later.

In the selection of employees, the consideration of personal interests may be of great practical significance to employers and therefore circumstantially relevant to the job candidates as well. We may sketch out a rough equation as follows: productivity = ability x interest. In other words, high ability in a specific field does not guarantee success; neither does high interest level. The best predictions are possible when both variables are considered together. Thus, employers have good reason to determine whether a potential employee is well matched to the position; the employee should like to know as well.

We begin with a critical examination of major interest tests. The six instruments chosen for review include the following:

- The Strong Interest Inventory (SII), the latest revision of the well-known Strong Vocational Interest Blank (SVIB)
- The Jackson Vocational Interest Survey (JVIS), a test that embodies modern methods for scale construction
- The Kuder General Interest Survey (KGIS), an instrument that incorporates a divergent philosophy of test construction
- The Vocational Preference Inventory (VPI), which measures six widely, used vocational themes
- The Self-Directed Search (SDS), a self-administered and self-scored guide to exploring career options
- The Campbell Interest and Skill Survey (CISS), a recent and appealing test that is simple in format but sophisticated in execution

The review of prominent interest tests is followed by the related topic of assessment in career and work values.

INVENTORIES FOR INTEREST ASSESSMENT

Strong Interest Inventory (SII)

The Strong Interest Inventory (SII) is the latest revision of the Strong Vocational Interest Blank (SVIB), one of the oldest and most prominent instruments in psychological testing (Strong, Hansen, & Cam-bell, 1994). We can best understand the SII by studying the history of its esteemed predecessor, the SVIB. In particular, we need to review the guiding assumptions used in the construction of the S' that have been carried over into the SII.

The first edition of the SVIB appeared in 1927, eight years after E. K. Strong formulated the essential procedures for measuring occupational interests while attending a seminar at the Carnegie Institute of Technology (Campbell. 1971; Strong, 1927). In constructing the SVIB, Strong employed two little used techniques in measurement. First, the examinee was asked to express liking or disliking for a large and varied sample of occupations, educational disciplines, personality types, and recreational activities. Second, the responses were empirically keyed for specific occupations. In an empirical key, a specific response (e.g. liking to roller skate) is assigned to the scale for a particular occupation only if successful persons in that occupation tend to answer in that manner more often than comparison subjects.

Although Strong did not express his underlying assumptions in a simple and straightforward manner, it is clear that the theoretical foundation for SVIB derives from a typological, trait-oriented conception of personality. Tzeng (1987) has identified the following basic assumptions in the development and application of the SVIB:

1. Each occupation has a desirable pattern of interests and personality characteristics among its workers. The ideal pattern is represented by successful people in that occupation.
2. Each individual has relatively stable interests and personality traits. When such interests and traits match the desirable interest patterns of the occupation the individual has a high probability to enter that occupation and be more likely to succeed in it.
3. It is highly possible to differentiate individuals in a given occupation from others-in-general in terms of the desirable patterns of interests and traits for that occupation.

Strong constructed the scales of his inventory by contrasting the responses of several specific occupational criterion groups with those of a people-in-general group. The subjects for each criterion group were workers in that occupation who were satisfied with their jobs and who had been so employed for at least three years. The items that differentiated the two groups, keyed in the appropriate direction, were selected for each occupational scale. For example, if members of a specific occupational group disliked "buying merchandise for a store" more often than people in general, then that item (keyed in the dislike direction) was added to the scale for that occupation.

The first SVIB consisted of 420 items and a mere handful of occupational scales (Strong. 1927). Separate editions for men and women followed shortly. The inventory has undergone numerous revisions over the years (Tzeng, 1987) culminating in the modern instrument known as the Strong Interest Inventory (Campbell, 1974; Hansen, 1992; Hansen & Campbell, 1985).

Although the Strong Interest Inventory (SII) was fashioned according to the same philosophy as the SVIB, the latest revision departs from its predecessors in three crucial ways:

1. The SII merges the men's and women's forms into a single edition.
2. The SII introduces a theoretical framework to guide the organization and interpretation of scores, as discussed later.

3. The SII incorporates a substantial increase in the number of occupational scales, particularly in the vocational/technical areas underrepresented in the SVIB.

The SII consists of 317 items grouped into seven sections. In the first five sections, the examinee records "Like," "Indifferent," or "Dislike" for

Table 12.2 Characteristic Items from the Strong Interest Inventory

Mark Like, Indifferent, or Dislike next to the following items.

1. Driving a truck __________
2. Being a fish and game officer __________
3. Chemistry __________
4. Doing applied research __________
5. Acting in a drama __________
6. Magazines about music __________
7. Sociology __________
8. Fundraising for charities __________
9. Buying goods for a store __________
10. People who are leaders __________
11. Regular work hours __________
12. Assertive people __________

occupations, school subjects, activities, leisure activities, and contact with different types of persons (Table 12.2). A sixth part requires the examinee to express a preference between paired items (e.g., dealing with things versus dealing with people). The seventh section consists of self-descriptive statements which the examinee marks "Yes," "No," or "?".

The SII can only be scored by prepaid answer sheets or booklets that are mailed or faxed to the publisher, or through purchase of a software system that provides on-site scoring for immediate results. The results consist of a lengthy printout that is organized according to several themes. All scores are expressed as standard scores with a mean of 50 and an SD of 10. Normative results for men and women are reported separately, but cross-sex comparisons can be achieved by simple visual transposition.

At the most global level are the six General Occupational Theme Scores, namely. Realistic, Investigative, Artistic, Social, Enterprising, and Conventional. These theme scores were based upon the theoretical analysis of Holland (1966, 1985ab), whose work we discuss later. Each theme score pertains to a major interest area that describes both a work environment and a type of person. For example, persons scoring high on the Realistic theme are generally quite robust, have difficulty expressing their feelings, and prefer to work outdoors with heavy machinery. Within the theme scores can be found 25 Basic Interest Scales such as Adventure, Mathematics, and Social Science. The interest scales are empirically derived and consist of substantially intercorrelated items.

The most specific results consist of 211 scores for the Occupational Scales. In the 1985 revision of the SII these scales were constructed in the usual manner by comparing responses of persons employed in the given occupation versus samples of men-in-general and women-in-general (Hansen, 1992; Hansen & Campbell, 1985). Sample sizes for the criterion groups ranged from 60 to 420, with most groups containing 200 or more persons. The criterion groups consisted of persons between the ages of 25 and 60 years, satisfied with their occupation, meeting certain minimum standards of successful employment, and employed in the given occupation for at least three years. Standardization of the 1985 version involved the testing of over 140,000 persons, of whom only 50.000 met the criteria for scale development.

1. A recent innovation on the SII is the addition of personal style scales (Harmon. Hansen, Borgen, & Hammer, 1994). These are designed to measure preferences for broad styles of living and working. These scales assist in vocational guidance by showing level of comfort with distinctive styles. The four style scales are
2. *Work Style*, on which a high score indicates a preference to work with people and a low score signifies an interest in ideas, data, and things:
3. *Learning Environment*, on which a high score indicates a preference for academic learning environments and a low score indicates a preference for more applied learning activities;
4. *Leadership Style*, on which a high score indicates comfort in taking charge of others and a low score indicates, uneasiness. and
5. *Risk Taking/Adventure*, on which a high score indicates a preference for risky and adventurous activities as opposed to safe and predictable activities.

The personal style scales each have a mean of 50 and a standard deviation of 10. Note that these are truly bipolar scales for-which each pole is distinct and meaningful.

Evaluation of the SII

The SII represents the culmination of over 50 years of study, involving literally thousands of research reports and hundreds of thousands of respondents. In evaluating this instrument, we can only outline basic trends in the research, referring the reader to other sources for details (Savickas, Taber, & Spokane, 2002; Tzeng, 1987; Campbell & Hansen, 1981; Hansen, 1984, 1987, 1992; Hansen & Campbell, 1985). We should also point out that evaluations of the reliability and validity of the SII are based in part upon its similarity to the SVIB. for which a huge amount of technical data exists.

Based upon test-retest studies, the reliability of the SI1-SVIB has proved to be exceptionally good in the short run, with one- and two-week stability coefficients for the occupational scales generally in the .90s. When the test-retest interval is years or decades, the correlations drop to the .60s and .70s for the occupational scales, except for respondents who were older (overage 25) upon first testing. For younger respondents first tested as adolescents, the median test-retest correlation after 15 years is around .50 (Lubinski, Benbow, & Ryan. 1995). But for older respondents, first tested after the age of 25, the median test-retest correlation 10 to 20 years later is a phenomenal .80 (Campbell, 1971). Apparently, by the time we pass through young adulthood, personal interests become extremely stable. The questions on the SII-SVIB capture that stability in the occupational scores, providing support for the trait conception of personality upon which these instruments were based.

The validity of the SII-SVIB is premised largely on the ability of the initial occupational profile to predict the occupation eventually pursued. Strong (1955) reported that the chances were about two in three that people would be in occupations predicted by high occupational scale scores, and about one in five that respondents would be in occupations for which they had shown little interest when tested. Although other researchers have quibbled with the exact proportions (Dolliver, Irvin, & Bigley, 1972), it is clear that the SII-SVIB has impressive hit rates in predicting occupational entry. The instrument functions even better in predicting the occupations that an examinee will not enter. In a recent study, Donnay and Borgen (1996) provide evidence for construct validity by demonstrating strong overall differentiation between 50 occupational groups on the SII:

> *The big picture is that people in diverse occupations show large and predictable differences in likes and dislikes, whether in terms of vocational interests or in terms of personal styles. And the Strong provides valid, structural, and comprehensive measures of these differences, (p. 290)*

The SII is used mainly with high school and college students and adults seeking vocational guidance or advice on continued education. Because most students' interests are undeveloped and unstabilized prior to age 13 or 14, the SII is not recommended for use below high-school level. As evident in the reliability data reported, the SII becomes increasingly valuable with older subjects, and it is not unusual to see middle-aged persons use the results of this instrument for guidance in career change.

Jackson Vocational Interest Survey (JVIS)

The Jackson Vocational Interest Survey (JVIS) is a relatively new instrument that contrasts sharply in several respects with the SII (Jackson, 1977; Verhoeve, 1993). The 34 basic interest scales on the JVIS are composed of two different types, work role scales and work style scales. The 26 work role scales measure specific interests pertinent to broad occupational themes such as mathematics, life science, adventure, business, and teaching. The 8 work style scales were designed to measure preferences for working in environments that require particular modes of behavior, such as job security, dominant leadership, accountability, and stamina. The JVIS may be hand scored, but computer scoring is probably preferable since the user then obtains several additional groups of scales, including data on examinees' similarity to college students majoring in specific academic disciplines. The JVIS is suited to high school age and older.

Several features distinguish the JVIS from the SII and other interest inventories. First, the JVIS employs a forced-choice ipsative format whereby examinees must select their preferred choice from two alternatives. Items on the JVIS resemble the following:

A. Acting in a school drama.
B. Teaching kids how to write.
A. Quilting bedspreads with ornate designs.
B. Buying furniture for a chain of stores.
A. Writing a mathematics text for grade school children.
B. Studying the financial growth of a local bank.

Although rarely used, the forced-choice item format has the advantage of reducing the impact of social desirability upon test results. A second distinctive feature of the JVIS is that Jackson used a rational and theory-guided method in the derivation of scales, as opposed to the empirical approach found in most other instruments. As a result of these two features, the JVIS scales possess a greater independence from one another than found on other instruments and are also quite factorially pure. As evidence of factorial homogeneity of the scales, biserial correlations between item endorsements and scale scores are typically in the high .60s and low .70s.

The JVIS is normed on a very large sample approximately 8,000 high school and college students. However, these subjects consist mainly of students from Pennsylvania and the Province of Ontario, so their representation of the general population is questionable. Reliability is excellent, at least in the short range, with one- to two-week test-retest coefficients typically in the mid-.80s. Based on an eclectic group of studies reported in the manual, concurrent and predictive validity appear promising, but additional studies are needed to bolster confidence in this instrument (Shepard, 1989).

Kuder General Interest Survey

The Kuder General Interest Survey (KGIS) represents the most recent evolution of a series of highly respected Kuder vocational interest inventories developed over the last 50 years. The first of these instruments, the Kuder Preference Record, was published in 1939. This instalment introduced an interesting forced-choice response format that has survived into the present (discussed later). The Preference Record underwent several revisions and emerged in 1979 as the Kuder Occupational Interest Survey-Revised (KOIS-R; Kuder & Diamond. 1979). The KOIS-R is a well-known test that produces scores for over 100 specific occupational groups and nearly 50 college majors. The target population for the KOIS-R is roughly the same as for the SII and the JVIB. For purposes of presenting a diversity of interest tests, it is more instructive to discuss the KGIS here.

The KGIS is unique among interest inventories in that its target population is restricted to adolescents in grades six through twelve (Kuder, 1975). The test requires only a sixth-grade reading level and may be administered by the classroom teacher and hand scored on site. Thus, the KGIS is well suited to the development of educational and vocational goals in the early formative years of adolescence.

The KGIS is also unusual in its methodology: The inventory uses a forced-response triad format to measure interests. Specifically, each item on the test requires the examinee to indicate most- and least-liked alternatives from three statements. This forced-choice approach is particularly suited to identifying examinees who have not answered the items sincerely.

The 168-item inventory produces 10 interest scores that are largely ipsative in nature. The reader will recall that scores on an ipsative test reflect intraindividual variability rather than interindividual variability. With the KGIS, comparison to an external reference group is of secondary importance in determining scores. Thus a high score in one interest area mainly means that the examinee ferred that area more often than the others in the forced-choice items.

The 10 scales reflect broad areas of interest: Outdoor, Mechanical, Computational, Scientific, Persuasive, Artistic, Literary, Musical, Social Service, and Clerical. An eleventh scale, the Verification Scale, is designed to determine the sincerity of the responses. The manual reports extensive test-retest, internal consistency, and stability data based on a sample of 9,819 students in grades 6 through 12. The six-week test-retest and internal consistency data are generally acceptable, with the older students showing higher test-retest correlations. The possible exception to good reliability is the Persuasive Scale (pertinent to sales positions), which shows test-retest correlations of .69 and .73, respectively, for boys and girls in grades 6 through 8.

Stability data over a four-year follow-up are less impressive. The mean stability coefficient is only .50, and for low-IQ subjects (below 100) it is even lower as low as .19 for the Clerical Scale. This is unfortunate because low-IQ adolescents would be more likely to enter clerical fields than high-IQ adolescents. Yet, measurement of clerical interests is highly unstable for precisely this group.

Comment on the KGIS and Other Interest Inventories

Considering the difficulty of the task it undertakes—measuring the broad interest patterns of adolescents—the KGIS performs at an acceptable level. In grades 6 through 8, results of the KGIS may spur students to explore new experiences pertinent to their measured interests; in grades 9 and 10, results may help students plan high school courses; and in grades 11 and 12 the results can help students make tentative vocational choices.

But the KGIS suffers the same pivotal shortcoming of all existing interest inventories, a total inattention to opportunity. Williams and Williams (1985) have expressed this point well:

> *For those specifically looking for a measure of interest, the Kuder is definitely an acceptable mea-:1 sure. But interest is only one prong in the triumvirate of interest-ability-opportunity. The most important prong, opportunity, has generated the least psychometric interest. That this would be so is not surprising. Opportunity is by far the hardest construct to define, but those who deal in career counseling should never ignore it, regardless of the difficulty in measurement and definition.*

We remind the reader that the inattention to opportunity is common to all interest measures, although it is perhaps a more serious problem for the KGIS because this instrument is used with persons who have not yet entered the job market.

Vocational Preference Inventory

The Vocational Preference Inventory is an objective, paper-and-pencil personality interest inventory used in vocational and career assessment (Holland, 1985c). The VPI measures eleven dimensions, including the six personality-environment themes of Realistic, Investigative, Artistic, Social, Enterprising, and Conventional, and five additional dimensions of Self-Control, Masculinity/Femininity, Status, Infrequency, and Acquiescence. The test items consist of 160 occupational titles toward which the examinee expresses a feeling by marking y (yes) or n (no). The VPI is a brief test (15 to 30 minutes) and is intended for persons 14 years and older with normal intelligence.

Holland proposes that personality traits tend to cluster into a small number of vocationally relevant patterns, called types. For each personality type there is also a corresponding work environment best suited to that type. According to Holland, there are six types: Realistic, Investigative, Artistic, Social, Enterprising, and Conventional. This is sometimes known as the RIASEC model, in reference to the first letters of the six types. The types are idealizations that few people (or environments) fit completely. Nonetheless, Holland believes that most individuals tend to resemble one type more than the others. In addition, individuals show a lesser degree of resemblance to a second and third type as well.
We can summarize the personality-environment types as follows:

- *Realistic*: athletic, lacks verbal and interpersonal skills, and prefers "hands-on" or outdoors vocations such as mechanic, farmer, or electrician
- *Investigative*: task-oriented thinker with unconventional attitudes who fits well in scientific and scholarly positions such as chemist, physicist, or biologist
- *Artistic*: individualistic, avoids conventional situations, and prefers aesthetic pursuits
- *Social*: uses social competencies to solve problems, likes to help others, and prefers teaching or helping professions
- *Enterprising*: a leader with good selling skills who fits well in business and managerial positions
- *Conventional*: conforming and prefers structured roles such as bank teller or computer operator

The six themes in the RIASEC system can be arranged in a hexagon with similar themes side by side and dissimilar themes opposite one another, as depicted in Figure 12.1.

Test-retest reliability coefficients for the six major scales range from .89 to .97. VPI norms are based upon large convenience samples of college students and employed adults from earlier VPI editions. The characteristics of the standardization sample are not well defined, which makes the norms somewhat difficult to interpret (Rounds, 1985).

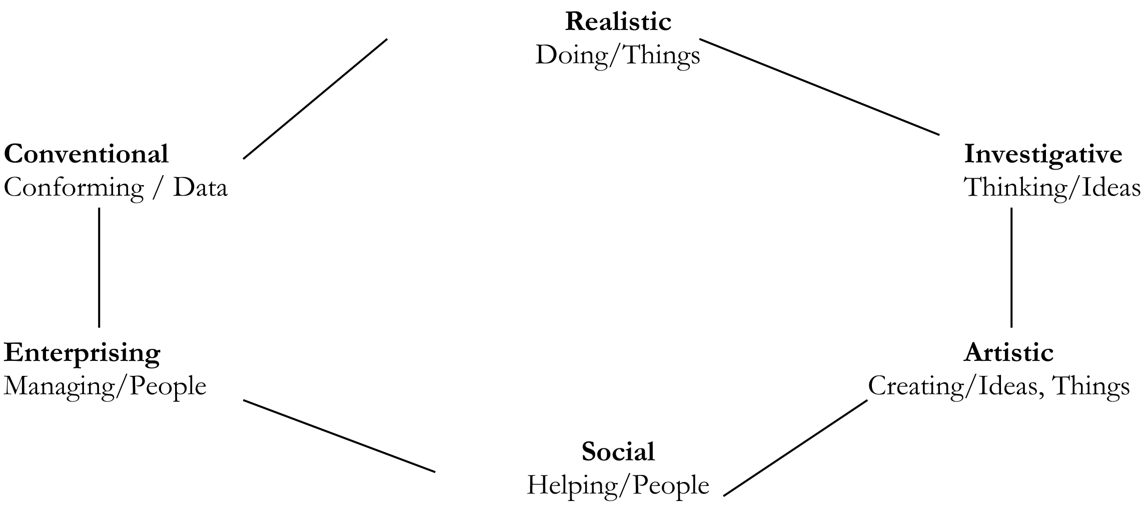


Figure 12.1 Holland's Hexagonal Model of Occupational Themes

The validity of the VPI is essentially tied to the validity of Holland's (1985a) hexagonal model of vocational interests. Literally hundreds of studies have examined this model from different perspectives. We will cite trends and representative studies. The reader is referred to Holland (1985c) and Walsh and Holland (1992) for more details.

Several VPI studies have investigated a key assumption of Holland's theory—that individuals tend to move toward environments that are congruent with their personality types. If this assumption is correct, then the real-world match between work environments and personality types of employees should be substantial. We should expect to find that Realistic environments have mainly Realistic employees, Social environments have mainly Social employees, and so on. Research on this topic has followed a straightforward methodology: Subjects are tested with the VPI and classified by their Holland types (using up to six letters); the work environments of the subjects are then independently classified by an appropriate environmental measure; finally, the degree of congruence between persons and environments is computed. In better studies, a correction for chance agreement is also applied.

Using his hexagonal model. Holland has developed occupational codes as a basis for classifying work environments (Gottfredson & Holland, 1989; Holland, 1966, 1978. 1985c). For example, landscape architect is eoded as RIA (Realistic. Investigative. Artistic) because this occupation is known to be a technical, skilled trade (Realistic component) that requires scientific skills (Investigative component) and also demands artistic aptitude (Artistic component). The Realistic component is listed first because it is the most important for landscape architect, whereas the Investigative and Artistic components are of secondary and tertiary importance, respectively. Some other occupations and their codes are taxi driver (RSE), mathematics teacher (ISC), reporter (ASE), police officer (SRE), real estate appraiser (ECS), and secretary (CSA). In a similar manner, Holland has also worked out codes for different college majors.

One approach to congruence studies is to compare VPI results of students or workers with the Holland codes that correspond to their college majors or occupations. For example, VPI Holland codes for a sample of police officers should consist mainly of profiles that begin with S and should contain a larger-than-chanee proportion of specifically SRE profiles. Furthermore, the degree of congruence should be related to the degree of expressed satisfaction with that line of work or study.

Research with college students provides strong support for the congruence prediction: Students tend to select and enter college majors that are congruent with their primary personality types (Holland. 1985a: Walsh & Holland, 1992). Thus, Artistic types tend to major in art. Investigative types tend to major in biology, and Enterprising types tend to major in business, to cite just a few examples. These results provide strong support for the VPI and the theory upon which it is based.

This short review has barely touched the surface of supportive validity studies with the VPI. Walsh and Holland (1992) cite several additional lines of research that buttress the validity of test. But not all studies of the VPI affirm its valkity, Furnham, Toop, Lewis, and Fisher (1995) failed to find a relationship between person-environment (P-E) "fit" and job satisfaction, a key theoretical underpinning of the test. According to Holland's theory, the better the P-E fit, the greater should job satisfaction. In three British samples, the relationships were weak or nonexistent, suggesting that the VPI does not "travel well" in cultures outside the United States.

Although we have emphasized mainly the strengths of the VPI up to this point, even the authors of the test acknowledge that there is room f< improvement. For example, Walsh and Holland (1992) cite the following weaknesses of the VPI: (1) the notions about vocational environments only partially tested; (2) the hypotheses about that person-environment interactions need consider able additional research work; (3) the formulation; about personal development have received some support but need more comprehensive examinations; (4) the classification of occupations may differ depending on the device used to assess the personality types; and (5) there are personal and environmental contingencies that are currently outside the scope of the theory.

The last weakness is perhaps the most serious. After all, the VPI assessment approach does not currently recognize any role for education, intelligence, and special aptitudes except insofar as these factors might indirectly bear upon personality and vocational interests. Yet, common sense dictates that intellectual ability will have a great deal to do with vocational satisfaction for some professions, independent of the match between personality type and work environment. For further discussion of the VPI and the theory upon which it is based, the interested reader is referred to Gottfredson (1990), Holland (1990), and Holland and Gottfredson (1990).

Self-Directed Search

Holland has always shown a keen interest in the practical applications of his research on vocational development. Consistent with this interest, he developed the Self-Directed Search, a highly practical, brief test that is appealing in its simplicity (Holland 1985ab). As the name suggests, the Self-Directed Search is designed to be a self-administered, self-scored, and self-interpreted test of vocational interest. The SDS measures the six RIASEC vocational themes described previously.

The SDS consists of dichotomous items that the examinee marks "like" or "dislike" (or "yes" or "no") in four sections: (1) Activities (six scales of 11 items each); (2) Competencies (six scales of 11 items each); (3) Occupations (six scales of 14 items each); and (4) Self-Estimates (two sets of six ratings). For each section, the face-valid items are grouped by RIASEC themes. For each theme, the total number of "like" and "yes" answers is combined with the self-estimates of ability to come up with a total theme score. The SDS takes 30 to 50 minutes for completion and is intended for persons 15 years and older.

The RIASEC themes on the SDS showed test-retest reliabilities that range from .56 to .95 and internal consistencies that range from .70 to .93. Norms for SDS scales and codes are reported for pooled convenience samples of 4,675 high school students, 3,355 college students, and 4,250 employed adults ages 16 through 24 (Holland, 1985ab). However, SDS results are typically interpreted in an individualized, ipsative manner ("Is this occupation a good fit for this client?"), so normative data are of limited relevance.

The SDS is available in a hand-scored paper-and-pencil version and a computerized version as well. Unfortunately, the paper-and-pencil version is prone to a 16 percent clerical error rate when used by high school students (Holland. 1985ab). The user-friendly microcomputer test is probably the preferred version because of the ease of administration and the error-free scoring and interpretation.

When a subject takes the SDS, the three highest theme scores are used to denote a summary code. For example, a person whose three highest scores were on Investigative, Artistic, and Realistic would have a summary code of IAR. In a separate booklet distributed with the test—the Occupations Finder—-the examinee can look up his or her summary code and find a list of occupations that provide the best "fit." For example, an examinee with an IAR summary code would learn that he or she most closely resembles persons in the following occupations: anthropologist, astronomer, chemist, pathologist, and physicist. The test booklet contains additional information which helps the examinee explore relevant career options.

The SDS serves a very useful purpose in providing a quick and simple format for prompting young persons to examine career alternatives. By eliminating the time-consuming process of administration, scoring, interpretation, and counselor feedback, the test makes it possible for a wide audience to receive an introductory level of career counseling. Holland (1985ab) proposes that the SDS is appropriate for up to 50 percent of students and adults who might desire career guidance. Presumably, the other 50 percent would find the SDS an insufficient basis for career exploration. Holland (1985ab) rightfully warns users to consider many sources of information in career choice and not to rely too heavily on test scores per se. Levinson (1990) discusses the integration of SDS data with other psychoeducational data to make specific vocational recommendations for high school students.

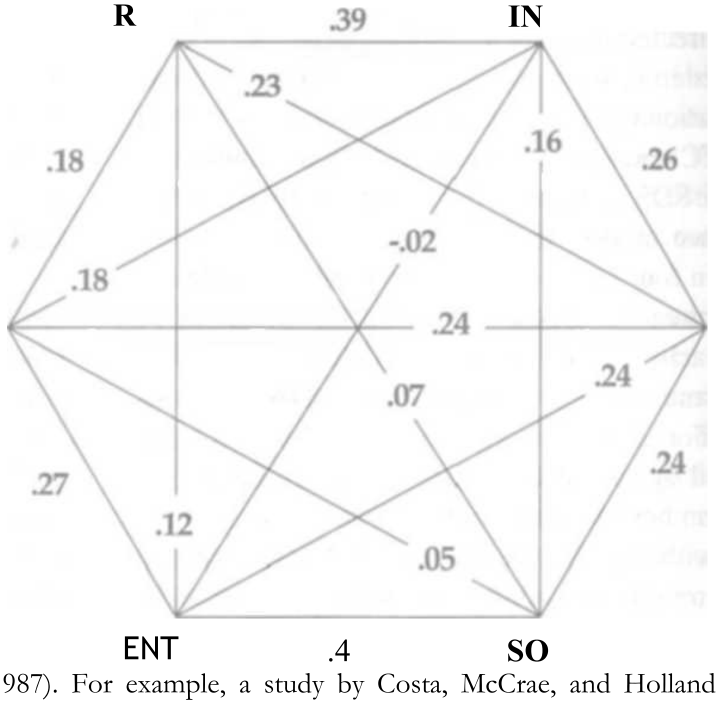
The validity of the SDS is linked to the validity of the hexagonal model of personality and environments upon which the test is based. One aspect of validity, then, is whether the model makes predictions which are confirmed by SDS results in the real world. In general, the results from over 400 studies support the construct validity of the SDS (Dumenci. 1995: Holland, 1985ab, 1987).

One approach to construct validity is to determine whether the relationships between SDS scales make theoretical sense. As is true of the VPI, the six RIASEC themes of the SDS can be arranged in a hexagon with similar themes side by side and dissimilar themes opposite one another. For example, in Figure 12.2, Artistic and Investigative themes are adjacent. It is not difficult to imagine one person combining these two themes in personality and work environment, so we would predict a moderate positive correlation between them. In a general reference sample of 175 women ages 26 to 65 years, Holland (1985ab) found that scores on these two themes correlated modestly, r - .26, as would be expected. The reader will also notice that the Investigative and Enterprising themes are opposite one another, signifying the huge disparities in these two occupational motifs. These themes should be uncorrelated. In fact, scores on these two themes correlated very little, r = -.02. In general, the correlations found in Figure 12.2 make theoretical sense; these findings support the construct validity of the SDS.

The predictive validity of the SDS has been investigated in several dozen studies, which are summarized by Holland (1985ab, 1987). The typical methodology for these studies is that SDS high point codes for large samples of students are compared with the first letter of their occupational choices (or aspirations) one to three years later. Overall, the findings indicate that the SDS has moderate to high predictive efficiency, depending upon the age of the sample (hit rates go

up with age), the length of the time interval (hit rates go down with time), and the specific category predicted (hit rates are better for Investigative and Social predictions) (Gottfredson & Holland, 1975).

Correlations between SDS scales and a wide range of other psychological measures (e.g., personality, aptitudes, and values) also serve to define the meaning of SDS scales and therefore help to validate the test (Holland, 1985ab, 1987). For example, a study by Costa, McCrae, and Holland (1984) investigated the relationship between SDS scales and the NEO Personality Inventory for a sample of 217 men and 144 women ages 21 to 89 years. The Investigative and Artistic scales from the SDS showed strong positive correlations with the NEO Openness scale—a measure of openness to experience in the areas of fantasy, feelings, actions, and ideas. The Social and Enterprising scales from the SDS showed strong correlations with the NEO Extraversion scale—a measure of outward directness and sociability. The Realistic and Conventional scales from the SDS revealed only trivial correlations with the NEO scales. Overall, the observed correlations were consistent with the interpretation of the I, A, S, and E scales and provide good



support for their validity. Although results for this study failed to support the validity of the R and C scales, many other investigations have yielded confirmatory findings (Holland, 1985ab). Schinka, Dye, and Curtiss (1997) provide a thoughtful analysis and discussion of the relationship between NEO dimensions and the SDS scales.

Campbell Interest and Skill Survey

The Campbell Interest and Skill Survey (CISS; Campbell, Hyne, & Nilsen, 1992) is a newer measure of self-reported interests and skills. The test is designed to help individuals make better career choices by describing how their interests and skills match the occupational world. The primary target population for the CISS is students and young adults who have not entered the job market, but the test is also suitable for older workers who are considering a change in careers. The test is appropriate for persons 15 years of age and older with a sixth-grade reading level, although younger children can be tested in exceptional circumstances.

The CISS consists of 200 interest items and 120 skill items. The interest items include occupations, school subjects, and varied working activities that the examinee rates on a six-point scale from strongly like to strongly dislike. The interest items resemble the following:

A pilot, flying commercial aircraft
A biologist, working in a research lab
A police detective, solving crimes

The skill items include a list of activities that the examinee rates on a six-point scale from expert (widely recognized as excellent in this area) to none (have no skills in this area). The skill items resemble the following:

Helping a family resolve its conflicts
Making furniture, using woodworking and power tools
Writing a magazine story

CISS results are scored on several different kinds of scales: Orientation Scales, Basic Interest and Skill Scales, Occupational Scales, Special Scales, and Procedural Checks. All scale scores are reported as T scores, normed to a population average of 50, with a standard deviation of 10.

The Orientation Scales serve to organize the CISS profile—the interest, skill, and occupational scales are reported under the appropriate Orientations. The seven Orientations are as follows (Campbell et al., 1992, pp. 2-3):

- *Influencing*—influencing others through leadership, politics, public speaking, and marketing
- *Organizing*—organizing the work of others, managing, and monitoring financial performance
- *Helping*—helping others through teaching, healing, and counseling
- *Creating*—creating artistic, literary, or musical productions, and designing products or environments
- *Analyzing*—analyzing data, using mathematics, and carrying out scientific experiments
- *Producing*—producing products, using "hands-on" skills in farming, construction, and mechanical crafts
- *Adventuring*—adventuring, competing, and risk taking through athletic, police, and military activities

There are 29 pairs of Basic Scales, each pair consisting of parallel interest and skill scales. The Basic Scales are clustered within the seven Orientations, based upon their intercorrelations. For example, the Helping Orientation contains the following Basic Scales, each with separate interest and skill components: Adult Development, Counseling, Child Development, Religious Activities, and Medical Practice.

The 58 pairs of Occupational Scales, each with separate interest and skill components, provide feedback on the degree of similarity between the examinee and satisfied workers in that occupation. These scales were constructed empirically by contrasting the responses of happily employed persons in specific occupations with responses of a general reference sample drawn from the working population at large.

In addition to Basic and Occupational Scales, the CISS incorporates three special scales: Academic Focus, a measure of interest and confidence in intellectual, scientific, and literary activities; Extraversion, a measure of social extraversion: and Variety, a measure of the examinee's breadth of interests and skills. Finally, the CISS reports a variety of Procedural Checks to detect possible problems in test taking such as random responding or excessive omissions.

Overall, the reliability of CISS scales is exceptionally strong. For example, coefficient alpha for the Orientation Scales is typically in the high .80s. and three-month test-retest reliabilities for 324 respondents are in the mid- to high .80s. Similar findings for reliability are reported for the Basic and Occupational Scales. Norms for the CISS are based upon 5,000 subjects spread over the 58 occupations. The authors report extensive validity data for the Occupational Scales, including sample means for each occupational sample as well as lists of the three highest- and lowest-scoring occupations for each scale (Campbell et al., 1992). These data document that the scales do discriminate between occupations in an effective and meaningful way. For example, the average T score on accountant by accountants is 75.8. Statisticians, bookkeepers, and financial planners achieve the next three highest scores for this scale, with average T scores in the low 60s. Commercial artists, professors, and social workers obtain the three lowest scores, with average Tscores around 40. Because these results tit well with our expectations about occupational interest and skill patterns, they provide support for the validity of the CISS.

Independent correlational studies also support the validity of the CISS. For example, in a sample of 118 adults. Savickas et al. (2002) correlated scores from individual occupational scales of the CISS with scores from the scales of other mainstream instruments such as the Strong Interest Inventory. They found strong support for both convergent validity (i.e., modest correlations for same-named pairs of scales) and discriminant validity (i.e., negligible correlations for unlike pairs of scales). In a sample of 128 college students, Hansen and Neuman (1999) confirmed the concurrent validity of the CISS by finding a good fit between occupational scale scores and students' chosen majors. The tit was considered "excellent" or "moderately good" for more than 70 percent of the students. Boggs (1999) provides a review and critique of the CISS. Campbell (2002) presents the history and development of the instrument.

This instrument will almost certainly receive increased attention in the years ahead. One noteworthy feature of the CISS is the comprehensiveness and clarity of the profile report form. The report consists of 11 user-friendly pages. This format is preferable to the detail-rich but eye-straining graphs encountered with many instruments. The CISS promises to rival the Strong Interest Inventory for vocational guidance of young adults.

CAREER AND WORK VALUES ASSESSMENT

In Working, his monumental discourse about Americans on the job, Studs Terkel concluded that work is a search

For daily meaning as well as daily bread, for recognition as well as cash for astonishment rather than torpor; in short, for a sort of life rather than a Monday through Friday sort of dying. Perhaps immortality, too, is part of the quest. To be remembered was the wish, spoken and unspoken, of the heroes and heroines of this book. (Terkel, 1974)

People seek meaning in their work. After interviewing hundreds of workers, Terkel concluded that only a lucky few find this meaning. Everyone can recall such fulfilled souls: the minister who is adored by his flock, the landscaper who proudly leaves an enduring legacy, the auto mechanic who delights in the perfectly tuned engine, or the oral historian who rescues a piece of the past. But contrasted with these few, Terkel discovered that the vast majority harbor a "hardly concealed discontent" about work.

Whether we agree or disagree with this pessimistic position, it is clear that values play an important role in work satisfaction, career choice, and career development. This is especially evident when a mismatch arises between personal values and the dominant values required by a career. In her book on career changes, Jones (1980) relates the story of an advertiser who came to a painful midlife realization: "I disliked the focus of my work. The advertising of bad products is damaging the country. . . The whole idea of advertising seemed wrong" (p. 27). This person was so dissatisfied with his vocation that he switched to another field in midlife. Apparently, he valued service to others—a work value that collided head-on with the amoral stance prevalent in advertising. We can only wonder how he picked such a mismatched career in the first place, but it seems unlikely that it was a rational choice based on an assessment of his work values.

It is also worth asking about the source of the discontent that Terkel discovered. Is this discontent largely unavoidable, inherent to the very nature of work? Or does it arise, at least in part, from the millions of individual mismatches between what a job offers and what a worker needs?

In this section we expand upon the theme developed in the preceding topic—that career choice can be enhanced through the appropriate application of career assessment tools. The reader will first encounter individual tests of work values and career development. Next, we discuss the integrative career assessment model. In this approach, abilities, interests, and personality characteristics are integrated in vocational guidance. The chapter closes on the related topic of consumer assessment.

Work values refer to needs, motives, and values that influence vocational choice, job satisfaction, and career development. Even when background factors such as intelligence, education, and ability are held constant, it is clear that individuals differ in their work values. What one person desires from his or her work might be positively poisonous to another individual of equal intelligence, education, and ability.

Here is a true story to illustrate this point. On behalf of several families, an attorney specializing in personal injury filed a lawsuit against a large mining corporation. The mining company was accused of spewing poisonous lead smelter emissions into the air breathed by hundreds of small-town residents, causing subtle neurological damage to dozens of children. The lawsuit involved more than 20 expert witnesses and dragged on for years. The lawyer was deep in debt from financing the protracted litigation—he faced bankruptcy if the lawsuit failed. Yet, he was ecstatic as he approached the final showdown in U.S. Federal Court, his entire career on the line. For this individual, perilous risk taking was a cherished work value. In contrast, most persons would actively avoid this kind of high-stakes gamble with their careers.

A proper match between work values and career choice is essential for job satisfaction. Some people succeed in finding such a match. For example, the intrepid lawyer mentioned here was well suited to his career path. But for those uncertain about career choice, feedback about work values can provide much-needed guidance. There are several assessment tools that might be helpful in this regard, but three instruments deserve special mention: the Minnesota Importance Questionnaire, the Work Values Inventory, and the Values Scale are reviewed in the following sections.

Minnesota Importance Questionnaire

The Minnesota Importance Questionnaire (MIQ) was developed to measure vocational needs and values of adults from high school age on up. The test has a solid foundation in a theory of work adjustment that emphasizes the importance of person-environment correspondence in determining satisfaction from work (Dawis, England, & Lofquist. 1964; Dawis & Lofquist, 1984; Lofquist & Dawis, 1991). According to the theory, work satisfaction is directly related to the correspondence between the worker's needs and the rewards or reinforces available from the job. For example, a prospective employee who has a strong need to help other people will probably find satisfaction in a job that provides plentiful opportunities for social service; conversely, he or she might be miserable in a position that emphasizes solitary work.

In its most popular form—which consists of paired-comparison items—the MIQ measures 20 needs organized into six underlying values relevant to work satisfaction. The test also comes in a ranked format that we do not discuss here. The six values emerged from factor analyses of the needs. The values and their component needs are listed in Table 12.3. It is important to emphasize that each need "scale" actually consists of a single need statement. For example, the Independence need "scale" actually consists of a single statement resembling the following: "Could make my own decisions."

The MIQ consists of 210 items. These include 190 items that pair each of the 20 needs with every other need. An additional 20 items require absolute judgments of the importance of each need dimension. The paired-comparison items are in reference to the examinee's "ideal job" and resemble the following:

___________ could give me a sense of accomplishment OR
__________ could make my own decisions

Table 12.3 Values and Components of the Minnesota Importance Questionnaire

Values	Components
Achievement	Ability Utilization Achievement
Comfort	Activity Independence Variety
	Compensation
	Security
	Working Conditions
Status	Advancement Recognition
	Authority
Altruism	Coworkers
	Social Service
	Moral Feelings
Safety	Company Policies
	Supervision—Human Relations
	Supervision—Technical
Autonomy	Creativity
	Responsibility

The examinee is instructed to select the alternative of greater personal importance in a job—hence the reference to Importance in the title of the instrument. In the preceding example, the achievement need (Could give me a sense of accomplishment) is matched with the responsibility need (Could make my own decisions). In order to pair each need with every other need, 190 items are required. Each of the 20 work needs is also rated individually on an absolute scale of importance, which results in a total of 210 items. These absolute judgments permit comparisons across examinees or across scales within examinees.

The MIQ is interpreted in reference to occupational reinforcer patterns (ORPs) for nearly 200 occupations. The ORPs were derived from a parallel research program using the Minnesota Job Description Questionnaire (MJDQ), a scale that resembles the MIQ. The MJDQ requires current job holders to rate the perceived presence or absence of reinforcers in a given job. Of course, these reinforcers are simply the 20 work needs appropriately restated so as to capture occupational requirements.

By comparing the profile of examinee needs and values with known reinforcer patterns for representative occupations, MIQ results can be used to predict satisfaction in specific jobs. This is done by means of the C-Index, or correspondence index, which is the correlation coefficient between the individual's MIQ profile and the ORP for each occupation. Satisfaction is predicted in an occupation when the correlation exceeds .50.

The paired-comparison format of the MIQ permits the examiner to evaluate the consistency of responses. Consider any three needs, designated as A, B, and C. Suppose an examinee prefers A to B and also prefers B to C. Logically, this person also should prefer A to C (transitivity). This is an example of a logically consistent triad (LCT). The LCT score is the percentage of all triads that are logically consistent. This score provides an index of response consistency that is one measure of test taking validity. LCT scores below 33 raise a suspicion that the examinee has responded carelessly or randomly. In test-retest studies, the higher the LCT the more stable the examinee's MIQ profile.

Reliability of the MIQ is fair to excellent, depending upon the retesting interval. The median test-retest correlation for the 20 scales is reported to be .89 for immediate retesting, but only .53 for retesting after 10 months. Internal consistency reliabilities are typically around .80 (Rounds et al. 1981).

Approximately 200 studies bear upon the validity of the MIQ, so it is difficult to summarize trends (Layton, 1992). The results indicate that the 20 MIQ scales discriminate among distinct occupational groups; that correlations with the Strong Vocational Interest Blank are significant and theory consistent; and that the MIQ has appropriately low correlations with abilities as measured by the General Aptitude Test Battery (Benson, 1985; Lay-ton, 1992). In an affirming study, scores on the MIQ Independence scale moderately predicted whether graduate students in counseling psychology would become scientists or practitioners when they entered the job market (Tinsley, Tinsley, Boone, & Shim-Li, 1993).

The MIQ is a well-respected instrument that deserves to be broadly used. Curiously, the test has never really captured wide attention. Perhaps this is due to the format of the instrument, which might be an impediment to its adoption by human service personnel. The problem is that examinees encounter the same need statements time and again. In order to pair each need with every other need, it is necessary to use each individual need statement in 19 separate test items. Examinees feel like they encounter the same questions over and over, even though each item on the MIQ is, in fact, unique. Regardless of its psychometric soundness, from the standpoint of the examinee the MIQ is an unappealing instrument.

Work Values Inventory

The Work Values Inventory (WVI) is a short and simple instrument designed to measure 15 work values in individuals from junior high level through high school (Super, 1968, 1970). The test is the end product of decades of research on the goals that motivate individuals to work. The 15 work values were identified through a literature review that included the early, classic work of Spranger (1928) on Types of Men. Items, scales, and test formats were continually revised and refined until the current 5-point rating approach was selected (Super, 1970, 1973).

The WVI is a self-report instrument consisting of 45 items rated on a 5-point scale from '"Very Important" to "Unimportant." Test items resemble the following: "Become famous in your field," "Make your own job decisions," "Feel you have helped other people," and "Have a boss who is considerate." There are three items for each of the 15 scales. The 15 work values measured by the test include the following:

Altruism
Esthetic
Creativity
Intellectual
Achievement
Independence
Prestige
Management

Economic Returns
Security
Surroundings
Stimulation Supervisory Relations
Associates
Way of Life
Variety

These work values are described in detail in the manual. For example, Altruism is described as "present in work that enables one to contribute to the welfare of others" and Prestige is described as "associated with work that gives one standing in the eyes of others and evokes respect." The items and scales are transparent. For example, the item "Become famous in your field" would belong on the Prestige scale.

Results for the WVI are reported as 15 scale raw scores, each ranging from 3 (low score) to 15 (endorsing "Very Important" for all three items on the scale). These scores permit three strategies of interpretation. In a clinical analysis, the three highest scores are highlighted for purposes of discussion by counselor and examinee. The data can also be analyzed normatively with respect to results for other students of the same age. Most importantly, it is also possible to predict satisfaction in various occupations by use of occupational reinforcer patterns (ORPs). The approach is similar to that previously discussed for the MIQ, in which the counselor determines the degree of match between the examinee's work values and the known rein-forcers available in various occupations.

The technical aspects of the WVI are commendable. In a sample of 99 tenth-grade students, the instrument revealed two-week test-retest reliabilities in the .80s for all scales except Prestige (r = .76). The manual provides extensive normative data for junior and senior high school students. Validity also looks strong, as judged by correlations with other measures of work values, factor analysis of scales and items, and theory-confirming relationships with external criteria. Bolton (1985) provides an excellent review of validity evidence for the WVI. Perhaps the only cautionary note about the WVI is that the instrument is now somewhat dated. New norms and reliability data need to be provided.

Values Scale

The Values Scale (VS) was developed by a consortium of researchers under the direction of Super and Nevill (1986) to assess 21 values relevant to work and life roles. The test consists of five items per value, each rated from 1 ("Of little or no importance") to 4 ("Very important"). A final item is used when the scale is administered cross-culturally for a total of 106 items. The values measured by the VS include the following:

Ability Utilization
Achievement
Advancement
Aesthetics
Altruism
Authority
Autonomy
Creativity
Economic Rewards
Life Style
Personal Development
Physical Activity
Prestige
Risk
Social Interaction
Social Relations
Variety
Working Conditions
Cultural Identity
Physical Prowess
Economic Security

An unusual but highly desirable aspect to the VS is that the test was developed explicitly for use in cross-cultural research. An informal consortium of research teams from dozens of countries in North America, Europe, Australia, Asia, and Africa was involved in the definition, revision, and refinement of values measured by the test. Each national team translated the test into its own language and pilot-tested the items.

The reliability of the VS is only fair, which is understandable since the instruments contain only five items per scale. Alpha coefficients are above .70 for all scales, and test-retest reliabilities are above .50 in college student samples. Norms are provided for a convenience sample of 3,000 U.S. students and adults. Initial validity studies are promising, but more studies are needed before the test is used for individual guidance (Rousseau, 1989: Slaney, 1989).

The Values Scale represents the very best in test development ideals. By involving dozens of research teams around the globe, Super and Nevill (1986) have conceived a test with true cross-cultural appeal and utility. Perhaps their efforts will help forge a global perspective on the nature and value of work. Too often, test development has been a parochial activity restricted to Western industrialized cultures. We can only hope that other test developers will also value the cross-cultural perspective in assessment.

Assessment of Career Development

Super (1957, 1990) has emphasized that career choice is not a discrete decision but a continuing process. He argues that vocational development is characterized by stages: growth, exploration, establishment, maintenance, and decline. In the growth stage, an individual entertains fantasies, develops interests, and discovers his or her capacities. 'During the exploration stage in adolescence and early adulthood, the individual engages in tentative examination of careers. This is followed by stabilization and consolidation of a career in the establishment phase. At the age of approximately 50, most individuals enter the maintenance stage, characterized by innovation and updating for some, but stagnation and deceleration for others. The decline stage features disengagement for most, but career specialization for a few.

In the beginning stages of career development— the growth and exploration stages—traditional vocational tests may not provide the best kinds of guidance information, since they are usually founded on the premise that the examinee is knowledgeable about work and has well-established interest patterns. However, it is typical of individuals in these stages to have limited information about careers and minimal knowledge of their vocational interests and values. In these situations, specialized instruments are needed for effective career assessment.

Several vocational measures are based upon a recognition that career choice is an ongoing process rather than a single decision. These alternative instruments focus upon maturity of career knowledge, vocational planning, and decision-making skills. Several representative career development and career maturity tests are mentioned briefly in Table 12.4. For a more extended discussion, the reader is urged to consult Walsh and Betz (1995).

INTEGRATIVE MODEL OF CAREER ASSESSMENT

Practitioners of career assessment rarely rely upon information from a single source such as a survey

Table 12.4 Representative Measures of Career Deveolpment

Career Directions Inventory
(Jackson, 2000)
Consisting of 100 triads of statements describing job-related activities, the examinee marks his/her most-preferred and least-preferred activity. The 15 basic interest scales include both work roles and work styles, e.g., administration, food service, sales, outdoors, writing, assertive, persuasive, and systematic. Excellent reliability and validity; norms based upon 12,000 individuals from more than 150 educational and occupational specialties.

Career Beliefs Inventory
(Krumboltz, 1999)
Comprised of 96 items rated on a five-point scale from "strongly agree" to "strongly disagree," this inventory is intended to identify client beliefs that may be blocking his/her career goals. Examples of the 25 scales include: career plans, acceptance of uncertainty, intrinsic satisfaction, control, approval of others.

Career Thoughts Inventory
(Sampson, Peterson, Lenz, Reardon, & Saunders, 1998)
Based upon the principles of cognitive therapy, the CTI is a self-administered, objectively scored measure of dysfunctional thinking in career problem solving and decision-making. The 48 items assess decision-making confusion, commitment anxiety, and external conflict.

Career Development Inventory
(Super, Thompson, Lindeman, and others, 1981)
A comprehensive measure of career development and maturity, the CDI consists of five subtests: Career Planning, which measures extent of, and engagement in, career planning; Career Exploration, which

evaluates current and prospective attempts to obtain career information; Decision Making, which measures ability to apply knowledge and insight to career planning; World of Work Information, a measure of knowledge of occupational structure; and Knowledge of the Preferred Occupational Group, which provides an in-depth assessment of knowledge about the examinee's single, preferred occupational group.

---

of interests or work values. The effective vocational counselor uses an integrative model in which information from interest, ability, and personality domains is considered simultaneously. Lowman (1991) has presented the elements of this approach, and much of our discussion is based upon his analysis. Practitioners of this method do not minimize the importance of interests in determining career choice and satisfaction. However, they tend to assign primary importance to ability patterns and certain personality characteristics in vocational assessment and guidance. We discuss ability patterns first.

Ability Patterns in Career Assessment

Depending upon the career goals of the client, several ability dimensions might be relevant to career assessment. A partial list includes general intelligence (g), mechanical and physical abilities, spatial analysis, verbal intelligence, investigative skills, artistic abilities, and social intelligence (Gottfredson, 1986; Lowman, 1991). The importance of broad or primary mental abilities such as spatial analysis or verbal intelligence is fairly obvious. For example, architecture requires high levels of spatial analysis for success. In a prospective architect, no amount of interest in the field can compensate for low ability in spatial analysis. Likewise, verbal abilities are essential for journalism and other professions that demand language proficiency. The relevance of specialized aptitudes such as mechanical abilities (for a prospective mechanic) and artistic abilities (for a prospective artist) is likewise straightforward. But what about social intelligence? Is this relevant to career assessment? Can social intelligence be measured?

First identified by Thorndike (1920b), social intelligence refers to the capacity to understand other people and to relate effectively to them. Although there has been much ongoing controversy about the validity of the social intelligence construct, recent studies indicate that simple paper-and-pencil measures can be used to isolate this dimension of ability from other aspects of intelligence (Lowman & Leeman, 1988). We will review two studies to illustrate this point.

Getter and Nowinski (1981) developed the Interpersonal Problem Solving Assessment Technique (IPSAT), a semistructured free-response test of interpersonal effectiveness. In this test, the respondent is presented with a series of 46 problematic interpersonal situations and asked to imagine being in each situation. Examinees are instructed to write alternative ways of handling each situation and to indicate which of these potential solutions they would actually choose. An example of a situation is as follows:

Your boss (or teacher) has just criticized a piece of work that you've done, and you think the criticism is unjustified and unfair. What do you do?

Based upon a detailed scoring manual, each response chosen by the examinee is scored in one of these categories: Effective, Avoidant, Inappropriate, Dependent, and unscorable. The grand total number of responses is first tallied to provide an index of the examinee's ability to think of alternative courses of action. Then, the number of chosen solutions scored in each category is counted, providing a profile of the types of solutions preferred by the examinee. Interscorer reliability of IPSAT subscales is quite high, and correlations with other instruments strongly support the convergent and discriminant validity of this instrument.

A more recent and promising inventory of social intelligence is the 128-item, true-false Social Relations Survey (SRS) developed by Lorr, Youniss, and Stefic (1991). They used a rational scale construction method buttressed with factor analysis to produce an instrument that measures eight factors of social intelligence. The subscales and illustrative items are depicted in Table 12.5. For 49 subjects retested after two weeks, the median test-retest reliability of the subscales is an impressive .89. Norms are provided for 260 college women and 355 high-school women. Several approaches to concurrent and construct validity indicate that the SRS provides a useful and valid approach to the self-report assessment of social skills.

Beyond a doubt, social intelligence is highly relevant to career guidance. For example, a prospective nurse will need high levels of social intelligence to function effectively on the job. In contrast, a computer technician may need little in the way of social skills to excel in the work environment. Low-man (1991) presents a hypothetical taxonomy of social intelligence to illustrate its relevance to career assessment (Table 12.6). Although the measurement of social intelligence remains a challenge, practitioners would be foolish to ignore this ability factor in career assessment.

Table 12.5 Scales and Illustrative Items from the Social Relations Survey

Social Assertiveness
I find it easy to talk with other people that I have just met. (T)

When I meet new people I usually let them bring up things to talk about. (F)

Directiveness
I am at my best when I am the person in charge. (T)

I am comfortable letting others take the lead in a group. (F)

Defense of Rights
If someone breaks in line in front of me, I speak up. (T)

I am uncomfortable returning merchandise to a store. (F)

Confidence
I feel confident most of the time. (T)

I feel dissatisfied with my abilities. (F)

Perceived Approval
I am sure that people I know think well of me (T)

I sometimes feel a sense of disapproval from others around me (F)

Expression of Positive Feelings
I like to show my positive feelings for others. (T)

I am uncomfortable showing affection for a friend in public (F)

Social Approval Need
I make a deliberate effort to make myself popular. (T)

I am unconcerned with what people say about me. (F)

Empathy
I am strongly affected when friends tell me about their problems. (T)

I usually maintain an objective and detached feeling toward others. (F)

Personality Patterns in Career Assessment

Several personality dimensions are also highly relevant to career assessment. Personality testing is discussed in detail in later chapters, so we will only mention a few occupationally relevant personality dimensions here:

- Need for achievement is important for persons with business and managerial aspirations (e.g., Orpen. 1983).
- Ascendance or dominance is also important for success in managerial ranks (e.g., Bentz, 1985).

- Emotional stability predicts positive performance in a wide range of traditional jobs, whereas neuroticism is associated with success in some artistic professions (e.g.. Wills. 1984).
- Masculinity and femininity differ significantly between various occupational groups (e.g.. Gough, 1987).

Research on the relevance of personality dimensions to career assessment is still in its infancy. Nonetheless, preliminary trends such as those previously noted clearly demonstrate the relevance of personality variables to job success. Practitioners are advised to consider occupationally relevant personality dimensions in career assessment. In sum, career assessment is a multifaceted enterprise that must take into account not just interests, but also ability patterns and personality traits as well.

Table 12.6 Hypothetical Taxonomy of Social Demands of Jobs

Degree of Social Involvement	Social Job Dimensions
Very high	Therapeutic, educational or management roles such as business manager, nurse, or psychotherapist; high degree of social interaction
High	Social contact is not always primary, e.g., college professor, social science researcher, moderate degree of social interaction
Moderate	Minimal social interaction but social facilitation needed e.g., high level executive
Slight	Minimal social interaction and minimal need for concern with feelings and reactions of others, e.g., clerk in discount department store'
Low	Very limited interaction with people and no requirement for therapeutic or influencing roles, e.g., laboratory scientist or novelist
Very low	Social skills not needed; the work setting may be unsociable, such as computer programmer

Attitudes and the Assessment of Moral Concepts

The previous topic focused on interests and values and the issues raised in their assessment. In this topic we continue the discussion of additional loosely defined but nonetheless useful constructs such as attitudes, moral values, and spiritual concepts. We begin with attitudes because this concept is foundational, and the methods used in the assessment of attitudes can be widely applied. The topic then turns to assessment approaches in the moral, spiritual, and religious domains. This includes lengthy coverage of Kohlberg's (1981, 1984) classic method for the measurement of moral reasoning. Finally, we close with brief coverage of the overlooked literature on the measurement of spiritual and religious concepts.

ATTITUDES AND THEIR ASSESSMENT

Throughout the history of psychology, the notion of attitude has played an essential role in the explanation of behavior. Gordon Allport (1935), an early pioneer in attitude research, characterized the concept of attitude as "distinctive" and "indispensable" to social psychology. The importance of attitude as a psychological construct has not diminished in recent years. For example, a search of PsychlNFO with the keyword "attitude" revealed more than 12,000 articles published from 1992 to 2002.

Attitudes are closely linked to related concepts such as values, opinions, and beliefs, so it is important to distinguish these terms (Aiken, 2002)., discussed earlier, a value is a shared and enduring idea about what is ideal; that is, a value refers what is ultimate or best. In contrast, an opinion can be regarded as the overt, conscious, verbal demonstration of an attitude. Aiken (2002) notes that opinions are "less central, more specific, mc changeable and more factually based" than attitudes. Also, opinions are expressed in words, whereas attitudes may not be. Finally, a belief is a conviction that something is true, even though it cannot be rigorously proved. Religious beliefs certainly fall into this category. Thus, a belief is somewhere between knowledge and attitude.

Having said what attitudes are not, we now I to a positive definition:

> *Attitudes may be viewed as learned cognitive, affective, and behavioral predispositions to respond positively or negatively to certain objects, situations, institutions, concepts, or persons. Attitudes may be quite individual and thereby reflective of and related to personality characteristics such as a need for closure. A need for closure is expressed as a desire to complete a task, as in finding an answer to a question or a solution to a problem, (p. 3)*

The central constituent of attitudes is that they always have an evaluative component to them—attitudes involve positive or negative responses some kind. But it is also clear that the expression of attitudes can be multifactorial (i.e., cognitive, affective, or behavioral). Also, an attitude always has an object—it is in reference to something. Examples of attitudinal objects include the death penalty, handguns, Republicans, former president William Clinton, cold weather, telemarketers, slow drivers, and a late start to the school year. Finally, attitudes serve motivational functions by helping individuals organize their perceptions and make sense out of the world.

Assessment of Attitudes

Obviously, an attitude is an unobservable, hypothetical construct. As such, it must be inferred from measurable responses indicating positive or negative evaluations of the attitudinal object. Attitudes may be inferred from cognitive responses (e.g., knowledge of the attitudinal object), behavioral responses (e.g., intentions or actions with respect to the object), and affective responses (expressed feelings toward the object). Even though attitudes can be inferred from all three sources, psychologists generally regard the affective response as the most essential aspect of attitudes (Ajzen, 1996).

Attitude measurement has followed a different path than assessment in other areas such as personality or intelligence. In these other areas, researchers typically try to develop a small number of definitive instruments that will become widely adopted in the field. In contrast, the typical approach in attitude assessment is for most researchers to develop their own unique instruments. This is because attitudes come in a virtually infinite supply, depending upon the attitudinal object of interest to researchers.

Approaches to Attitude Assessment

There are three broad approaches to the measurement of attitudes: behavioral, covert, and questionnaire. We discuss the behavioral and covert approaches briefly before reviewing the mainstay of attitude assessment—questionnaire approaches.

The behavioral approach involves the direct measurement of intentions or actions in regard to the attitudinal object. For example, if a door-todoor canvasser asks for donations to the Jennifer Jones for Mayor fund, the willingness to donate (as well as the amount) would be an index of attitudes toward Ms. Jones as a mayoral candidate. Other examples of the behavioral approach to attitude assessment would be asking people to sign a petition or to circulate flyers regarding a certain cause (e.g., building a new municipal swimming pool). Declining to sign the petition or to circulate flyers would indicate a negative attitude toward the cause, whereas willingness to do either would signify a positive attitude.

Covert approaches to attitude assessment involve unobtrusive procedures and measurements. For example, in the lost-letter technique (Schwartz & Ames, 1977), a researcher would prepare hundreds of envelopes, each with a new stamp, ostensibly addressed to different organizations. In reality, it is only the name of the agency that differs among the envelopes (e.g., some addressed to "Campaign to End Capital Punishment" others addressed to "Campaign to Promote Capital Punishment"), whereas the street address is actually that of the researcher. These letters are then surreptitiously dropped on busy sidewalks throughout the city. On the assumption that individuals will rescue (and mail) the letters that appear to support their views (and may discard the others), the relative return rates for the two kinds of letters is then an index of city-wide attitudes toward the concept of interest, for example, capital punishment.

The implicit association test (IAT) is another example of a covert measure of attitudes. In an implicit association test, the researcher uses reaction time to measure the automatic or "unconscious" associations of individuals to different target concepts. Greenwald, McGhee, and Schwartz (1998) explain the rationale for this approach by contrasting a hypothetical experiment with their real study. In the hypothetical experiment, the examinee is asked to view a series of male and female faces, saying "hello" if the face is male and "goodbye" if it is female. Of course, the responses are timed. For a second task, the participant says "hello" for male names and "goodbye" for female names. Finally, the two tasks are combined with the four kinds of stimuli presented in a random manner. Of course, this would be an easy task, and response times would be quite fast. Greenwald et al. (1998) then explain the rationale for their real study as follows:

> *One might appreciate the IAT's potential value as a measure of socially significant automatic associations by changing the thought experiment to one in which the to-be-distinguished faces of the first task are Black or White (e.g. "hello" to African American faces and "goodbye" to European American faces) and the second task is to classify words as pleasant or unpleasant in meaning ("hello" to pleasant words, "goodbye" to unpleasant words). The two possible combinations of these tasks can be abbreviated as Black + pleasant and White + pleasant. Black + pleasant should be easier [faster reaction times-] than White + pleasant if there is a stronger association between Black Americans and pleasant meaning than between White Americans and pleasant meaning. If the preexisting associations are opposite in direction—which might be expected for White subjects raised in a culture imbued with pervasive residues of a history of anti-Black discrimination—the subject should find White + pleasant to be easier, (p. 1466)*

In an actual IAT study, respondents press computer keys rather than verbalizing their responses, permitting accurate timing to a millisecond. The advantage of the IAT approach is that it presumably gets around the social desirability bias encountered with paper-and-pencil measures. The procedure is designed to reveal attitudes—even when participants prefer not to express these attitudes. Currently, the IAT approach is used mainly for research to test theories in social psychology.

Psychophysiological measurements also have been used to assess attitudes. For example, pupillary response can be measured by unobtrusive cameras aimed at the pupil of the viewer's eye as he or she looks at different pictures. This is the science of pupillometrics (measurement of pupil size). When other factors are held constant (e.g., background light), a larger pupil is presumed to indicate a greater interest in the observed picture (Hess & Polt, 1960).

Pupillometrics reached its high point with the publication of The Tell-Tale Eye: How Your Eyes Reveal Hidden Thoughts and Emotions (Hess, 1975). In this book, Hess recounts an intriguing application of pupillometrics to advertising. A large number of observers looked at two different advertisements for Encyclopaedia Britannica. One was a new ad showing boys in a pool, the other a standard ad depicting a wholesome family scene. Based on a questionnaire, the observers expressed a preference for the new ad (a more favorable attitude). However, their pupils did not dilate at all for the new ad, whereas they dilated significantly for the standard ad. The two ads were placed in different copies of a magazine, together with a coded reply card. The two versions of the magazine sold approximately the same number of copies. However, the return rate for reply cards sent with the standard ad was far higher than for the new ad. Thus, pupillometrics predicted apparent attitudes much better than a traditional questionnaire technique. Even though this experiment was plagued with methodological weaknesses, the findings did serve to popularize the use of pupillometrics as an attitudinal measure. However, these techniques are expensive and therefore inefficient when the goal is to assess attitudes in a large group of individuals. Another concern is that pupil enlargement may signify not just a positive attitude, but also arousal or novelty of the stimulus picture.

Questionnaires in Attitude Assessment

The vast majority of attitude measures are questionnaires based upon established scaling methods. The reader will recall from Topic 4B (Test Construction) that a variety of scaling methods are available, including the method of equal-appearing intervals, the method of absolute scaling, the Likert scale, and the Guttman scale. Without a doubt, the Likert scale is the most popular in attitude measurement. In this approach, the examinee is offered five (or sometimes seven) responses ordered on an agree/disagree continuum. For example, one item on a scale to assess attitudes toward death might read:

It makes me anxious when people talk about death.
Do you:

II	II	II	II	II
Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree

In a Likert scale, the total score is obtained by adding up the scores (1 to 5) for individual items. Of course, scoring is reversed for negatively phrased items.

By definition, an attitude measure is supposed to tap a highly homogeneous construct, especially insofar as the affective response (extent of positive or negative feelings about the attitudinal object) is central to attitudes. For this reason, the most important psychometric quality of an attitude measure is that it should possess strong internal consistency as measured by coefficient alpha or a related index. In regard to validity of attitude measures, an important point is that attitudes are highly robust human characteristics. Thus, attitude scales measuring similar constructs should correlate very highly, even when the scales are developed by different researchers (Davis & Ostrum, 1996).

The characteristics of a good attitude-related measure can be illustrated with a specific instrument, the Gratitude Questionnaire-Six Item Form (GQ-6; McCullough, Emmons, & Tsang, 2002). The GQ-6 is a simple self-report measure of the disposition to experience gratitude (Figure 12.4). Strictly speaking, the GQ-6 is a trait measure of the grateful disposition. However, this trait is affective in nature; therefore the instrument fittingly illustrates the concepts and issues involved in developing an attitude measure.

The reader will notice that the GQ-6 is based on a Likert-type format with seven alternatives ranging from 1 (strongly disagree) to 7 (strongly agree). Two items are stated in the reverse (and therefore reverse scored) as a way of inhibiting response bias. The development and choice of specific test items was based on a thorough analysis of the many facets of the grateful disposition (McCullough et al., 2002). The authors determined that gratitude reflects intensity (feeling more intensely grateful), frequency (feeling grateful many times a day), span (grateful for many things), and density (grateful to many individuals). Initially, they proposed 39 items to measure these qualities. The GQ-6 is composed of the six best items, as determined by factor-analytic procedures performed with test results from two samples: 238 undergraduates and 1.228 adult volunteers surveyed via the Internet. Reliability of the instrument is good, with coefficient alphas

Using the scale below as a guide, write a number beside each statement to indicate how much you agree with it.

1 = strongly disagree
2 = disagree
3 = slightly disagree
4 = neutral
5 = slightly agree
6 = agree

7 = strongly agree

1. ________ I have so much in life to be thankful for.
2. ________ If I had to list everything that I felt grateful for, it would be a very long list.
3. ________ When I look at the world, I don't see much to be grateful for.
4. ________ I am grateful to a wide variety of people.
5. ________ As I get older I find myself more able to appreciate the people, events, and situations that
have been part of my life history.
6. ________ Long amounts of time can go by before I feel grateful to something or someone.

between .82 and .87. Validity of the GQ-6 is based upon numerous theory-confirming relationships with other measures. For example, self-ratings on the GQ-6 correlated modestly with external observers' perceptions of gratitude in the participants. Additional substudies indicated that the GQ-6 is positively related to optimism, hope, spirituality and religiousness, forgiveness, empathy, and prosocial behavior. The scale is negatively related to depression, anxiety, materialism, and envy.

Literally thousands of attitude measures have been proposed. Aiken (2002) provides information on dozens of carefully validated instruments. An Internet search using the phrase "Attitude Measures" revealed 647,000 sources, many citing unpublished instruments. Table 12.7 lists a sampling of the attitudinal objects surveyed by some of these measures.

Issues in Attitude Assessment

One of the major issues in attitude assessment is whether attitudes predict behavior. The literature on this topic is vast, and the findings are complex and multifaceted. In an early, classic study, LaPiere (1934) established that motel and restaurant owners in the United States answered attitude questionnaires one way, but behaved in another. Specifically, when these individuals were asked questions such as, "Will you accept persons of the Chinese race as guests," they answered yes, but when the researcher sent Chinese patrons to their establishments, they were refused service. Many other studies also point to a weak link, at best, between attitude measures and behavior (Aiken, 2002).

Table 12.7 Attitudinal Objects Surveyed by Unpublished Scales

Attitudes Toward…

Alcohol use among students
Body parts
Childhood illness
Christianity
Contraception
Electroconvulsive therapy
Engineering as a vocation

Gay/lesbian parenting
Herpes
HIV infection

Homeless people
Insurance
Late work arrival
Librarians
Organ donation
Overweight people
Physician-assisted suicide

Psychotherapy for self
School bullying
School subjects

More recently, researchers have focused on ways to increase the predictive validity of attitude measures. One genera] theme of this research is that attitudes will be predictive if they are strongly activated

(Greenwald & Banaji, 1995). Another finding is that attitudes will be predictive if the actor is highly conscious of them (Myers, 2002). A recent review of attitude research can be found in Ajzen (2001).

THE ASSESSMENT OF MORAL JUDGMENT

The Moral Judgment Scale

Kohl berg has proposed one of the few theories of moral development that is both comprehensive and empirically based (Colby, Kohlberg, Gibbs, & Lieberman, 1983; Kohlberg, 1958, 1981, 1984; Kohlberg & Kramer, 1969). Although he was more concerned with theory-based problems of moral development than with the nuances of standardized measurement, Kohlberg did generate a method of assessment that is widely used and intensely debated. We will review the underlying rationale for his measurement tool and discuss the psychometric properties of the instrument as well. In addition, we will take a brief look at a more objectively based adaptation of Kohlberg's approach known as the Defining Issues Test (Rest, 1979, Rest & Thoma, 1985).

Stages of Moral Development

Kohlberg's theory grew out of Piaget's (1932) stage theory of moral development in childhood. Kohlberg extended the stages into adolescence and adulthood. In order to explore reasoning about difficult moral issues, he devised a series of moral dilemmas. One of the most famous is the dilemma of Heinz and the druggist:

In Europe, a woman was near death from a special kind of cancer. There was one drug that the doctors thought might save her. It was a form of radium that a druggist in the same town had recently discovered. The drug was expensive to make, but the druggist was charging ten times what the drug cost him to make. He paid $200 for the radium and charged $2000 for a small dose of die drug. The sick woman's husband, Heinz, went to everyone he knew to borrow the money, but he could only get together about $1000 which is half of what it cost. He told the druggist that his wife was dying, and asked him to sell it cheaper or let him pay later. But the druggist said, "No, I discovered the drug and I'm going to make money from it." So Heinz got desperate and broke into the man's store to steal the drug for his wife. (Kohlberg & Elfenbein, 1975)

After reading or hearing this story, the respondent is asked a series of probing questions. The questions might be as follows: Should Heinz have stolen the drug? What if Heinz didn't love his wife? Would that change anything? What if the person dying was a stranger? Should Heinz steal the drug anyway? Based on answers to this and other dilemmas, Kohlberg concluded that there are three main levels of moral reasoning, with two substages within each level (Table 12.8). One use of his measurement instrument, the Moral Judgment Scale, is to determine a respondent's stage of moral reasoning.

The Moral Judgment Scale consists of several hypothetical dilemmas such as Heinz and the druggist, presented one at a time (Colby, Kohlberg, Gibbs, & others, 1978). In its latest revision, the Scale comes in three versions called Forms A, B, and C. Scoring is quite complex, based on the examiner's judgment of responses in relation to extensive criteria outlined in a detailed scoring manual (Colby & Kohlberg, 1987). Although there are several different dimensions to scoring, the one element most frequently cited in research studies is the overall stage of moral reasoning that characterizes a respondent.

Critique of the Moral Judgment Scale

Early versions of the Moral Judgment Scale suffered serious shortcomings of scoring and interpretation. For example, in his doctoral dissertation, Kohlberg (1958) proposed two scoring systems: one using the sentence or completed thought as the unit of scoring, the other relying upon a global rating of all the subject's utterances as the unit of analysis. Neither approach was fully satisfactory and early reviews of the scale were justifiably critical of its reliability and validity (Kurtines & Greif, 1974).

In response to these criticisms, Kohlberg and his associates developed a scoring system that is unparalleled in its clarity, detail, and sophistication (Rest, 1986). Fortuitously, since the moral dilemmas of the Moral Judgment Scale have remained constant over the years, it is possible to apply the new scoring system to old data. The capacity to reanalyze old data and compare it with new data is invaluable in determining the reliability and validity of an existing scale. A most important study in this regard has been published by Kohlberg and associates (Colby et al., 1983).

Table 12.8 Kohlberg's Levels and Stages of Moral Development

Level 1: Preconventional
Stage 1. Punishment and obedience orientation: The physical consequences determine what is good or bad.
Stage 2. Instrumental relativism orientation: What satisfies one's own needs is good.

Level 2: Conventional
Stage 3. Interpersonal concordance orientation: What pleases or helps others is good.
Stage 4. "Law-and-order" orientation: Maintaining the social order and doing one's duty is good.

Level 3: Postconventional or Principled
Stage 5. Social contract-legalistic orientation: Values agreed upon by society determine what is good.
Stage 6. Universal ethical-principle orientation: What is right is a matter of conscience derived from universal principles.

This investigation reports the results of using the new scoring system in a longitudinal study spanning more than 20 years. The results are impressive and offer strong support for the reliability and validity of the instrument. Test-retest correlations for the three forms were in the high .90s as were interrater correlations. Longitudinal scores of subjects tested at three- to four-year intervals over 20 years revealed theory-consistent trends. Fifty-six of 58 subjects showed upward change, with no subjects skipping any stages. Furthermore, only 6 percent of the 195 comparisons showed backward shifts between two testing sessions. The internal consistency of scores was also excellent: about 70 percent of the scores were at one stage, and only 2 percent of the scores were spread further than two adjacent stages. Cronbach's alpha was in the mid-. 90s for the three forms. These findings have been corroborated by Nisan and Kohlberg (1982). Heilbrun and Georges (1990) also report favorably upon the validity of the Moral Judgment Scale, insofar" as postconventional development is correlated with higher levels of self-control, as would be predicted from the fact that morally mature persons often oppose social pressure or legal constraints. In sum, the Moral Judgment Scale is reliable, internally consistent, and possesses a theory-confirming developmental coherence.

The Defining Issues Test

The Defining Issues Test (DIT) is similar to the Moral Judgment Scale, but incorporates a much simpler and completely objective scoring format (Rest, 1979, 1986). The examinee reads a series of moral dilemmas similar to those designed by Kohlberg, and then chooses a proper action for each. For example, one dilemma involves a patient dying a painful death from cancer. In her lucid moments, she requests an overdose of morphine to hasten her death. What should the doctor do? Three options of the following kind are listed:

__________ He should give the woman a fatal overdose
__________ Should not give the overdose
__________ Can't decide

The examinee's choice does not enter directly into the determination of the moral judgment score. The real purpose in forcing a choice is to cause the examinee to think about the importance of various factors in making the decision. Following the choice of proper action, the examinee rates the importance of several factors on a five-point Likert scale: great, much, some, little, or no importance. The factors are distinct for each dilemma. The factors differ in the level of moral judgment they signify, ranging from Kohlberg's stage 1 through stage 6. In the case of the preceding dilemma, the factors include such matters as follows:

__________Whether the doctor can make it look like an accident
__________Can society afford to let people end their lives when they want to
__________Whether the woman's family favors giving the overdose or not

These ratings form the basis for generating several quantitative scores that pertain to the moral judgment of the examinee. The most widely used score is the P score, which is a percentage of principled thinking. Reliability of the P score ranges from .71 to .82 in test-retest studies (Rest. 1979, 1986). Validity has been studied by contrasting group known to differ on principled thinking. For example, graduate students in moral philosophy and political science, general college students, high school seniors, and ninth-grade students were found to differ appropriately and systematically on the P score. In longitudinal studies, significant upward trends were found over six years and four testings. Recently, Rest has recommended a new measure moral judgment, the N2 index, calculated on the basis of several complex formulas that use both ranking and rating data. The two indices are highly correlated in the .90s. Nonetheless, in a retrospective analysis of previous studies, the N2 index performed the P index by a substantial margin (Rest, Thoma, Narvaez, & Bebeau, 1997).

Over 600 articles have been published on the Defining Issues Test (McCrae, 1985; Morelanc 1985; Sutton, 1992). In general, the instrument considered a useful alternative to Kohlberg's Moral Judgment Scale, particularly for research on group differences in moral reasoning. However, reviewer sdo note several cautions about the DIT (Sutton, 1992; Westbrook & Bane, 1992). First, the test us two moral dilemmas from the Vietnam War and: therefore somewhat dated. Many young examine have little knowledge of (and perhaps no interest in) this topic and may find it difficult to identify with these questions. Another dilemma —the classic case of whether Heinz should steal a drug to save his wife's life—is also of dubious value since it ha been widely publicized and reprinted in college textbooks. A significant proportion of prospective examinees are no longer naive about this moral dilemma.

Richards and Davison (1992) have pressed the-point that the DIT is biased against conservativel; religious individuals. Certainly, it is well established that conservative or fundamentalist religious people tend to score lower than average on the P score of the Defining Issues Test (Getz, 1984; Richards, 1991). According to Richards and Davison (1992), the reason for this is that stage 3 and stage 4 items (unintentionally) possess strong theological implications that cause fundamentalist individuals to endorse the items, thereby lowering their score on the test. Consider items that tap stage 4 reasoning, which is the "law and order" orientation that equates "moral" with doing one's duty and maintaining the social order. Whereas nonreligious persons might support the laws of the land (and endorse stage 4 items) because they believe that legal authorities define what is right and moral, religious minorities such as Mormons believe that supporting the laws of the land is a theological and religious obligation that flows directly from articles of faith in their religion:

> *While Mormons place a high value on obeying the law and supporting legal authorities, this value is due to their theological belief that God has commanded them to do so, and not because they believe, as do true Stage 4 thinkers, that the laws of the land or legal authorities define what is right or moral. (Richards & Davison, 1992, 470)*

These researchers demonstrate empirically that certain DIT items measure a different construct for conservative religious persons than for the general population. As a consequence, the validity of the test in these groups is open to question.

A related criticism of the DIT is the dearth of norms pertinent to minority groups. Finally, Westbrook and Bane (1992) argue that the technical manual for the DIT lacks essential details needed to evaluate the adequacy of the test. In spite of these criticisms, the DIT is a widely respected test, particularly for research on moral reasoning.

Lesson 13

PERSONALITY TESTING

Origins of Personality Testing

Theories and the Measurement of Personality

In psychological testing a fundamental distinction often is drawn between ability tests and personality tests. Defined in the broadest sense, ability tests include the plethora of instruments for measuring intelligence, achievement, aptitude, and neuropsychological functions. In the preceding 12 chapters we have explored the nature, construction, application, reliability, and validity of these instruments. In the next two chapters we shift the emphasis to personality tests. Personality tests seek to measure one or more of the following: personality traits, dynamic motivation, personal adjustment, psychiatric symptomatology, social skills, and attitudinal characteristics. This chapter investigates the origins of personality testing. In Topic 13 A, Theories and the Measurement of Personality, the different ways in which researchers have conceptualized personality are surveyed to illustrate how their theories have impacted the design of personality tests and assessments. In Topic 13B, Projective Techniques, we examine the multiplicity of instruments based upon the turn-of-the-twentieth-century psychoanalytic hypothesis that responses to ambiguous stimuli reveal the innermost, unconscious mental processes of the examinee. The coverage of personality assessment continues in the next chapter with a review of objective tests and procedures, including self-report inventories and behavioral assessment approaches.

PERSONALITY: AN OVERVIEW

Although personality is difficult to define, we can distinguish two fundamental features of this vague construct. First, each person is consistent to some extent; we have coherent traits and action patterns that arise repeatedly. Second, each person is distinctive to some extent; behavioral differences exist between individuals. Consider the reactions of three graduate students when their midterm examinations were handed back. Although all three students received nearly identical grades (solid Bs), personal reactions were quite diverse. The first student walked off sullenly and was later overheard to say that a complaint to the departmental administrator was in order. The second student was pleased, stating out loud that a B was, after all, a respectable grade. The third student was disappointed but stoical. He blamed himself for not studying harder.

How are we to understand the different reactions of these three persons, each of whom was responding to an identical stimulus? Psychologists and laypersons alike invoke the concept of personality to make sense out of the behavior and expressed feelings of others. The notion of personality is used to explain behavioral differences between persons (for example, why one complains and another is stoical) and to understand the behavioral consistency within each individual (for example, why the complaining student noted previously was generally sour and dissatisfied).

In addition to understanding personality, psychologists also seek to measure it. Literally hundreds of personality tests are available for this purpose; we will review historically prominent instruments and also discuss some promising new approaches. However, in order that the reader can better comprehend the diversity of instruments and approaches, we begin with a more fundamental question: How is personality best conceptualized?

As the reader will discover, in order to measure personality we must first envision what it is we seek to measure. The reader will better appreciate the multiplicity of tests and procedures if we also briefly describe the personality theories which comprise the underpinnings for these instruments. We close out this topic by raising a general question pertinent to all theories and testing approaches: How stable and predictable is behavior?

Although we partition personality tests separately from the ability tests, the distinction between these two kinds of instruments is far from absolute. Intellectual ability is, in part, a characterological feature based on such attributes as perseverance and self-control. Thus, ability tests inevitably tap important dimensions of personality, albeit in an indirect and imperfect manner. Often, the converse is also true: Personality tests may be saturated with ability factors. For example, certain personality dimensions such as openness to experience probably correlate positively with intelligence. As the reader will discover in the next chapter, some true-false personality inventories incorporate a very robust intelligence factor (e.g., Cattell, Eber, & Tatsuoka, 1970).

PSYCHOANALYTIC THEORIES OF PERSONALITY

Psychoanalysis was the original creation of Sigmund Freud (1856-1939). While it is true that many others have revised and adapted his theories, the changes have been slight in comparison to the substantial foundations that can be traced to this singular genius of the Victorian and early twentieth century era. Freud was enormously prolific in his writing and theorizing. We restrict our discussion to just those aspects of psychoanalysis that have influenced psychological testing. In particular, the Rorschach, the Thematic Apperception Test, and most of the projective techniques critiqued in the next topic dictate a psychoanalytic framework for interpretation. Readers who wish a more thorough review of Freud's contributions can start with the New Introductory Lectures on Psychoanalysis (Freud, 1933). Reviews and interpretations of Freud's theories can be found in Stafford-Clark (1971) and Fisher and Greenberg (1984).

Origins of Psychoanalytic Theory

Freud began his professional career as a neurologist, but was soon specializing in the treatment of hysteria, an emotional disorder characterized by histrionic behavior and physical symptoms of psychic origin such as paralysis, blindness, and loss of sensation. With his colleague Joseph Breuer, Freud postulated that the root cause of hysteria was buried memories of traumatic experiences such as childhood sexual molestation. If these memories could be brought forth under hypnosis, a release of emotion called abreaction would take place and the hysterical symptoms would disappear, at least briefly (Studies on Hysteria, Breuer & Freud, 1893-1895).

From these early studies Freud developed a general theory of psychological functioning with the concept of the unconscious as its foundation. He believed that the unconscious was the reservoir of instinctual drives and a storehouse of thoughts and wishes that would be unacceptable to our conscious self. Thus, Freud argued that our most significant personal motivations are largely beyond conscious awareness. The concept of the unconscious was discussed in elaborate detail in his first book (The Interpretation of Dreams, Freud, 1900). Freud believed that dreams portray our unconscious motives in a disguised form. Even a seemingly innocuous dream might actually have a hidden sexual or aggressive meaning, if it is interpreted correctly.

Freud's concept of the unconscious penetrated the very underpinnings of psychological testing early in the twentieth century. An entire family of projective techniques emerged, including inkblot tests, word association approaches, sentence completion techniques, and story-telling (apperception) techniques (Frank, 1939, 1948). Each of these methods was predicated on the assumption that unconscious motives could be divined from an examinee's responses to ambiguous and unstructured stimuli. In fact, Rorschach (1921) likened his inkblot test to an X ray of the unconscious mind. Although he patently overstated the power of projective techniques, it is evident from Rorschach's view that the psychoanalytic conception of the unconscious had a strong influence on testing practices.

The Structure of the Mind

Freud's views on the structure of the mind and the operation of defense mechanisms also influenced psychological testing and assessment (New Introductory Lectures on Psychoanalysis, Freud, 1933). Several tests and assessment approaches discussed in this chapter are predicated upon the psychoanalytic conception of defense mechanisms, so this topic deserves brief summary.

Freud divided the mind into three structures: the id, the ego, and the superego. The id is the obscure and inaccessible part of our personality that Freud likened to "a chaos, a cauldron of seething excitement." Because the id is entirely unconscious, we must infer its characteristics indirectly by analyzing dreams and symptoms such as anxiety. From such an analysis, Freud concluded that the id is the seat of all instinctual needs such as for food, water, sexual gratification, and avoidance of pain. The id has only one purpose, to obtain immediate satisfaction for these needs in accordance with the pleasure principle. The pleasure principle is the impulsion toward immediate satisfaction without regard for values, good or evil, or morality. The id is also incapable of logic and possesses no concept of time. The chaotic mental processes of the id are therefore unaltered by the passage of time, and impressions that have been pushed down into the id "are virtually immortal and are preserved for whole decades as though they had only recently occurred" (Freud, 1933).

If our personality consisted only of an id striving to gratify its instincts without regard for reality, we would soon be annihilated by outside forces. Fortunately, soon after birth part of the id develops into the ego or conscious self. The purpose of the ego is to mediate between the id and reality. The ego is part of the id and servant to it, but the ego "interpolates between desire and action the procrastinating factor of thought" (Freud, 1933). Thus, the ego is largely conscious and obeys the reality principle; it seeks realistic and safe ways of discharging the instinctual tensions which are constantly pushing forth from the id.

The ego must also contend with the superego, the ethical component of personality that starts to emerge in the first five years of life. The superego is roughly synonymous with conscience and comprises the societal standards of right and wrong that are conveyed to us by our parents. The superego is partly conscious; but a large part of it is unconscious; that is, we are not always aware of its existence or operation. The function of the superego is to restrict the attempts of the id and ego to obtain gratification. Its main weapon is guilt, which it uses to punish the wrongdoings of the ego and id. Thus, it is not enough for the ego to find a safe and realistic way for the gratification of id strivings. The ego must also choose a morally acceptable outlet, or it will suffer punishment from its overseer, the superego. This explains why we may feel guilty for immoral behavior such as theft even when getting caught is impossible. Another part of the superego is the ego ideal, which consists of our aims and aspirations. The ego measures itself against the ego ideal and strives to fulfill its demands for perfection. If the ego falls too far short of meeting the standards of the ego ideal, a feeling of guilt may result. We commonly interpret this feeling as a sense of inferiority (Freud, 1933).

The Role of Defense Mechanisms

The ego certainly has a difficult task, acting as mediator and servant to three tyrants: id, superego, and external reality. It may seem to the reader that the task would be essentially impossible and that the individual would therefore be in a constant state of anxiety. Fortunately, the ego has a set of tools at its disposal to help carry out its work, namely, mental strategies collectively labeled defense mechanisms.

Defense mechanisms come in many varieties, but they all share three characteristics in common. First, their exclusive purpose is to help the ego reduce anxiety created by the conflicting demands of id, superego, and external reality. In fact, Freud felt that anxiety was a signal telling the ego to invoke one or more defense mechanisms in its own behalf. Defense mechanisms and anxiety are therefore complementary concepts in psychoanalytic theory, one existing as a counterforce to the other. The second common feature of defense mechanisms is that they operate unconsciously. Thus, even though defense mechanisms are controlled by the ego, we are not aware of their operation. The third characteristic of defense mechanisms is that they distort inner or outer reality. This property is what makes them capable of reducing anxiety. By allowing the ego to view a challenge from the id, superego, or external reality in a less-threatening manner, defense mechanisms help the ego avoid crippling levels of anxiety. Of course, because they distort reality, the rigid, excessive application of defense mechanisms may create more problems than it solves.

Assessment of Defense Mechanisms and Ego Functions

Although Freud introduced the concept of defense mechanisms, it was left to his followers to elucidate these unconscious mental strategies in more detail (Paulhus, Fridhandler, & Hayes, 1997). An early portrayal of defense mechanisms was provided by Freud's daughter, Anna (The Ego and the Mechanisms of Defense, A. Freud, 1946). However, the application of these concepts to psychological measurement and assessment is much more recent. For example, Loevinger (1976, 1979, 1984) has produced a sentence completion technique for measuring ego development that is based, indirectly, on the analysis of defense mechanisms. This interesting approach to personality measurement is outlined briefly in the next unit. Here we will present Vaillant's (1977, 1992) work to illustrate the measurement of defense mechanisms and the application of this information to the understanding of personality.

Vaillant (1971) developed a hierarchy of ego adaptive mechanisms based on the assumption that some defensive mechanisms are intrinsically healthier than others. In his view, defense mechanisms can be grouped into four different types. Listed in order of increasing healthiness, the types are psychotic, immature, neurotic, and mature (Table 13.1). Psychotic mechanisms such as gross denial of external reality are the least healthy because they distort reality to an extreme degree. They appear "crazy" to the beholder. Immature mechanisms such as the projection of one's own unacknowledged feelings to others are healthier than psychotic mechanisms. Nonetheless, they are easily detected by outside observers and seen as undesirable. Neurotic defense mechanisms typically alter private feelings so that they are less threatening. An example is intellectualization, a defense mechanism in which threatening matters are analyzed in bland terms that are void of feelings. For example, a physician whose mother died recently might talk at great length about the medical characteristics of her cancer, thereby easing his sense of loss. Mature mechanisms of defense appear to the beholder as convenient virtues. An example is certain forms of humor which do not distort reality but which case ease the burden of matters "too terrible to be borne" (Vaillant, 1977).

The application of defense mechanisms to the understanding of personality is illustrated in the Grant Study, a 45-year follow-up study conducted by Vaillant and others (Vaillant, 1977; Vaillant & Vaillant, 1990). These researchers used structured interviews to obtain evidence of unconscious adaptive mechanisms from a sample of 95 men. The subjects were from an original sample of 268 students from Harvard University's classes of 1939 through 1944. At follow-up, Vaillant interviewed each participant for two hours, using a semi structured interview schedule (Vaillant, 1977, App. B). In addition, the subjects filled out autobiographical questionnaires and provided other sources of information. The entire protocol for each subject was then evaluated by Vaillant and other raters according to the extent that each defense mechanism characterized the individual's adaptation to life. Defense mechanisms were scored from 1 (absent) to 5 (major). Here is an example of one unconscious adaptive behavior:

> *A California hematologist developed a hobby of cultivating living cells in test tubes. In a recent interview, he described with special interest and animation an unusually interesting culture that he had grown from a tissue biopsy from his mother. Only toward the end of the interview did he casually reveal that his mother had died from a stroke only three weeks previously. His mention of her death was as bland as his description of her still-living tissue culture had been affectively colored. Ingeniously and unconsciously, he had used his hobby and his special skills as a physician to mitigate temporarily the pain of his loss. Although his mother was no longer alive, by shifting his attention he was still able to care for her. There was nothing morbid in the way he told the story; and because ego mechanisms are unconscious, he had no idea of his defensive behavior. Many of the healthiest men in the Study used similar kinds of attention shifts or displacement. Unless specifically looked for by a trained observer, such behavior goes unnoticed more often than not. (Vaillant, 1977)*

Most likely, this individual would receive a rating of 5 (major) for the neurotic defense mechanism of displacement.

Considering the degree of skilled judgment required by the evaluation task, the interrater reliability of the defense mechanism ratings was—with a few exceptions—respectable. The individual defense mechanisms possessed reliabilities that ranged from .53 (Fantasy) to .96 (Projection); most reliabilities were in the .70s and .80s. Reliability of a global rating (reflecting the ratio of mature to immature ratings) was .77.

The validity of defense mechanism ratings hinges mainly on the demonstration that developmental changes and group differences are consistent with psychoanalytic theory regarding these constructs. We would expect, for example, that the Grant Study subjects would use fewer immature and more mature defense mechanisms as they grew into middle age, and this is precisely what Vaillant discovered. In addition, we would expect that persons found to be maladjusted by other criteria (e.g., frequent divorce, underachievement) would rate less favorably on defense mechanisms in comparison to adjusted persons, and this is also what Vaillant observed. In sum, the analysis of defense mechanisms is a promising approach to personality assessment. However, this approach does have two drawbacks: The examiner needs specialized training to recognize defense mechanisms, and the process of collecting relevant information from examinees is very time-consuming.

Table 13.1 Levels of Defense Mechanisms Proposed by Vaillant (1977)

I. Psychotic

Delusional Projection: frank delusions about external reality, usually of a persecutory nature
Denial: denial of external reality; e.g., failing to acknowledge that one has a terminal illness
Distortion: grossly reshaping external reality to suit inner needs; e.g., wish-fulfilling delusions

II. Immature

Projection: attributing one's own unacknowledged feelings to others; e.g., "You're angry, not me!"
Schizoid Fantasy: use of fantasy and inner retreat for the purpose of conflict resolution and gratification
Hypochondriasis: transforming reproach toward others first into self-reproach then into complaints of physical illness
Passive-Aggressive Behavior: aggression toward others expressed indirectly and ineffectively through passivity or directed against the self
Acting Out: direct expression of an unconscious wish or impulse in order to avoid being conscious of the feeling that accompanies it

III. "Neurotic"

Intellectualization: thinking about wishes in formal, unfeeling terms, but not acting upon them
Repression: seemingly inexplicable memory lapses or failure to acknowledge information; e.g., "forgetting" a dental appointment
Displacement: directing of feelings toward something or someone other than the real object; e.g., kicking the dog when angry with the boss
Reaction Formation: unconsciously turning an impulse into its opposite; e.g., over-solicitousness to a hated coworker
Dissociation: temporary but drastic modification of one's character to avoid emotional distress: e.g. a brief devil-may-care attitude

IV. Mature

Altruism: vicarious but constructive and gratifying service to others; e.g., philanthropy
Humor: playful acknowledgment of ideas and feelings without discomfort and without unpleasant effects on others; does not include sarcasm
Suppression: conscious or semiconscious decision to postpone paying attention to a conscious conflict or impulse
Anticipation: realistic anticipation of or planning for future inner discomfort; e.g., realistic anticipation of surgery or separation
Sublimation: indirect expression of instinctual wishes without adverse consequences or loss of pleasure; e.g., channeling aggression into sports

TYPE THEORIES OF PERSONALITY

The earliest personality theories attempted to sort individuals into discrete categories or types. For example, the Greek physician Hippocrates (ca. 460-377 B.C.) proposed a humoral theory with four personality types (sanguine, choleric, melancholic, and phlegmatic) that was too simplistic to be useful. In the 1940s, Sheldon and Stevens (1942) proposed a type theory based upon the relationship between body build and temperament. Their approach stimulated a flurry of research and then faded into obscurity. Nonetheless, typological theories have continued to capture intermittent interest among personality researchers. We will illustrate type theories by reviewing contemporary research on coronary-prone personality types.

Type A Coronary-Prone Behavior Pattern

Friedman and Rosenman (1974) investigated the psychological variables that put individuals at higher risk of coronary heart disease. They were the first to identify a Type A coronary-prone behavior pattern, which they described as "an action-emotion complex that can be observed in any person who is aggressively involved in a chronic, incessant struggle to achieve more and more in less and less time, and if required to do so, against the opposing efforts of other things or persons" (Friedman & Rosenman, 1974). At the opposite extreme is the Type B behavior pattern, characterized by an easygoing, noncompetitive, relaxed lifestyle. Of course, people vary along a continuum from "pure" Type A to "pure" Type B.

Friedman and Ulmer (1984) have listed the specific components of the full-fledged Type A behavior pattern:

- **Insecurity of status:** A hidden lack of self-esteem seems to plague many Type A persons. No matter how successful, they often compare themselves unfavorably to other superachievers.
- **Hyperaggressiveness:** A desire to dominate others and damage their self-esteem is part of the pattern. Type A persons are often indifferent to the feelings or rights of competitors.
- **Free-floating hostility:** The Type A person finds too many things to get upset about, and the anger is out of proportion to the situation.
- **Sense of time urgency (hurry sickness):** This includes two basic strategems: speeding up daily activities (one Type A used an electric shaver in each hand!), and doing two things at once such as conversing on the phone while reviewing correspondence.

Type A behavior can be diagnosed from a short interview consisting of questions about habits of working, talking, eating, reading, and thinking (Friedman, 1996). The more flagrant cases of Type A behavior can also be detected by paper-and-pencil tests (Jackson & Gray, 1987; Jenkins. Zyzanski, & Rosenman, 1971, 1979) which we discuss in the next chapter. However, the questionnaire approach is limited because it cannot reveal the facial, vocal, and psychomotor indices of hostility and time urgency that are usually evident in interview (Friedman & Ulmer, 1984).

Early studies indicated that persons who exhibited the Type A behavior pattern were at greatly increased risk of coronary disease and heart attack. In one 9-year study of more than 3,000 healthy men persons with the Type A behavior pattern were 2!/2 times more likely to suffer heart attacks than those with Type B behavior pattern (Friedman & Ulmer, 1984). In fact, not one of the "pure" Type Bs—the extremely relaxed, easygoing, and noncompetitive members of the study—had suffered a heart attack. In the famous Framingham longitudinal study, Type A men ages 55 to 64 were about twice as likely at 10-year follow-up to develop coronary heart disease as Type B men (Haynes, Feinleib, & Eaker. 1983). In this study, the link between Type A behavior and heart disease was especially strong for white-collar workers.

In more recent studies, researchers have found only a weak relationship)—or no relationship at all between Type A behavior and coronary heart disease (e.g., Eaker & Castelli, 1988; Mathews & Haynes, 1986; Smedslund & Rundmo, 1999). Other researchers have found that heart disease is linked not so much with the full-blown Type A behavior pattern as it is with specific components such as being anger-prone (Dembroski, Mac-Dougall, Williams, & Haney, 1985) or possessing time urgency (Wright, 1988). Certainly, there is a need to sort out the specific risk factors in this area of investigation. In a review of current thinking, Wielgosz and Nolan (2000) identify hostility, cynicism, and suppression of anger, as well as stress, depression, and social isolation, as significant risk factors in Type A behavior. Good reviews of the complex and confusing research on Type A behavior can be found in Brannon and Feist (1992) and Wiebe and Smith (1997).

Research on Type A behavior has sparked a renewed and more sophisticated interest in typological conceptions of personality. Rather than viewing types as separate pigeonholes, psychologists have come to view them as idealized examples that occupy the end points of continuous dimensions. Individuals can thus differ with respect to how much of an idealized personality type that they possess. This is similar to the trait conception of personality discussed later. Perhaps the main difference is that modern type theorists tend to believe that most individuals are near to the idealized types at the end of each dimension, whereas trait theorists argue that people are more likely to be found at all points along each personality continuum. In practice, then, the modern distinction between types and traits is relative, not absolute.

PHENOMENOLOGICAL THEORIES OF PERSONALITY

Phenomenological theories of personality emphasize the importance of immediate, personal, subjective experience as a determinant of behavior. Some of the theoretical positions subsumed under this title have been given other labels also, such as humanistic theories, existential theories, construct theories, self-theories, and fulfillment theories (Maddi, 2000). Nonetheless, these approaches share a common focus on the person's subjective experience, personal worldview, and self-concept as the major wellsprings of behavior.

Origins of the Phenomenological Approach

The orientation briefly reviewed in this section has numerous sources that reach back to turn-of-the twentieth century European philosophy and literature. Nonetheless, two persons, one a philosopher and the other a writer, stand out as seminal contributors to the modern phenomenological viewpoint. The German philosopher Edmund Husserl (1859-1938) invented a complex philosophy of phenomenology that was concerned with the description of pure mental phenomena. Husserl's approach was heavily introspective and nearly inscrutable. More approachable was the Danish writer Soren Kierkegaard (1813-1855), well known for his contributions to existentialism. Existentialism is the literary and philosophical movement concerned with the meaning of life and an individual's freedom to choose personal goals. The phenomenology of Husserl and the existentialism of Kierkegaard influenced dozens of prominent philosophers and psychologists. Vestiges of these early viewpoints are evident in virtually every contemporary phenomenological personality theory (Maddi, 2000).

Carl Rogers, Self-Theory, and the Q-Technique

The most influential phenomenological theorist was Carl Rogers (1902-1987). His contributions to personality theory, known as self-theory, are extensive and generally well appreciated by students of psychology (Rogers, 1951, 1961, 1980). But it is also true, albeit little recognized that Rogers helped shape a small part of psychological testing by popularizing the Q-technique.

The **Q-technique** is a procedure for studying changes in the self-concept, a key element in Rogers's self-theory. The technique was developed by Stephenson (1953) but a series of studies by Rogers and his colleagues served to popularize this measurement approach (Rogers & Dymond, 1954). Also known as a Q-Sort, the Q-technique is a generalized procedure that is especially useful for studying changes in self-concept. The Q-sort consists of a large number of cards, each containing a printed statement such as the following:

I am poised
I put on a false front
1 make strong demands on myself
I am a submissive person
I am likeable

The examinee is asked to sort a hundred or so statements into nine piles, putting a prescribed number of cards into each, thus forcing a near-normal distribution. The instructions specify that the examinee put the cards most descriptive of him or her at one end, those least descriptive at the opposite end, and those about which he or she is indifferent or undecided around the middle of the distribution. The required distribution might look like this:

	Least	**Like Me**				**Most Like Me**			
Pile No.	1	2	3	4	5	6	7	8	9
No. of cards	1	4	11	21	26	21	11	4	1

The nature of the items is determined by the needs of the researcher or practitioner. Rogers used a set of items devised by Butler and Haigh (Rogers & Dymond, 1954, chap. 4) to tap the self-concept. These statements were taken at random from available therapeutic protocols; their Q-sort items represented actual client statements, reworded for clarity. But a special virtue of the Q-technique is that other researchers or practitioners are free to craft their own items. For example, Marks and Seeman (1963) used a psychodynamic perspective in devising items for the therapist description of patient groups. Examples of their items include the following:
Utilizes acting out as a defense mechanism
Tends to be flippant in both word and gesture
Genotype has paranoid features
Appears to be poised, self-assured, socially at ease
Exhibits depression (manifest sad mood)

Scoring a Q-sort is usually a matter of comparing or correlating the distribution of items against an established norm. For example, well-adjusted persons might be asked to sort the items so as to derive an average pile placement number (ranging from 1 to 9) for each item. An individual examinee would be considered more- or less-adjusted according to the resemblance between his or her sortings and the average sorting for adjusted persons. We will refer the reader to Block (1961) for details.

Another way to use the Q-sort is to compare an examinee's self-sort with his or her ideal sort. Rogers used the discrepancy between these two sortings as an index of adjustment. His subjects were required to sort the items twice, according to the following instructions:

Self sort. Sort these cards to describe yourself as you see yourself today, from those that are least like you to those that are most like you.

Ideal sort. Now sort these cards to describe your ideal person—the person you would most like within yourself to be (Rogers & Dymond, 1954).

Using the item pile numbers, Rogers then correlated the two sorts for each subject separately. Consider what these data mean: If the self-sort and the ideal sort are highly similar, the correlation of Q-sort data will approach 1.0; if the two sorts are opposite one another, the correlation will approach -1.0. Of course, most sorts will be somewhere in between but typically on the positive side. Butler and Haigh found that psychotherapy clients increased their congruence between self and ideal (Rogers & Dymond, 1954, chap. 4). Even so, adjusted control subjects possessed a greater congruence (Table 13.2).

Table 13.2 Average Self Ideal Correlations for Client and Control Groups

	Precounseling	Postcounseling	Follow-Up
Client Group (N=25)	-.01	.36	.32
Control Group (N= 16)	.58		.59

BEHAVIORAL AND SOCIAL LEARNING THEORIES

Behavioral and social learning theories have their origins in laboratory studies on operant learning and classical conditioning. A fundamental assumption of all behavioral theorists is that many of the behaviors that make up personality are learned. To understand personality, then, we must know about the learning history of the individual. Behavioral theorists also believe that the environment is of supreme importance in shaping and maintaining behavior. Behavioral inquiry therefore seeks to identify the specific components of the current environment that are controlling a person's behavior. The behavioral approach to personality has produced a variety of direct assessment methods, which we discuss in the next chapter.

Behavioral theorists disagree mainly on the role that cognitions play in determining behavior. Cognitions are inferred mental processes such as problem solving, judging, or reasoning. Radical behaviorists believe that resorting to mentalistic explanations of any kind is futile: "When what a person does is attributed to what is going on inside him, investigation is brought to an end" (Skinner, 1974). By contrast, social learning theorists make cautious reference to cognitions in explaining what it is, specifically, that a person learns. A social learning theorist might argue that we learn expectations or rules about the environment, not just stimulus and response connections.

Modern social learning theory can be viewed as a cognitive variant of the strict behaviorism that was dominant in U.S. psychology early in the twentieth century. Social learning theorists accept the Skinnerian premise that external reinforcement is an important determinant of behavior. But they also maintain that cognitions have a critical influence on our actions as well. For example, Rotter (1972) has popularized the view that our expectations about future outcomes are the primary determinants of behavior. The probability that a person will behave self-assertively, for example, depends upon his or her expectations about the likely results of self-assertiveness. If the expected outcome is valued by the person, the behavior is more likely. Of course, expectations are a function of the person's history of reinforcement, so Rotter's social learning perspective is similar to the behavioral viewpoint. But the implication of social learning theory is that behavior is the result of a belief, in particular, a belief that the behavior will result in a desired outcome. Thus, cognitions are assumed to affect actions.

Based on his social learning views, Rotter (1966) developed the Internal-External (I-E) Scale, an interesting measure of internal versus external locus of control. The construct of locus of control refers to the perceptions that individuals have about the source of things that happen to them. In particular, the I-E Scale seeks to assess the examinee's generalized expectancies for internal versus external control of reinforcement. The purpose of the I-E Scale is to determine the extent to which the examinee believes that reinforcement is contingent upon his/her behavior (internal locus of control) as opposed to the outside world (external locus of control). The instrument is a forced-choice self-report inventory. For each item, the examinee chooses the single statement (from a pair) with which he/she more strongly concurs. Items resemble the following:

In general, most people get the respect they deserve.

OR

In reality, a person's worth often passes unrecognized.

For the preceding item, the first alternative indicates an internal locus of control, whereas the second alternative signifies an external locus of control. The balance of internal to external responses determines the overall score on the scale. The I-E Scale is a reliable and valid instrument that has stimulated a huge body of research on the nature and meaning of locus of control and related variables. Research indicates that locus of control has a strong relationship to occupational success, physical health, academic achievement, and numerous other variables. As the reader might suspect, an internal locus of control generally predicts a more positive outcome than an external locus of control. The interested reader can consult Lefcourt (1991) and Wall, Hinrichsen, and Pollack (1989) for further details.

Important contributions to social learning theory have also been made by Albert Bandura. In his early studies, Bandura examined the role of observational learning and vicarious reinforcement in the development of behavior (Bandura, 1965, 1971; Bandura & Walters, 1963). More recently, he has proposed that perceived self-efficacy is a central mechanism in human action (Bandura, 1982; Bandura. Taylor, Ewart, Miller, & DeBusk, 1985). Self-efficacy is a personal judgment of "how well one can execute courses of action required to deal with prospective situations" (Bandura, 1982). The concept of self-efficacy is useful in explaining why correct knowledge does not necessarily predict efficient action. For example, two boys may be equally convinced that a garden snake in the bathtub presents no hazard, but one will pick it up while the other runs out the door. These differences in behavior illustrate the role of self-referential thought as a mediator between knowledge and action. The boy who ran out the door did not believe he could deal with the situation effectively. He had little perceived self-efficacy for snake handling. Bandura would argue that the primary determinant of the boy's behavior is a self-judgment about personal capabilities. Cognitions are therefore assumed to be a major determinant of behavior.

Bandura has developed an interesting instrument for the assessment of self-efficacy expectancies (Bandura, Taylor, Ewart, Miller, & BeBusk, 1985). For a variety of situations that might arouse anxiety, annoyance, or anger, the examinee checks whether he or she "can do" the task, and also rates the degree of confidence using a number from 10 to 100. The format of the checklist is as follows:

10	20	30	40	50	60	70	80	90	100
Quite			Moderately					Certain	
Uncertain			Certain						

Can Do Confidence

Go to a party at which there is no one you know. ____________
Complain about poor food at a restaurant. ____________

Bandura's instrument is essentially a criterion-referenced tool for use in psychotherapy and research.

TRAIT CONCEPTIONS OF PERSONALITY

A trait is any "relatively enduring way in which one individual differs from another" (Guilford, 1959). Psychologists developed the concept of trait from the ways people describe other people in everyday life. As language evolved, people found words to portray the consistencies and differences they encountered in their daily interactions with others. Thus, when we say one person is sociable and another is shy we are using trait names to describe consistencies within individuals and also differences between them (Goldberg, 1981a; Fiske, 1986).

Trait conceptions of personality have been enormously popular throughout the history of psychological testing, so the coverage here is necessarily selective. We will review three prominent and influential positions from the dozens of trait theories that have been proposed. These approaches differ primarily in terms of whether traits are split off into finely discriminable variants or grouped together into a small number of broad dimensions:

1. Cattell's factor-analytic viewpoint identifies 16 original 20 bipolar trait dimensions.

2. Eysenck's trait-dimensional approach coalesces dozens of traits into two overriding dimensions.

3. Goldberg and others have sought a modern synthesis of all trait approaches by proposing a five factor model of personality.

For readers who desire a more detailed discussion of this topic, Pervin (1993) and Wiggins (1997) provide an excellent review of trait approaches to personality theory.

Cattell's Factor-Analytic Trait Theory

Cattell (1950. 1973) refined existing methods of factor analysis to help reveal the basic traits of personality. He referred to the more obvious aspects of personality as surface traits. These would typically emerge in the first stages of factor analysis when individual test items were correlated with each other. For example, true-false items such as "I enjoy a good prize fight." '"Getting stuck behind a slow driver really bothers me," and "It's important to let people know who is in charge" might be answered similarly by subjects, revealing a surface trait of aggressiveness.

But surface traits themselves tended to come in clusters, as revealed by Cattell's more sophisticated application of factor analysis. For Cattell, this was evidence of the existence of source traits, the stable and constant sources of behavior. Source traits are therefore less visible than surface traits but are more important in accounting for behavior.

Cattell (1950) was unrivaled in his use of factor analysis to discover how traits were organized and how they were related to each other. One approach was to have persons rate others they knew well by checking various adjectives such as aggressive, thoughtful, and dominating from a list of 171 choices. When the results from 208 subjects were subsequently factor analyzed, about 20 underlying personality factors or traits were tentatively identified. Another approach was to have thousands of persons answer questions about themselves and then factor analyze their responses. Sixteen of the original 20 personality traits were independently confirmed by this second approach (Cattell, 1973).

These 16 source traits have been incorporated into the Sixteen Personality Factor Questionnaire (16PF), a trait-based paper-and-pencil test of personality that is discussed in the next chapter, factor model of

Eysenck's Trait-Dimensional Theory

Eysenck used factor analysis to produce a parsimonious rapprochement between trait and dimensional approaches to personality (Eysenck & Eysenck, 1975, 1985). According to his system, personality consists of two basic dimensions, introverted-extraverted and emotionally stable emotionally unstable. These two dimensions are presumed to be biologically and genetically based. Furthermore, the dimensions subsume numerous specific traits. The positions of the 32 traits correspond to the direction and amount of the two basic dimensions. For example, a moderately extraverted person who was also moderately unstable might be characterized by these traits: aggressive, excitable, and changeable. An extremely introverted person who was also midway on the stable-unstable dimension might be viewed as unsociable, quiet, passive, and careful. Eysenck's trait-dimensional theory is incorporated in his personality inventory, the Eysenck Personality Questionnaire, which we review in the next chapter.

The Five-Factor Model of Personality

The five-factor model of personality has its origins in a review chapter by Goldberg (1981b). In his analysis of factor-analytic trait research, Goldberg identified several consistencies, which he referred to as the "Big Five" dimensions. Although researchers have used slightly different terms for these factors, the most common labels are

Neuroticism
Extraversion
Openness to Experience

Agreeableness
Conscientiousness

Rearranging the factors yields a simple acronym: OCEAN. The five-factor model is rapidly becoming the consensus model of personality. Support for the five-factor approach comes from several sources, including factor analysis of trait terms in language and the analysis of personality from an evolutionary perspective. Following, we discuss these perspectives.

The use of trait terms in the analysis of personality is based upon the fundamental lexical hypothesis. The essential point of this hypothesis is that trait terms have survived in language because they convey important information about our dealings with others:

> *The variety of individual differences is nearly boundless, yet most of these differences are insignificant in people's daily interactions with others and have remained largely unnoticed. Sir Francis Galton may have been among the first scientists to recognize explicitly the fundamental lexical hypothesis—namely that the most important individual differences in human transactions will come to be encoded as single terms in some or all of the world's languages. (Goldberg, 1990)*

When trait terms in English are distilled down to a reasonably distinct and nonoverlapping set of adjectives, a few hundred characteristics typically emerge (Allport, 1937). For decades, researchers have been asking individuals to rate themselves or others on these or similar traits. When these ratings are subjected to factor analysis, the "Big Five" dimensions previously listed usually appear in one guise or another. In sum, a mounting body of research indicates that the five-factor model captures a valid and useful representation of the structure of human traits.

The five-factor approach also possesses evolutionary plausibility. Specifically, the five factors of personality previously listed capture individual differences that relate to such basic evolutionary -13 functions as survival and reproductive success (Buss, 1997; Pervin, 1993). Goldberg (1981b) has theorized that people implicitly ask the following questions in their interactions with others:

1. Is X active and dominant or passive and submissive? (Can I bully X or will X try to bully me?)
2. Is X agreeable (warm and pleasant) or disagreeable (cold and distant)?
3. Can I count on X? (Is X responsible and conscientious or undependable and negligent?)
4. Is X crazy (unpredictable) or sane (stable)?
5. Is X smart or dumb? (How easy will it be for me to teach X?)

Directly or indirectly, each of these evaluations has a bearing upon survival and reproductive success. For example, point 3 (conscientiousness) involves a trait that might ensure group survival in a hostile world. A person low on this trait (undependable) would be a poor choice for guarding the food supply. The ability to discern conscientiousness in others therefore has adaptive value. Not surprisingly, the five points previously listed correspond to the five-factor personality model.

The five-factor model of personality has inspired several personality scales and other systems for assessment (deRaad & Perugini, 2002). For example, Costa and McCrae have developed two personality tests based upon the five-factor model (Costa, 1991; McCrae & Costa. 1987). The Revised NEO Personality Inventory (NEO-PI-R) contains 240 items rated on a five-point scale. In addition to the five major domains of personality, the inventory measures six specific traits (called facets) within each domain. A shortened 60-item version known as the NEO Five-Factor Inventory (NEO-FFI) also is available. Trull, Widiger, Useda, and others (1998) have published a semi structured interview for the assessment of the five-factor model of personality. These tests are discussed in the next chapter.

Comment on the Trait Concept

The challenge faced by trait theorists is that psychologists have long known that thousands of trait names can be found in any standard English dictionary. For example, in an early and influential study, Allport and Odbert (1936) tallied over 18,000 trait names. This is obviously too many to be useful in any theory of personality or testing, so theorists are required to search for a smaller, more manageable number of basic traits. Until recently, there was no consensus whatever on the number of fundamental traits. Some theorists proposed two or three overriding trait factors, whereas others divided the personality domain into sixteen or twenty trait dimensions. Many personality theorists—perhaps a majority—now concede that the five factors previously noted (Neuroticism, Extraversion, Openness, Agreeableness, Conscientiousness) provide a parsimonious and useful way to look at personality. But this model is very recent, and it will take time to confirm its utility. For example, there is still debate about whether Openness to Experience belongs on the list of fundamental dimensions of personality (Digman, 1990). Also, why is Intellect not included in the five-factor model?

All trait approaches to personality share certain problems in common. First, there is disagreement whether traits cause behavior or merely describe behavior (Fiske, 1986). It can be persuasively argued that invoking traits as causes is an empty form of circular reasoning. For example, a person with extremely high standards might be said to possess the trait of perfectionism. But when asked to explain what is meant by perfectionism, we invariably end up referring to a pattern of extremely high standards. Thus, when we assert that someone is perfectionist, are we really doing anything more than providing a short-hand description of their past behavior? Miller (1991) has voiced this criticism of the five-factor approach, noting that the model merely describes psychopathology but does not explain it.

A second problem with traits is their apparently low predictive validity. Mischel (1968) is credited with the first effective disparagement of the trait concept in his influential book Personality and Assessment. He stated that "while trait theory predicts behavioral consistency, it is behavior inconsistency that is typically observed" (Mischel, 1968).

Projective Techniques

Frank (1939, 1948) introduced the term projective method to describe a category of tests for studying personality with unstructured stimuli. In a projective test the examinee encounters vague, ambiguous stimuli and responds with his or her own constructions. Disciples of projective testing are heavily vested in psychoanalytic theory and its postulation of unconscious aspects of personality. These examiners believe that unstructured, vague, ambiguous stimuli provide the ideal circumstance for revelations about inner aspects of personality. The central assumption of projective testing is that responses to the test represent projections from the innermost unconscious mental processes of the examinee. We introduce this topic with some preliminary concepts and distinctions relevant to projective testing.

THE PROJECTIVE HYPOTHESIS

The assumption that personal interpretations of ambiguous stimuli must necessarily reflect the unconscious needs, motives, and conflicts of the examinee is known as the projective hypothesis. Frank (1939) is generally credited with popularizing the projective hypothesis:

> *When we scrutinize the actual procedures that may be called projective methods we find a wide variety of techniques and materials being employed for the same general purpose, to obtain from the subject, "what he cannot or will not say," frequently because he does not know himself and is not aware what he is revealing about himself through his projections.*

The challenge of projective testing is to decipher underlying personality processes (needs, motives, and conflicts) based on the individualized, unique, subjective responses of each examinee. In the sections that follow we will examine how well projective tests have met this portentous assignment.

A PRIMER OF PROJECTIVE TECHNIQUES

Origins of Projective Techniques

Projective techniques date back to the nineteenth century. By way of quick review, Galton (1879) developed the first projective technique, a word association test. This procedure was adapted to testing by Kent and

Rosanoff (1910) and used in therapy by C. G. Jung and others. Meanwhile, Ebbinghaus (1897) used a sentence completion test as a measure of intelligence, but others soon realized the method was better suited to personality assessment (Payne. 1928; Tendler, 1930). Heavily influenced by psychoanalytic formulations of personality, Rorschach published his famous inkblot test in 1921. In 1905, Binet invented a precursor to story telling or thematic apperception techniques when he used verbal responses to pictures as a measure of intelligence. These and other endeavors form the cornerstone of modern projective testing.

The Popularity of Projective Tests: A Paradox

The widespread use of projective tests has continued unabated from the early twentieth century to present times (Louttit & Browne, 1947; Lubin, Wallis, & Paine, 1971; Watkins, Campbell, & McGregor, 1988). Recently, Watkins, Campbell, Nieberding, and Hallmark (1995) surveyed more than 400 psychologists who practiced assessment to estimate the frequency of use of various prominent tests. They discovered that 5 of 15 most frequently used tests are projective techniques (Table 13.3).

Paradoxically, from the standpoint of traditional psychometric criteria, projective tests do not fare nearly as well as the objective tests discussed in the next chapter. The essential puzzle of projective tests is how to explain the enduring popularity of these instruments in spite of their sometimes questionable psychometric quality. After all, psychologists are not uniformly dense, nor are they dumb to issues of test quality. So why do projective techniques persist? We return to this puzzle— which might be called the projective paradox— after we familiarize the reader with prominent approaches to projective testing.

A Classification of Projective Techniques

Lindzey (1959) has offered a classification of projective techniques that we will follow here. Based on the response required, he divided projectives into five categories:

- Association to inkblots or words
- Construction of stories or sequences
- Completion of sentences and stories
- Arrangement/selection of pictures or verbal choices
- Expression with drawings or play

Table 13.3 The 15 Most Frequently Used Tests in the United States

Test	Rank
Wechsler Adult Intelligence Scale-Revised	1
Minnesota Multiphasic Personality Inventory-2	2
Sentence Completion Methods	3
Thematic Apperception Test	4
Rorschach	5
Bender-Gestalt	6
Projective Drawings	7
Beck Depression Inventory	8
Wechsler Intelligence Scale for Children-III	9
Wide Range Achievement Test-Revised	10
Wechsler Memory Scale-Revised	11
Peabody Picture Vocabulary Test-Revised	12
Millon Clinical Multiaxial Inventory-II	13
Wechsler Preschool and Primary Scale of Intelligence-R	14
Children's Apperception Test	15

Association techniques include the widely used Rorschach inkblot test and its psychometrically superior cousin the Holtzman Inkblot Test, as well as word association tests. Construction techniques include the Thematic Apperception Test and the many variations upon this early instrument. Completion techniques consist mainly of sentence completion tests, discussed later. Arrangement/ selection procedures such as the Szondi test (discussed in the first chapter) are currently seldom used. Finally, expression techniques such as the Draw-A-Person or House-Tree-Person test are very popular among clinicians in spite of dubious validity data.

We will review prominent techniques within each category except the antiquated arrangement/ selection approaches, which are almost never used. However, the literature on major projective techniques is simply overwhelming, running to perhaps tens of thousands of articles on the Rorschach alone. We can suggest major trends in the research, but the reader will need to consult other sources for comprehensive reviews.

ASSOCIATION TECHNIQUES

The Rorschach

The Rorschach consists of 10 inkblots devised by Herman Rorschach (1884-1922) in the early 1900s. He formed the inkblots by dribbling ink on a sheet of paper and folding the paper in half, producing relatively symmetrical bilateral designs. Five of the inkblots are black or shades of gray, while five contain color; each is displayed on a white background. An inkblot of the type employed by Rorschach is shown in Figure 13.1. The Rorschach is suited to persons age five and up, but is most commonly used with adults.

In administering the Rorschach, the examiner sits by the examinee's side to minimize body language communication. Administration consists of two phases. In the free association phase, the examiner presents the first blot and asks, "What might this be?" If the examinee asks for clarification (e.g., "Should I use the whole blot or only part of it?"), the examiner always responds in a nondirective manner ("It's up to you"). The test proceeds at a leisurely pace, so there is an implicit expectation that the examinee will give more than one response per card. However, this is not required; it is even permissible for the examinee to reject a card entirely, although this rarely happens. All 10 cards are presented in a similar manner.

Next, the examiner begins the inquiry phase. In this phase the examiner asks questions to clarify the exact blot location of each percept and to determine which aspects of the blot, such as the form or color, played a part in the creation of the response. Based on the information collected during the inquiry phase, the examiner can then code the location, determinants, form quality, and content of

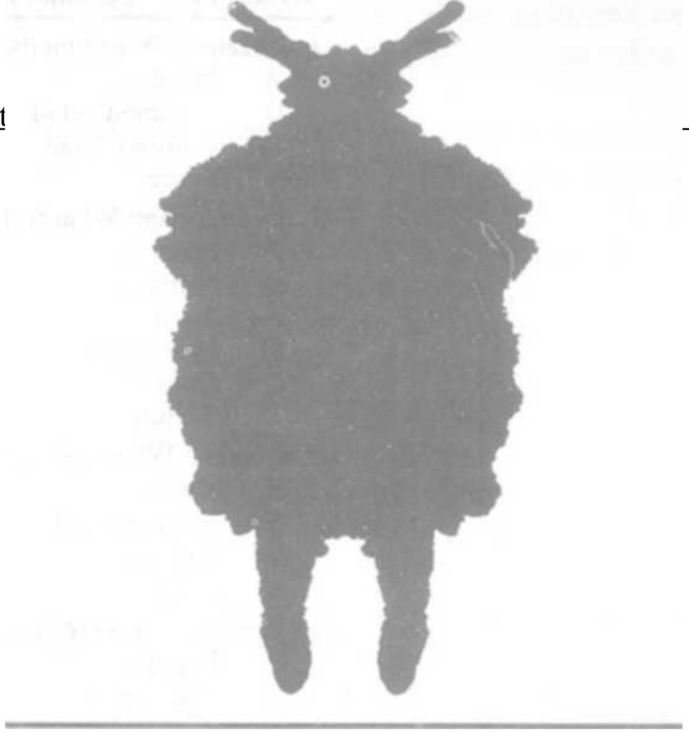


Figure 13.1 An Inkblot Similar to Those Found on the Rorschach

each response according to one or more formal scoring systems. For example, if the examinee used the entire blot for a percept, the response is coded W (whole); if the form of the blot was important in the percept, the response is further coded F (form); if human movement is depicted in the percept, the response is coded M (movement); the use of color in a percept is coded C (color), CF (color/form), or FC (form/color), depending upon whether form is totally absent, primary, or secondary to color as a determinant. The content of the percept is also coded, for example, H (human), Hd (human detail), An (anatomy), Cg (clothing), and so on (Table 13.4). Proper scoring of the Rorschach requires extensive training and supervision; we have touched on just a few basic aspects here.

Regrettably, Rorschach died before he could complete his scoring methods, so the systematization of Rorschach scoring was left to his followers.

Table 13.4 Summary of Major Rorschach Scoring Criteria

I.	Location: Where on the blot was the percept located?	
W	Whole	Entire inkblot used
D	Common detail	Well defined part used
Dd	Unusual detail	Unusual part used
S	Space	Percept defined by white space
II.	Determinant: What feature of the blot determined the response?	
F	Form	Shape or outlined used
F+	Form+	Excellent match of percept and inkblot
F-	Form-	Very poor match of percept and inkblot
M	Movement	Movement seen or applied in percept
C	Color	Color helped determined the response
T	Texture	Shading involved in the response
III.	Content: What was the percept?	
H	Human	Percept of a whole human form
Hd	Human	detail Human form incomplete in any way

Ex	Explosion	An actual explosion
Xy	X-Ray	X-Ray of any human part; involves shading
IV.	Popular versus original	
P	Popular	Response given by many normal persons
O	Original	Rare and creative response

Five American psychologists produced overlapping but independent approaches to the test—Samuel Beck, Marguerite Hertz, Bruno Klopfer, Zygmunt Piotrowski, and David Rapaport (Erdberg, 1985). Predictably, the nuances of scoring vary from one scoring method to another. Fortunately, Exner and his colleagues have synthesized these earlier approaches into the Comprehensive Scoring System (Exner, 1991, 1993; Exner & Weiner, 1994). The Comprehensive Scoring System is better grounded in empirical research and clearly has supplanted all other approaches to Rorschach scoring.

Once the entire protocol has been coded, the examiner can compute a number of summary scores that form the primary basis for hypothesizing about the personality of the examinee. For example, the F+ percent is the proportion of the total responses that uses pure form as a determinant. A voluminous literature exists on the meaning of this index, but it seems safe to hypothesize that when the F+ percentage falls below 70 percent, the examiner should consider the possibility of severe psychopathology, brain impairment, or intellectual deficit in the examinee (Exner, 1993). The F+ percent is also considered to be an index of ego strength, with higher scores indicating a greater capacity to deal effectively with stress. However, support for this conjecture is mixed at best.

Frank (1990) has emphasized that formal scoring of the Rorschach is insufficient for some purposes such as the diagnosis of schizophrenia. He stresses that an analysis of the patient's thinking for the presence of highly personal, illogical, and bizarre associations to the blots is essential for psychodiagnosis. In his approach, the Rorschach is really an adjunct to the interview, and not a test per se.

Comment on the Rorschach

For a variety of reasons, it is difficult to offer concise generalizations about the reliability, validity, and clinical utility of the Rorschach. Even simple questions provoke complex answers. For example, What is the purpose of a Rorschach evaluation? In successive research epochs, the Rorschach has been used to derive a psychiatric diagnosis, estimate prognosis for psychotherapy, obtain an index of primary process thinking, predict suicide, and formulate complex personality structures, to name just a few applications (Peterson, 1978). The purpose of the Rorschach is so ill-defined that some adherents even decline to regard it as a test, preferring instead to call it a method for generating information about personality functioning (Weiner, 1994). When the purpose of an instrument is unclear, objective research on its psychometric attributes is both risky and difficult. Worse yet, objective research may be pointless since supporters will ignore contrary findings and detractors don't use the test anyway.

A study by Albert, Fox, and Kahn (1980) on the susceptibility of the Rorschach to faking is typical of research on this instrument. We remind the reader that thousands of research studies exist in the literature, including many with positive, supportive findings (e.g., Hilsenroth, Fowler, Padawer, & Handler, 1997; Smith, Gacono, & Kaufman, 1997; Weiner, 1996). But the mixed results reported by Albert, Fox, and Kahn (1980) are not unusual. They submitted the Rorschach protocols of 24 persons to a panel of experts, asking for psychiatric diagnoses of each examinee. The 24 Rorschach protocols consisted of results from four groups of six persons each:

- Mental hospital patients with a diagnosis of paranoid schizophrenia
- Uninformed fakers given instructions to fake the responses of a paranoid schizophrenic
- Informed fakers who listened to a detailed audiotape about paranoid schizophrenia
- Normal controls who took the test under standard instructions

The uninformed fakers, informed fakers, and normal controls were students who had passed an MMPI screening and were judged reasonably normal during interview. Each protocol was rated by six to nine judges, all fellows of the Society for Personality Assessment. The-judges were told to provide a psychiatric diagnosis as well as other information not reported here. The judges were not informed as to the purpose of the study, but were told to assess whether any profiles appeared to be malingered.

The informed fakers must have done an excellent job, for they were more likely to be diagnosed psychotic than the real patients themselves (72 percent versus 48 percent, respectively). The uninformed fakers were fairly convincing, too, with a 46 percent rate of diagnosed psychosis. The normal controls were diagnosed as psychotic 24 percent of the time. Granted that the diagnostic challenge in this study was immense, it is still disturbing to find that the expert judges rated 24 percent of the normal protocols as psychotic, while correctly identifying psychosis in only 48 percent of the actual psychotic protocols. A more recent study by Netter and Viglione (1994) also concluded that the Rorschach was susceptible to the faking of psychosis.

Although there are noteworthy exceptions in Rorschach testing, a substantial number of studies point to low reliability and a general lack of predictive validity (Carlson. Kula, & St. Laurent. 1997; Peterson. 1978; Lanyon, 1984; Wood, Nezworski. & Stejskal. 1996: Lilienfeld. Wood, & Garb, 2000). In a meta-analytic review. Garb, Florio, and Grove (1998) concluded" that the Rorschach explained a dismal 8 to 13 percent of the variance in client characteristics, as compared to the MMPI, which explained 23 to 30 percent of the variance. On the positive side, recent studies based upon improvements in scoring offered by the Exner approach are more optimistic in outcome (see Exner, 1995; Exner & Andronikof-Sanglade, 1992; Meyer, 1997; Ornberg & Zalewski, 1994; Piotrow-ski, 1996). Even so, the Rorschach has not yet gained the status of scientific respectability enjoyed by many other personality tests, and perhaps it never will.

Holtzman Inkblot Technique

Wayne H. Holtzman sought to overcome the major limitations in the Rorschach by developing a completely new technique using more inkblots with simplified procedures for administration and scoring. In the Holtzman Inkblot technique, the examinee is limited to one response per card, but views a series of 45 cards. Each response is followed with a very simple twofold question: Where was the percept represented in the blot, and what about the blot suggested the percept?

The HIT comes in two carefully constructed parallel forms. The existence of parallel forms is invaluable for test-retest studies, since examinees often remember their responses to a card and therefore mechanically offer the same answer when retested. The 45 responses to the HIT are scored for 22 different variables derived from early Rorschach scoring systems. The HIT scoring variables are described in Table 13.5.

Table 13.5 Names and Descriptions of the Holtzman Inkblot Technique Variables

Reaction Time	Time in seconds from the presentation of the inkblot to the beginning of the primary response.
Rejection	Subject fails to report anything or returns the inkblot to the examiner.
Location	Scored on a 3-point system: 0—whole blot, 1—large area, 2—smaller area.
Space	Scored when there is a true figure-ground reversal; the white part is the figure.
Form Definiteness	Scored on a 5-point system from 0 (formless concept—e.g., paint splatter) to 4 (highly formed concept—e.g., centaur).
Form Appropriateness	Goodness of fit of the concept to the form of the inkblot; 0—poor, 1—fair, 2—good.
Color	Color is a primary determinant usually mentioned by the subject; scored 0 to 3.

Shading	Subject refers to shading (fuzziness, texture) as a determinant: scored 0 to 2.
Movement	Scored when the response implies energy or dynamic movement quality; scored 0 to 4.
Pathognomic Verbalization	Incoherent, queer, absurd, self-referential, etc., verbalizations to cards.
Integration	Scored 1 if two or more blot elements are effectively integrated in the response; otherwise scored 0.
Content Scores	Each category (Human, Animal, Anatomy, Sex, Abstract) is scored 0 to 2 based on absence, partial, or full presence of the concept.
Anxiety	Each response is scored 0 to 2 for signs of anxiety (e.g., dark and dangerous cave).
Hostility	Each response is scored 0 to 3 for signs of hostility (e.g., mangled butterfly).
Barrier	Barrier refers to any protective covering, membrane, shell, or skin that might be symbolically related to body-image boundaries; 1 if present, 0 if absent.
Penetration	Scored 1 if the concept is symbolic of an examinee's feeling that his or her body exterior can be easily penetrated; otherwise 0
Balance	Scored 1 if examinee refers to presence or absence of symmetry in the design; otherwise scored 0.
Popular	Scored 1 if the response is common, observed in 1of 7 normative protocols.

The scoring system for the HIT is highly reliable, and the standardization of the instrument appears to be adequate. When well-trained scorers are used, interscorer agreement for the different categories is .95 to 1.00 for most categories: only Penetration and Integration fall below these standards. Split-half reliabilities are also acceptable, with median values in the .70s and .80s. Test-retest stability with parallel forms is generally fair, although some categories (Location, with r of .81) perform better than others (Popular, with r of .36). Percentile norms for each scoring category are reported separately for college students (N = 206), average adults (N = 252), seventh graders (N = 197), elementary schoolchildren (N = 132), five-year-olds (N = 122), chronic schizophrenics (N = 140), depressed patients (N = 90), and persons with mental retardation (N = 100).

The validity of the HIT has been addressed in several hundred research studies reporting on the relationships between HIT scores and independent measures of personality (Hill, 1972; Holtzman, 1988; Swartz, Reinehr, & Holtzman, 1983). In general, the relationships are modest but supportive of HIT validity, especially as an aid to psychodiagnosis. Holtzman (2000, 2002) describes the cross-cultural applications of the HIT and notes that the test has been featured in more than 800 publications.

A recent variant of the HIT requires two responses to each of a carefully selected subset of 25 cards from Form A. Called the HIT 25 to distinguish it from the standard HIT, this new test holds exceptional promise for helping make the diagnosis of schizophrenia. Using completely objective scoring criteria and simple decision rules, the HIT 25 correctly classified 26 of 30 schizophrenics and 28 of 30 normal college students (Holtzman, 1988). The decision criteria consist of four rules for normal findings scored +1 each, and 13 rules for schizophrenic findings scored -1 each. The total results are summed algebraically, yielding the "normalcy" score. This score is the basis for simple diagnostic decisions. Scores above zero suggest normalcy, whereas scores below zero indicate schizophrenia; a score of zero is indeterminate. The HIT 25 looks promising but cross-validation studies would be especially welcome.

COMPLETION TECHNIQUES

Sentence Completion Tests

In a sentence completion test, the respondent is presented with a series of stems consisting of the first few words of a sentence, and the task is to provide an ending. As with any projective technique, the examiner assumes that the completed sentences reflect the underlying motivations, attitudes, conflicts, and fears of the respondent. Usually, sentence completion tests can be interpreted in two different ways: subjective-intuitive analysis of the underlying motivations projected in the subject's responses, or objective analysis by means of scores assigned to each completed sentence.

An example of a sentence completion test is shown in Figure 13.2. This test is quite similar to existing instruments in that the stems are very short and restricted to a small number of basic themes. The reader will notice that three topics reoccur in this short test (the respondent's self-concept, mother, and father). In this manner the examinee has multiple opportunities to reveal underlying motivations about each topic. Of course, most sentence completion tests are much longer—anywhere from 40 to 100 stems—and contain more themes—anywhere from 4 to 15 topics.

Directions: Finish these sentences to indicate how you feel.

1. My best characteristic is
2. My mother
3. My father
4. My greatest fear is
5. The best thing about my mother was
6. The best thing about my father was
7. I am proudest about
8. I only wish my mother had
9. I only wish my father had

Figure 13.2 Example of a Short Sentence Completion Test

Dozens of sentence completion tests have been developed: most are unpublished and unstandardized instruments produced to meet a specific clinical need. Some representative sentence completion tests in current use are outlined in Table 13.6. Of these instruments, Loevinger's Washington University Sentence Completion Test is the most sophisticated and theory-bound (e.g., Weiss, Zilberg, & Genevro, 1989). However, the Rotter Incomplete Sentences Blank has the strongest empirical underpinnings and is the most widely used in clinical settings. We examine this instrument in more detail.

Table 13.6 Brief Outline of Representative Sentence Completion Tests

Sentence Completion Series
Psychological Assessment Resources

The SCS consists of 50 sentence stems designed to aid the clinician in identifying underlying concerns and specific areas of client distress. A unique feature of this instrument is the publication of eight different forms, parallel in content, which allow for repeated testing.

Forer Structured Sentence Completion Test
Western Psychological Services

This instrument is available in separate forms for men, women, adolescent boys, and adolescent girls. Each form contains 100 sentence stems designed to cover attitude-value systems, evasiveness, and defense mechanisms.

Geriatric Sentence Completion Form
Psychological Assessment Resources

The GSCF is a 30-item form specifically developed for use with older adult clients. The GSCF elicits personal responses to four content domains: physical, psychological, social, and temporal orientation. The test manual includes a number of clinical case illustrations.

Washington University Sentence Completion Test,
Privately published by Loevinger

The WUSC uses separate forms for men, women, and younger male and female subjects. This test is highly theory-bound; responses are classified according to seven stages of ego development: presocial and symbiotic, impulsive, self-protective, conformist, conscientious, autonomous, integrated.

Rotter Incomplete Sentences Blank

The Rotter Incomplete Sentences Blank (RISB) consists of three similar forms—high school, college, and adult—each containing 40 sentence stems written mostly in the first person (Rotter & Raf-ferty, 1950; Rotter, Lah, & Rafferty, 1992). Although the test can be subjectively interpreted in the usual manner through qualitative analysis of needs projected in the subject's responses, it is the objective and quantitative scoring of the RISB that has drawn the most attention.

In the objective scoring system each completed sentence receives an adjustment score from 0 (good adjustment) to 6 (very poor adjustment). These scores are based initially on the categorizing of each response as follows:

Omission—no response or response too short to be meaningful
Conflict response—indicative of hostility or un-happiness
Positive response—indicative of positive or hopeful attitude
Neutral response—declarative statement with neither positive nor negative affect

Examples of the last three categories include:

I hate ... the entire world, (conflict response)
The best... is yet to come, (positive response)
Most girls ... are women, (neutral response)

Conflict responses are scored 4, 5, or 6, from lowest to highest degree of the conflict expressed. Positive responses are scored 2, 1, or 0, from least to most positive response. Neutral responses and omissions receive no score. The manual gives examples of each scoring category. The overall adjustment score is obtained by adding the weighted ratings in the conflict and positive categories. The adjustment score can vary from 0 to 240, with higher scores indicating greater maladjustment.

The reliability of the adjustment score is exceptionally good, even when derived by assistants with minimal psychological expertise. Typically, interscorer reliabilities are in the .90s and split-half coefficients are in the .80s (Rotter et al., 1992; Rotter, Rafferty, & Schachtitz, 1965). The validity of this index has been investigated in numerous studies using the RISB as a screening device with a "maladjustment" cutoff score. For example, a cut¬off score of 135 has been found to correctly screen delinquent youths 60 percent of the time while identifying nondelinquent youths correctly 73 percent of the time (Fuller, Parmelee, & Carroll, 1982). The same cutoff identifies heavy drug users 80 to 100 percent of the time (Gardner, 1967). These and similar findings support the construct validity of the adjustment index, but also indicate that classification rates are much lower than needed for individual decision making or effective screening. It also appears that the norms for the adjustment index are outdated. Lah and Rotter (1981) found that current student scores differ significantly from those obtained in the original study by Rotter and Rafferty (1950). Lah (1989) and Rotter et al. (1992) provide new normative, scoring, and validity data for the RISB.

As discussed by R Goldberg (1965), the simplicity of the single adjustment score is both the test's strength and weakness. True, the test provides a quick and efficient method for obtaining an overall index of how respondents are functioning on a day to day basis. However, a single score cannot possibly capture any nuances of personality functioning. In addition, the RISB is subject to the same types of bias as other self-report measures, namely, the information will reflect mainly what the respondent wants the examiner to know (Phares, 1985).

Rosenzweig Picture Frustration Study

Often considered a semiprojective technique, the Rosenzweig Picture Frustration Study (P-F Study) requires the examinee to produce a verbal response to highly structured verbal-pictorial stimuli. The P-F Study comes in three forms—child, adolescent, and adult—each consisting of 24 comic-strip pictures depicting a frustrating circumstance (Rosenzweig, 1977, 1978a). Each picture contains two people, with the person on the left uttering words that provoke or describe a frustrating situation to the person on the right (Figure 13.3). The examinee is requested to indicate, by writing in the balloon above the frustrated person's head, the first verbal response that comes to mind as being uttered by the anonymous cartoon figure. In the case of younger examinees, the examiner writes down the subject's response.

The purpose of the P-F Study is to assess the examinee's characteristic manner of reacting to frustration. Frustration is defined as occurring whenever the organism encounters an obstacle or obstruction en route to the satisfaction of a need (Rosenzweig, 1944). In a general sense, it is well known that persons react to frustration with aggression. The value of the P-F Study is its multi-faceted conceptualization of aggression according to three directions and three types. The direction of aggression can be extraggressive, it is turned onto the environment; intraggressive, it is turned by the examinee onto the self; or unaggressive, it is evaded in an attempt to gloss over the frustration.

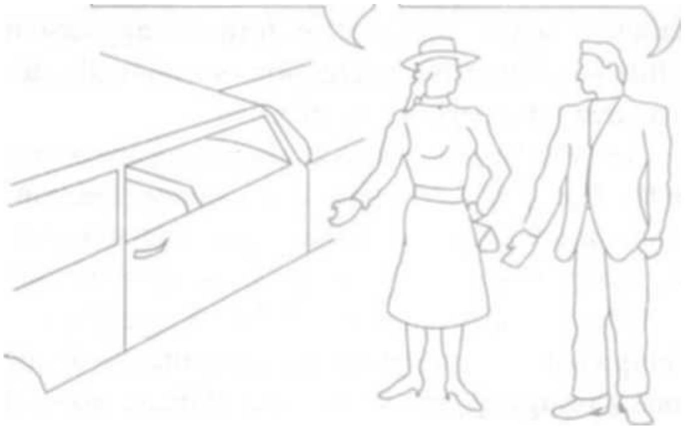


Figure 13.3 Sample Item from the Rosenzweig Picture-Frustration Study

The type of aggression can be obstacle-dominant, in which the barrier that occasions the frustration stands out in the response; ego-defensive, in which the organizing capacity of the examinee predominates in the response; or need-persistent, in which the solution of the frustrating situation is emphasized by pursuing the goal despite the obstacle (Rosen-zweig, 1978b). It is important to point out that aggression is not necessarily a negative construct. Need-persistent types of aggression represent constructive, sometimes creative, forms of aggression while ego-defensive aggression is frequently destructive (of others or oneself).

The P-F Study is scored by detecting one or two of the factors in each individual response. Deep interpretations are avoided; the manual contains scoring samples to aid in decision making. When the item scores have been tallied, the scoring blank is completed by computing the percentages of the nine scoring categories which occur in the protocol of the examinee. The overall types and directions of aggression are also tallied, resulting in 15 indices. In addition, a Group Conformity Rating (GCR) can be computed. The GCR indicates how closely the examinee's responses correspond to those given most frequently by a norm sample. All the indices can be compared to results from appropriate standardization samples. Of course, in addition to quantitative scoring, responses to the P-F Study can be evaluated impressiomstically.

The interscorer reliability of the P-F Study is reportedly in the range of .80 to .85 for well-trained, conscientious examiners. However, the test-retest stability of the instrument is somewhere between fair and marginal. For example, retest correlations for scoring categories on the adult form of the P-F Study range from .21 to .71, with most values in the .40s (Rosenzweig, 1978b). A huge body of validational research has been summarized in several publications (Rosenzweig, 1977, 1978b; Rosenzweig & Adelman, i977). Based on the very modest reliabilities of the scoring categories, we concur that the P-F Study is more appropriate for research than individual assessment (Graybill & Heuvelman, 1993).

CONSTRUCTION TECHNIQUES

The Thematic Apperception Test (TAT)

The TAT consists of 30 pictures that portray a variety of subject matters and themes in black-and white drawings and photographs; one card is blank. Most of the cards depict one or more persons engaged in ambiguous activities. Some cards are used for adult males (M), adult females (F), boys (B), or girls (G), or some combination (e.g., BM). As a consequence, exactly 20 cards are appropriate for every examinee.

A picture similar to those on the TAT is shown in Figure 13.5. In administering the TAT, the examiner requests the examinee to make up a dramatic story for each picture, telling what led up to the current scene, what is happening at the moment, how the characters are thinking and feeling, and what the outcome will be. The examiner writes down the story verbatim for later scoring and analysis.

The TAT was developed by Henry Murray and his colleagues at the Harvard Psychological Clinic (Morgan & Murray, 1935; Murray, 1938). The test was originally designed to assess constructs such as needs and press, elements central to Murray's personality theory. According to Murray, needs organize perception, thought, and action and energize behavior in the direction of their satisfaction. Examples of needs include the needs for achievement, affiliation, and dominance. In contrast, press refers to the power of environmental events to influence a person. Alpha press is objective or "real" external forces, whereas beta press concerns the subjective or perceived components of external forces. Murray (1938. 1943) developed an elaborate TAT scoring system for measuring 36 different needs and various aspects of press, as revealed by the examinee's stories.

Almost as soon as Murray released the TAT, other clinicians began to develop alternative scoring systems (e.g., Dana, 1959; Eron, 1950; Shneidman, 1951; Tomkins. 1947). Literature on the

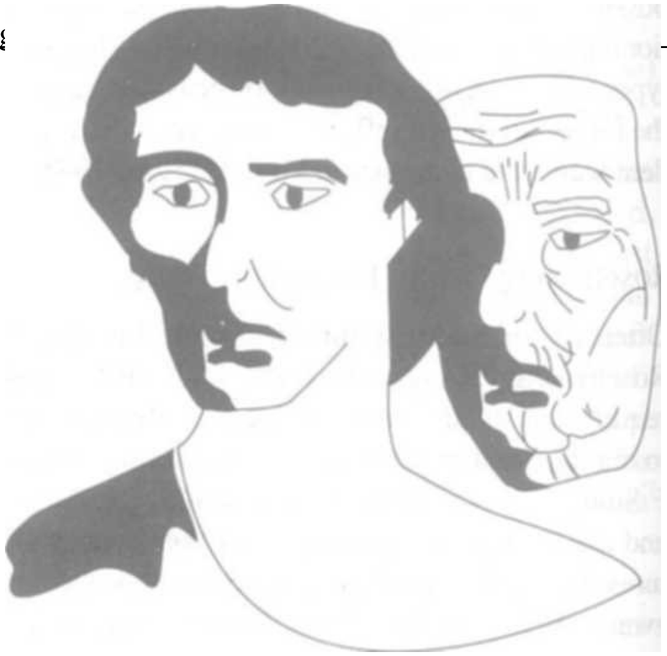


Figure 13.4 A Picture Similar to Those on the Thematic Apperception Test

administration, scoring, and interpretation of the TAT burgeoned extensively, as documented by recent reviews (Aiken, 1989, chap. 12; Groth-Marnat, 1997; Ryan. 1987; Weiner & Kuehnle, 1998). By the 1950s, there was no single preferred mode of administration, no single preferred system of scoring, and no single preferred method of interpretation, a predicament that still endures today. Clinicians even vary the wording of the instructions and commonly select an individualized subset of TAT cards for each client. Indeed, the absence of standardized procedures is such that we should rightly regard the TAT as a method, not a test.

It is worth mentioning that Murray's instructions included a statement that the TAT was "a test of imagination, one form of intelligence" and further stipulated:

I am going to show you some pictures, one at a time; and your task will be to make up as dramatic a story as you can for each. Tell what has led up to the event shown in the picture, describe what is happening at the moment, what the characters are feeling and thinking; and then give the outcome. Speak your thoughts as they come to your mind. Do you understand? Since you have fifty minutes for ten pictures, you can devote about five minutes to each story. Here is the first picture. (Murray, 1943)

Currently, clinicians downplay the emphasis upon imagination and intelligence when giving instructions. Surely, this omission must influence the quality of the stories produced.

Even though more than a dozen scoring systems have been proposed, interpretation of the TAT is usually based upon a clinical-qualitative analysis of the story productions. A central consideration harks back to Murray's "hero" assumption. According to this viewpoint, the hero is the protagonist of the examinee's story. It is assumed that the examinee clearly identifies with this character and projects his or her own needs, strivings, and feelings onto the hero. Conversely, thoughts, feelings, or actions avoided by the hero may represent areas of conflict for the examinee. A specific example will help clarify these points. Consider the response to Card 3 BM given by depressed examinee?

Looks like ... I can't tell if it's a girl or boy. Could be either. I guess it doesn't matter. This person just had a hard physical workout. I guess it's a her. She's just tired. No trauma happened or anything. She was sitting around a table with friends and *she got real tired. She's not in a health danger or anything. These are her keys. Her friends drag her back to her room and put her to bed. She's O.K. the next day. No trauma. She's tired physically, not mentally. (Ryan. 1987)*

What stands out in this response is the repetitive denial of danger or trauma. But later in the testing, the denial of trauma is no longer maintained. Read how the examinee responded to the blank card, relating a story of a young man traumatized at school, who takes his car down to the river:

He sees the bridge, he's really down. He remembers that he's heard stories about people jumping off and killing themselves. He could never understand why they did that. Now he understands, he jumps and dies ... he should have waited 'cause filings always get better sometime. But he didn't wait, he died. (Ryan, 1987)

Most clinicians would conclude that the examinee who produced these stories had been traumatized and was defending against self-destructive impulses. Correspondingly, the clinician would be well advised to explore these issues in psychotherapy.

The psychometric adequacy of the TAT is difficult to evaluate because of the abundance of scoring and interpretation methods. Clinicians defend the test on an anecdotal basis, pointing out remarkable and confirmatory findings such as illustrated here. However, data-minded researchers are more cautious. One problem is that formally scored TAT protocols possess very low test-retest reliability, with a reported median value of r = .28 (Winter & Stewart, 1977). Furthermore, an astonishing 97 percent of test users employ subjective and "personalized" procedures for interpreting the TAT: that is, only a tiny fraction of clinical practitioners rely upon a standardized scoring system (Lilienfeld, Wood, & Garb. 2001). This is troubling because a consistent theme in research on projective testing is that intuitive interpretations are likely to overdiagnose psychological disturbance.

In large measure, then, the interpretation of the TAT is based on strategies with unknown and untested reliability and validity. Even so, advocates of the test remain undaunted, proposing that practitioners with psychodynamic expertise can use the instrument as a "magic set of optics without which psychologists have only partial psychological vision ... a means of inferring the vital secret wishes and unconscious fantasies that participants are not able to communicate directly" (Schneidman, 1999, p. 87). Obviously, a large chasm separates enthusiastic clinical practitioners from skeptical empirical researchers in their assessment of the TAT. The latter group has made an occasional effort to develop new TAT scoring approaches that might provide a solid empirical foundation for the test (McGrew & Teglasi. 1990: Ronan. Colavito, & Hammontree, 1993). However, there is surprisingly little ongoing research on TAT scoring systems.

The Picture Projective Test

The Picture Projective Test (PPT) is a long overdue attempt to construct a general-purpose instrument with improved psychometric qualities (Ritzier, Sharkey, & Chudy, 1980: Sharkey & Ritzier. 1985). The developers of the PPT note that the majority of the TAT pictures exert a strong negative stimulus "pull" on storytelling. The TAT cards are cast in dark, shaded tones and most scenes portray persons in low-key or gloomy situations. It is not surprising, then, that projective responses to the TAT are strongly channeled toward negative, melancholic stories (Goldfried & Zax, 1965).

In contrast, the PPT uses a new set of pictures taken from the Family of Man photo essay published by the Museum of Modern Art (1955). The following criteria were used in selecting 30 pictures:

- The pictures had to show promise of eliciting meaningful projective material.
- Most but not all of the pictures had to include more than one human character.
- About half of the pictures had to depict humans showing positive affective expression (e.g., smiling, embracing, and dancing).
- About half of the pictures had to depict humans in active poses, not simply standing, sitting, or lying down.

In an initial pilot study, the authors compared TAT and PPT story productions of eight undergraduates on several variables such as length of stories, emotional tone, and activity level (Ritzier, Sharkey, & Chudy, 1980). Compared to the TAT productions, the PPT stories were of comparable length but were much more positive in thematic content and emotional tone. The PPT stories were also much more active, meaning that the central character had an active, self-determined effect on the situation in the story. Furthermore, the PPT stories placed greater emphasis upon interpersonal rather than intrapersonal themes. In other words, the PPT stories placed more emphasis on "healthy," adaptive aspects of personality adjustment than did the TAT productions.

The PPT developers also compared their instrument against the TAT in a diagnostic validity study (Sharkey & Ritzier, 1985). PPT and TAT story productions of 50 subjects were compared: normals, nonhospitalized depressives, hospitalized depressives, hospitalized psychotics with good premorbid histories, and hospitalized psychotics with poor premorbid histories (10 subjects in each group). Although the TAT and PPT were essentially equal in their capacity to discriminate normal from depressed subjects, the PPT was superior in differentiating psychotics from normals and depressives. On the PPT, depressives told stories with gloomier emotional tone and psychotics made more perceptual distortions, and thematic/interpretive deviations. The PPT appears to be a very promising instrument, although it is obvious that further research is needed on its psychometric qualities. One noteworthy feature is that anyone can purchase the PPT stimuli at their local bookstore. The requisite materials are found in the Family of Man photo collection (Museum of Modern Art, 1955).

CHILDREN'S APPERCEPTION TEST

Designed as a direct extension of the TAT, the Children's Apperception Test (CAT) consists of 10 pictures and is suitable for children 3 to 10 years of age. The preferred version for younger children (CAT-A) depicts animals in unmistakably human social settings (Bellak & Bellak, 1991). The test developers used animal drawings on the assumption that young children would identify better with animals than humans. A human figure version (CAT-H) is available for older children (Bellak & Bellak, 1994). No formal scoring system exists for the CAT and no statistical information is provided on reliability or validity. Instead, the examiner prepares a diagnosis or personality description based upon a synthesis of 10 variables recorded for each story: (1) main theme; (2) main hero; (3) main needs and drives of hero; (4) conception of environment (or world); (5) perception of parental, contemporary, and junior figures; (6) conflicts; (7) anxieties; (8) defenses; (9) adequacy of superego; (10) integration of ego (including originality of story and nature of outcome) (Bellak, 1992). The lack of attention to psychometric issues of scoring, reliability, and validity of the CAT is troublesome to most testing specialists.

Other Variations on the TAT

The TAT has inspired a number of similar tests designed for children and older adults (Table 13.7). In addition, modifications and variations of the TAT have been developed for ethnic, racial, and linguistic minorities. One of the first was the Thompson TAT (T-TAT) in which 21 of the original TAT pictures were redrawn with African American figures (Thompson, 1949). This TAT modification incorporated certain unintended changes—for example, in facial expressions and the situations portrayed. As a result, the T-TAT should be considered a new test and not just a TAT translation suited to African American individuals (Aiken, 1989).

Another specialized TAT-like test is the TEMAS, which consists of 23 colorful drawings that depict Hispanic persons interacting in contemporary, inner-city settings (Aiken, 1989; Constantino, Malgady, & Rogler, 1988). TEMAS is Spanish for themes and an acronym for "tell me a story." The thematic content of TEMAS stories is scored for 18 cognitive functions, 9 personality (ego) functions, and 7 affective functions. The test can also be scored for various objective indices such as reaction time, fluency, unanswered inquiries, and stimulus transformations (e.g., a letter is transformed into a bomb). Hispanic children respond well to the TEMAS, even though they may be inarticulate in response to traditional projective tests.

Table 13.7 Thematic Apperception Tests for Specific Populations

Adolescent Apperception Cards

This is the only thematic apperception test designed specifically for adolescents (12- to 19-year-olds). The 11 cards represent contemporary issues relevant to adolescents; themes include loneliness, parenting styles, domestic violence, gang activity, and drug abuse (Silverton, 1993). Problems with this instrument include the negative themes depicted in the cards (which preclude positive associations) and the absence of any objective approach to scoring. Like many thematic apperception techniques, the AAC is really an idiographic clinical tool, not a test.

Blacky Pictures

For children ages 5 and older, the Blacky Pictures test was also based on the premise that children identify more readily with animals than humans. The 11 cartoon stimuli depict the adventures of the dog Blacky and his family (Mama, Papa, and sibling Tippy). In addition to requesting a story for each card, the examiner also presents multiple-choice questions based on stages of psychosexual development derived from psychoanalytic theory (Blum, 1950). Although the test was originally developed with adults, children enjoy taking the Blacky and are quite responsive to the pictures. Problems with this test include the absence of norms, especially for children, and poor stability of scores (LaVoie, 1987).

Michigan Picture Test-Revised

For older children ages 8 to 14 years, the MPT-R consists of 15 pictures and a blank card. Responses are scored for Tension Index (e.g., portrayal of personal adequacy). Direction of Force (whether the central figure acts or is acted upon), and Verb Tense (e.g., past, present, future). These three scores can be combined to yield a Maladjustment Index. Reliability and norms are adequate, although evidence of validity is unsatisfactory. A major problem with this test is that the cards portray interpersonal relationships so vividly that little is left to the child's imagination (Aiken, 1989).

Senior Apperception Test (SAT)

Although the 16 situations depicted on the SAT cards include some positive circumstances, the majority of pictures were designed to reflect themes of helplessness, abandonment, disability, family problems, loneliness, dependence, and low self-esteem (Bellak, 1992). Critics complain that the SAT stereotypes the elderly and therefore discourages active responding (Schaie, 1978; Klopfer & Taulbee, 1976).

The inconsistent reliability of the TEMAS is a source of concern, because reliability constrains validity. The manual reports that Cronbach's alpha for the 34 scoring functions ranged from .31 to .98 with half below .70. Test-retest reliabilities were even lower; the highest correlation was r = .53 and for 26 of the 34 functions the correlations were near zero! In spite of the questionable reliability of the instrument, several studies provide support for its concurrent and predictive validity. For example, in a clinical sample of 210 Puerto Rican children. TEMAS scale scores predicted independent criteria of ego development, trait anxiety, and adaptive behavior reasonably well, with correlations ranging from .27 to .51 (Malgady, Constantino, & Rogler, 1984). A steady stream of research has continued to bolster the utility of this instrument, as surveyed by Constantino & Malgady (1996). Flanagan and di Guiseppe (1999) provide a critical review of the TEMAS; Constantino and Malgady (2000) describe recent developments with the test

EXPRESSION TECHNIQUES

The Draw-A-Person Test

As the reader will recall from an earlier chapter, Goodenough (1926) used the Draw-A-Man task as a basis for estimating intelligence. Subsequently, psychodynamically minded psychologists adapted the procedure to the projective assessment of personality. Karen Machover (1949, 1951) was the pioneer in this new field. Her procedure became known as the Draw-A-Person Test (DAP). Her test enjoyed early popularity and is still widely used as a clinical assessment tool. Watkins, Campbell, Nieberding, and Hallmark (1995) report that projective drawings such as the DAP rank eighth in popularity among clinicians in the United States.

The DAP is administered by presenting the examinee with a blank sheet of paper and a pencil with eraser, then asking the examinee to "draw a person." When the drawing is completed the examinee usually is directed to draw another person of the sex opposite that of the first figure. Finally, the examinee is asked to "make up a story about this person as if he [or she] were a character in a novel or a play" (Machover, 1949).

Interpretation of the DAP proceeds in an entirely clinical-intuitive manner, guided by a number of tentative psychodynamically based hypotheses (Machover, 1949, 1951). For example, Machover maintained that examinees were likely to project acceptable impulses onto the same-sex figure and unacceptable impulses onto the opposite-sex figure. She also believed that the relative sizes of the male and female figures revealed clues about the sexual identification of the examinee. Several of Machover's interpretive hypotheses are listed in Table 13.8.

Table 13.8 Illustrative Interpretations of the Draw-A-Person Test

Sign	Hypothesized Interpretive Significance
Disproportionately large head	Organic brain disease; previous brain surgery; preoccupation with headaches
Deliberate omission of facial features	Evasive about highly conflictual interpersonal relationships
Mouth drawn with heavy line slash	Verbally aggressive, over-critical, and sometimes sadistic personality
Chin changed, erased, or reinforced	Compensation for weakness, indecision, and fear of responsibility
Large male eyes with lashes	Homosexually inclined male, often very extraverted
Hair emphasis, e.g., a beard	An indication of a striving for virility
Graphic emphasis of the neck	Disturbed about the lack of control over impulses
Conspicuous treatment of index finger, thumb	Preoccupation with masturbation
Anatomical indications of internal organs	Found only in schizophrenic or actively manic patients

These interpretive premises are colorful, interesting, and plausible. However, they are based entirely upon psychodynamic theory and anecdotal observations. Machover made little effort to validate the interpretations. The empirical support for her hypotheses is somewhere between meager and nonexistent (Swensen, 1957, 1968). In favor of the DAP, the overall quality of drawings does weakly predict psychological adjustment (Lewinsohn, 1965; Yama, 1990). However, judged by contemporary standards of evidence, the sweeping and cavalier assessments of personality so often derived from the DAP are embarrassing. Some reviewers have concluded that the DAP is an unworthy test that should no longer be used (Gresham, 1993; Motta, Little, & Tobin, 1993).

Rather than using the DAP to infer nuances of personality, a more appropriate application of this test is in the screening of children suspected of behavior disorder and emotional disturbance. For this purpose, Naglieri, McNeish, and Bardos (1991) developed the Draw A Person: Screening Procedure for Emotional Disturbance (DAP:SPED). In one study, diagnostic accuracy of problem children was significantly improved by application of the DAP:SPED scoring approach (Naglieri & Pfeiffer, 1992).

The House-Tree-Person Test (H-T-P)

The H-T-P is a projective test that uses freehand drawings of a house, tree, and person (Buck, 1948, 1981). The examinee is given almost complete freedom in sketching the three objects; separate pencil and crayon drawings are requested. Although the examiner can improvise an H-T-P Test with mere blank pieces of paper, Buck (1981) recommends the use of a four-page drawing form with identification information on the first page. Pages two, three, and four are titled House, Tree, and Person. Two drawing forms are needed for each examinee, one for pencil drawings and the other for crayon drawings. Buck (1981) also provides a separate four-page form for a postdrawing interrogation phase, which consists of 60 questions designed to elicit the examinee's opinions about elements of the drawings. Many practitioners feel the post-drawing interrogation phase is not worth the extended effort. Also, the value of separate crayon drawings is questioned (Killian, 1987).

The House-Tree-Person Test has much the same familial lineage as the Draw-A-Person Test. Like the DAP Test, the H-T-P Test was originally conceived as a measure of intelligence, complete with a quantitative scoring system to appraise an approximate level of ability (Buck, 1948). However, clinicians soon abandoned the use of the H-T-P as a measure of intelligence, and it is now used almost exclusively as a projective measure of personality.

Although we will not delve into any details here, the interpretation of the H-T-P rests upon three general assumptions: the House drawing mirrors the examinee's home life and intrafamilial relationships; the Tree drawing reflects the manner in which the examinee experiences the environment; and the Person drawing echoes the examinee's interpersonal relationships. Buck (1981) provides numerous interpretive hypotheses for both quantitative and qualitative aspects of the three drawings.

The H-T-P is an alluring test that has fascinated clinicians for more than 40 years. Unfortunately, Buck (1948, 1981) has never provided any evidence to support the reliability or validity of this instrument. Indeed, he is perhaps his own worst critic. At one point in his test manual, he even asserts that validational research is not possible with the H-T-P (Buck, 1981, p. 164). Among the impediments to such research, he cites the following points:

1. No single sign itself is an infallible indication of any strength or weakness in the S.
2. No H-T-P sign has but one meaning.
3. The significance of a sign may differ markedly from one constellation to another.
4. The amount of diagnostic and prognostic data derivable from each of the points of analysis may vary greatly from S to S.
5. Colors do not have any absolute and universal meaning.
6. Nothing in the quantitative scoring system can be taken automatically at face value (Buck, 1981).

In general, attempts to validate the H-T-P as a personality measure have failed miserably (for reviews see Ellis, 1970; Hayworth, 1970; Krugman, 1970; Killian, 1987). Thoughtful reviewers have repeatedly recommended the abandonment of the H-T-P and similar figure-drawing approaches to personality assessment. But these pronouncements apparently fall on deaf ears. The popularity of the H-T-P and other projective techniques continues unabated. In the final section of this chapter, we offer some reflections on the continued acceptance of projective techniques.

REPRISE: THE PROJECTIVE PARADOX

The evidence is quite clear that personality inferences drawn from projective tests often are wrong. In the face of negative validational findings, the enduring practitioner acceptance of these tests constitutes what we have referred to as the projective paradox. How do we explain the continued popularity of instruments for which the validity evidence is at best mixed, often marginal, occasionally nonexistent, or even decisively negative?

We offer two explanations for the projective paradox. The first is that human beings cling to preexisting stereotypes even when exposed to contradictory findings. Decades ago, Chapman and Chapman (1967) demonstrated this phenomenon with projective tests, naming it illusory validation. These researchers asked college students to observe several human figure drawings similar to those obtained from the Draw-A-Person Test (DAP). The students were naive with respect to projective tests and knew nothing about traditional DAP interpretive hypotheses. Each drawing was accompanied by brief descriptions of two symptoms which supposedly characterized the patient who produced the drawing. Actually, the symptoms were assigned randomly t: drawings and consisted of the bits and pieces of DAP clinical lore that had been gleaned from an earlier mail questionnaire to clinical psychologists. For example, two of the symptoms used were these:

1. Is worried about how manly he is
2. Is suspicious of other people

Each student received a different combination of draawings and randomly assigned symptoms.

Later, the students were asked to demonstrate what they had learned by describing, for several drawings, the symptoms they had observed to be linked with that kind of drawing. Of course, in reality there was no learning to be demonstrated, since symptoms and drawings were randomly combined. Nonetheless, the participants responded in terms of popular clinical stereotypes (e.g., unusual eyes indicate suspiciousness; large head suggests a concern with intelligence). Apparently, the commonsense stereotypes held by participants emerged robust and unscathed—in spite of an abundance of disconfirming examples. Perhaps something similar occurs in all fields of projective testing: clinicians notice the confirming instances, but ignore the more numerous findings which contradict expectations.

The second explanation for the projective paradox is that many clinicians do not use projective methods as tests at all, but as auxiliary approaches to the clinical interview. These practitioners use projective techniques as clinical tools to derive tentative hypotheses about the examinee. Most of these hypotheses will turn out to be false when examined more closely. However, the few that are confirmed may have important implications for the clinical management of the examinee. Furthermore, we suspect that these fruitful hypotheses might not emerge—or might emerge more slowly—if the practitioner relied entirely upon the interview or used only formal tests with established reliability and validity. However, this assertion is difficult to test empirically. We remain open to the possibility that clinically successful applications of projective techniques largely provide further evidence of illusory validation.

CASE EXHIBIT

PROJECTIVE TESTS AS ANCILLARY TO THE INTERVIEW

A specific example may help to clarify the role of projective techniques as ancillary to the clinical interview. During the Vietnam War, a Veteran's Administration psychologist tested a young soldier who had accidentally shot himself in the leg with a forty-five calibre pistol while practicing quick draw in the jungle. Surgeons found it necessary to amputate the soldier's leg from the knee down. He was quite depressed, and everyone assumed that he suffered from grief and guilt over his great personal tragedy. He was virtually mute and nearly untestable. However, he was persuaded to complete a series of figure drawings. In one drawing he depicted himself as a helicopter gunner, spraying bullets indiscriminantly into the jungle below. When questioned about this drawing, he became quite animated and confessed that he relished combat. Guided by the possible implications of the morbid drawing, the psychologist sought to learn more about the veteran's attitudes toward combat. In the course of several interviews, the veteran revealed that he particularly enjoyed firing upon moving objects—animals, soldiers, civilians—it made no difference to him. Gradually, it became clear that the young veteran was an incipient war criminal who was depressed because his injury would prevent him from returning to the front lines. Needless to say, this information had quite an impact on the tenor of the psychological report.

Structured Personality Assessment

The history of personality assessment can be characterized by two overlapping trends. First, unstructured projective techniques such as the Rorschach test dominated personality testing in the early twentieth century and then waned in popularity. Second, structured approaches such as self-report inventories and behavioral ratings gained prominence in midcentury and then rapidly expanded in popularity. In the previous topic we introduced the reader to the many varieties of projective techniques. These methods are resplendent in the richness of the hypotheses they yield; however, projective techniques largely lack the approval of psychometrically oriented clinicians. In this chapter, we focus on the more objective methods for personality assessment favored by measurement-minded psychologists. In Topic 14A, Self-Report Inventories, we review true-false and forced-choice instruments, including the most widely used personality test ever, the Minnesota Multiphasic Personality Inventory (MMPI), and its recent revision, the MMPI-2. In Topic 14B, Behavioral Assessment and Related Approaches, we examine more recent approaches that rely upon behavioral observations and ratings.

Contemporary psychometricians have relied upon three tactics for test development: theory-bounded approaches, factor-analytic strategies, and criterion-key methods. We will organize the discussion of self-report inventories around these three categories. Of course, the boundaries are somewhat artificial and many test developers use a combination of methods.

The structured approaches to personality testing discussed in the following sections are steeped in the details of psychometric methodology. These tests feature prominent references to reliability indices, criterion keying, factor analysis, construct validation, and other forms of technical craftsmanship. For this reason, the approaches discussed here are often considered objective—as contrasted with projective. However, whether they are objective in any meaningful sense is really an empirical question that must be answered on the basis of research. Perhaps it is more accurate to call these methods structured. They are structured in the sense that highly specific rules are followed in the administration, scoring, and interpretation of the tests. In fact, some of the approaches are so completely structured that an examinee can answer questions presented on a computer screen and observe a computer-generated narrative report spewed forth from the printer seconds later.

THEORY-GUIDED INVENTORIES

The construction of several self-report inventories was guided closely by formal or informal theories of personality. In these cases, the test developer designed the instrument around a preexisting theory. Theory-guided inventories stand in contrast to factor-analytic approaches which often produce a retrospective theory based upon initial test findings. Theory-guided inventories also differ from the stark atheoretical empiricism found in criterion-key instruments such as the MMPI and MMPI-2. Examples of theory-guided inventories include the Edward Personal Preference Schedule (EPPS) and the Personality Research Form (PRF), both based on Murray's (1938) need-press theory of personality. Further examples include the Myers-Briggs Type Indicator (MBTI), which represents an application of Carl Jung's theory of personality types. The Jenkins Activity Survey, designed to assess the Type A coronary-prone behavior pattern, also epitomizes a theory-guided instrument. Finally, some theory-guided inventories such as the State-Trait Anxiety Inventory (STAI) attempt to measure very specific components of personality. Following we review each of these tests in more detail.

Edwards Personal Preference Schedule

The Edwards Personal Preference Schedule (EPPS) was the first attempt to measure Murray's (1938) manifest needs with a structured personality inventory (Edwards, 1959; Helms, 1983). The reader will recall from an earlier discussion that Murray posited 15 needs and developed a projective test, the Thematic Apperception Test, to tap those needs. Edwards, a consummative psychometrician well versed in the nuances of measurement theory, sought to develop an objective, structured test to measure those 15 needs in a more reliable and valid manner.

The EPPS consists of 210 pairs of statements in which items from each of the 15 scales are paired with items from the other 14. The inventory uses a forced-choice format in which the examinee must choose the one statement from each pair that is most personally representative. The forced-choice format of the EPPS is peculiar and uncomfortable to most test takers, because it often serves up the proverbial choice between a rock and a hard place. Here are three EPPS-like items; for each item, the examinee must choose the one statement that is most personally characteristic:

1. I like to talk in front of a group.
B. I like to work toward self-chosen goals.
2. I feel sad when I watch a tragic news story on TV.
B. I feel nervous when I have to speak before a group.
3. I wouldn't mind mopping up ten gallons of syrup.
B. I wouldn't mind scaling a steep cliff on a safety rope.

Why did Edwards adopt this awkward format for his test? The answer has to do with the problem of social desirability response set. Social desirability response set is the tendency of examinees to react to the perceived desirability (or undesirability) of a test item rather than responding accurately to its content.
Put simply, examinees tend to endorse socially desirable statements and tend not to endorse socially undesirable statements—regardless of the truth value of the responses. Most persons would respond true to a statement such as "I enjoy helping older persons across the street" because the item sanctions a socially desirable attribute; and most persons would respond false to a statement such as "At times I have fantasized about the death of my parents" because the item authorizes a socially undesirable quality. But for some persons, the socially desirable answer is not really accurate. After all, in truth many persons really do not enjoy helping others, and most individuals have fantasized about unpleasant possibilities.

The elegance of the EPPS is that pairs of statements in each item are matched for social desirability (Edwards. 1957).

1. Theory-guided self-report inventories rely upon explicit personality theories for their development. A good example of a theory-guided inventory is the Edwards Personal Preference Schedule (EPPS), a 210-item forced-choice instrument that attempts to measure Murray's manifest needs by self-report.

2. Jackson's Personality Research Form (PRF) is also based upon Murray's need system. The 20 personality scales on the PRF possess no item overlap and show exceptional internal consistency (median of .92). PRF validity is buttressed by confirmatory factor analysis and appropriate correlations with similar scales on other instruments.

3. The Myers-Briggs Type Indicator (MBTI) is a forced-choice self-report inventory based loosely upon Carl Jung's theory of personality types. The MBTI is scored for four dimensions: Extraversion-Introversion, Sensing-intuition, Thinking-Feeling, and Judging-Perceptive, yielding 16 different types, such as ENFP.

4. The Jenkins Activity Survey (JAS) is a 52-item multiple-choice questionnaire designed to identify the Type A coronary-prone behavior pattern. The three subscales include: Speed and Impatience, Job Involvement, and Hard-Driving and Competitiveness. The JAS has several limitations (e.g., unrepresentative norms, scoring complexities) and is therefore best suited to research.

5. A short, simple test that has received high marks for technical merit is the State-Trait Anxiety Inventory (STAI). The 40 items of the STAI are each rated on a four-point intensity scale. The STAI measures state anxiety, or transitory feelings of fear or worry; and trait anxiety, the relatively stable tendency to respond anxiously to stressful situations.

6. Cattell's Sixteen Personality Factor Questionnaire (16PF) is typical of factor-analytically derived instruments. The five forms of the 16PF (for different age groups) all encompass a forced-choice format. The 16 surveyed personality attributes (and four higher-order dimensions) have been repeatedly confirmed by factor analysis.

7. The Eysenck Personality Questionnaire (EPQ) proposes three major factor-analytically derived dimensions of personality: Psychoticism, Extraversion, and Neuroticism. Scale reliabilities are quite strong and the construct validity of the instrument is supported by dozens of studies.

8. The Comrey Personality Scales embody a short self-report instrument suitable for college students. The eight CPS scales consist of 20 items each and possess no overlap. The scales show excellent internal consistency. Extreme scores are especially predictive of psychological disturbance.

9. The NEO Personality Inventory-Revised (NEO PI-R) is based upon the five-factor model of personality described earlier. The five constructs measured by the test are Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. The NEO PI-R is available in two parallel forms consisting of 240 items rated on a five-point dimension.

10. The MMPI-2 consists of 567 true-false questions. The test is scored for four validity scales (?, L, F, and K) that assess unanswered questions, naive defensiveness, deviant responses, and subtle defensiveness, respectively. The 10 clinical scales are Hypochondriasis, Depression, Hysteria, Psychopathic Deviate, Masculinity-Femininity, Paranoia, Psychasthenia, Schizophrenia, Hypomania, and Social Introversion.

11. The California Psychological Inventory (CPI) is an MMPI-like instrument designed to measure the dimensions of normal personality. Three scales measure test-taking attitudes (e.g., "fake good" and "fake bad" tendencies). The 17 clinical scales are based upon "folk" concepts of personality easily recognized by laypersons.

12. The Millon Clinical Multiaxial Inventory, now in its third edition (MCMI-HI) is a short test (175 true-false items) designed as an aid to psychiatric diagnosis. The 27 scales are organized into four broad categories relevant to DSM-IV: clinical personality patterns, severe personality pathology, clinical syndromes, and severe clinical syndromes.

13. Designed to provide clinically relevant descriptions of child behavior and family characteristics, the Personality Inventory for Children-2 (PIC-2) consists of 275 true-false statements that are completed by a parent or parental surrogate. The test is suitable for children 5 through 19 years of age and yields scores on 9 adjustment scales and 21 subscales.

Lesson 14

APTITUDE AND ACHIEVEMENT TESTING

Aptitude Tests and Factor Analysis

In this chapter, we examine a variety of instruments traditionally grouped under the headings of aptitude tests and achievement tests. The coverage includes relevant instruments, but also embraces issues and applications in aptitude and achievement testing. Aptitude Tests and Factor Analysis, the use of factor analysis in the development of aptitude measures is described. This is followed by a review of typical instruments, including multiple aptitude test batteries and tests used to predict academic performance in college. Then in Group Tests of Achievement, we examine the educational achievement test batteries familiar to every student of American schooling. In addition, the reader will encounter a brief discussion of special-purpose tests for achievement as well as a review of troubling social issues that pertain to school system cheating on achievement tests.

Here we focus on aptitude tests, especially the multiple aptitude batteries commonly used to predict performance in school, employment, and military settings. Typically, multiple aptitude batteries perform a gatekeeper function. School admission, corporate employment, and military entry may hinge upon findings from the tests discussed here. Aptitude tests command great respect and therefore possess immense influence in modern society. The validity of aptitude tests is indeed consequential. The reader will learn more about the application of aptitude tests later in this topic.

Many aptitude tests arose as specialized offshoots of ability tests shortly after psychologists developed the necessary statistical tools for portioning general intelligence into its subcomponents. Put simply, most aptitude tests owe their origin to factor analysis, a family of procedures that researchers use to summarize relationships among variables that are correlated in highly complex ways. Because aptitude tests could not flourish without factor analysis, we begin this section with a primer of this useful statistical technique. The topic then continues with a discussion of prominent tests of aptitude, including multi-aptitude batteries useful for employment counseling (Differential Aptitude Test, General Aptitude Test Battery, and Armed Services Vocational Assessment Battery), tests used for college admission (Scholastic Assessment Tests and American College Test), and postgraduate admission tests (Graduate Record Exam, Medical College Admission Test, and Law School Admission Test).

A PRIMER OF FACTOR ANALYSIS

Broadly speaking, there are two forms of factor analysis: confirmatory and exploratory. In confirmatory factor analysis, the purpose is to confirm that test scores and variables fit a certain pattern predicted by a theory. For example, if the theory underlying a certain intelligence test prescribed that the subtests belong to three factors (e.g., verbal, performance, and attention factors), then a confirmatory factor analysis could be undertaken to evaluate the accuracy of this prediction. Confirmatory factor analysis is essential to the validation of many ability tests.

The central purpose of exploratory factor analysis is to summarize the interrelationships among a large number of variables in a concise and accurate manner as an aid in conceptualization (Gorsuch, 1983). For instance, factor analysis may help a researcher discover that a battery of 20 tests represents only four underlying variables, called factors. The smaller set of derived factors can be used to represent the essential constructs that underlie the complete group of variables.

Perhaps a simple analogy will clarify the nature of factors and their relationship to the variables or tests from which they are derived. Consider the track-and-field decathlon, a mixture of 10 diverse events including sprints, hurdles, and pole vault, shot put, and distance races, among others. In conceptualizing the capability of the individual decathlete, we do not think exclusively in terms of the participant's skill in specific events. Instead, we think in terms of more basic attributes such as speed, strength, coordination, and endurance, each of which is reflected to a different extent in the individual events. For example, the pole vault requires speed and coordination, while hurdle events demand coordination and endurance. These infer attributes are analogous to the underlying factors of factor analysis. Just as the results from the 10 events of a decathlon may boil down to a small number of underlying factors (e.g., speed, strength coordination, and endurance), so too may the results from a battery of 10 or 20 ability tests reflect the operation of a small number of basic cognitive attributes (e.g., verbal skill, visualization, calculation, and attention, to cite a hypothetical list). This example illustrates the goal of factor analysis: to help produce a parsimonious description of large, complex data sets.

We will illustrate the essential concepts of factor analysis by pursuing a classic example concerned with the number and kind of factors that best describe student abilities. Holzinger and Swineford (1939) gave 24 ability-related psychological to 145 junior high school students from Park, Illinois. The factor analysis described 1 was based upon methods outlined in Kinne Gray (1997).

It should be intuitively obvious to the reader that any large battery of ability tests will reflect a smaller number of basic, underlying abilities (factors). Consider the 24 tests depicted in Table 14.1. Surely some of these tests measure common underlying abilities. For example, we would expect Sentence Completion, Word Classification and Word Meaning (variables 7, 8, and 9) to assess a factor of general language ability of some kind. In like manner, other groups of tests seem likely to measure common underlying abilities. But how many abilities or factors? And what is the nature of these underlying abilities'7 Factor analysis the ideal tool for answering these questions. We follow the factor analysis of the Holzinger and Swineford (1939) data from beginning to end.

Table 14.1 The 24 Ability Tests Used by Holzinger and Swineford (1939)

1.	Visual Perception	13.	Straight and Curved Capitals
2.	Cubes	14.	Word Recognition
3.	Paper Form Board	15.	Number Recognition
4.	Flags	16.	Figure Recognition
5.	General Information	17.	Object-Number
6.	Paragraph Comprehension	18.	Number-Figure
7.	Sentence Completion	19.	Figure-Word
8.	Word Classification	20.	Deduction
9.	Word Meaning	21.	Numerical Puzzles
10.	Add Digits	22.	Problem Reasoning
11.	Code (Perceptual Speed)	23.	Series Completion
12.	Count Groups of Dots	24.	Arithmetic Problems

The Correlation Matrix

The beginning point for every factor analysis is the correlation matrix, a complete table of intercorrelations among all the variables. The correlations between the 24 ability variables discussed here can be found in Table 14.2. The reader will notice that variables 7, 8, and 9 do, indeed, intercorrelate quite strongly (correlations of .62, .69, and .53), as we suspected earlier. This pattern of inter-correlations is presumptive evidence that these variables measure something in common; that is, it appears that these tests reflect a common underlying factor. However, this kind of intuitive factor analysis based upon a visual inspection of the correlation matrix is hopelessly limited; there are just too many intercorrelations for the viewer to discern the underlying patterns for all the variables. Here is where factor analysis can be helpful. Although we cannot elucidate the mechanics of the procedure, factor analysis relies upon modern high speed computers to search the correlation matrix according to objective statistical rules and determine the smallest number of factors needed to account for the observed pattern of intercorrelations. The analysis also produces the factor matrix, a table showing the extent to which each test loads on (correlates with) each of the derived factors, as discussed in the following section.

The Factor Matrix and Factor Loadings

The factor matrix consists of a table of correlations called factor loadings. The factor loadings (which can take on values from -1.00 to +1.00) indicate the weighting of each variable on each factor. For example, the factor matrix in Table 14.3 shows that five factors (labeled I, II, III, IV, and V) were derived from the analysis. Note that the first variable, Series Completion, has a strong positive loading of .71 on factor I. indicating that this test is a reasonably good index of factor I. Note also that Series Completion has a modest negative loading of -.11 on factor II, indicating that, to a slight extent, it measures the opposite of this factor; that is high scores on Series Completion tend to signify low scores on factor II, and vice versa.

Table 14.2 The Correlation Matrix for 24 Ability Variables

2	32																				
3	40	32																			
4	47	23	31																		
5	32	29	25	23																	
6	34	23	27	33	62																
7	30	16	22	34	66	72															
8	33	17	38	39	58	53	62														
9	33	20	18	33	72	71	69	53													
10	12	06	08	10	31	20	25	29	17												
11	31	15	09	11	34	35	23	30	28	48											
12	31	15	14	16	22	10	18	27	11	59	43										
13	49	24	32	33	34	31	35	40	28	41	54	51									
14	13	10	18	07	28	29	24	25	26	17	35	13	20								
15	24	13	07	13	23	25	17	18	25	1524171437											
16	41	27	26	32	19	29	18	30	24	12	31	12	28	41	33						
17	18	01	18	19	21	27	23	26	27	29	36	28	19	34	35	32					
18	37	26	21	25	26	17	16	25	21	32	35	35	32	21	33	34	45				
19	27	11	31	14	19	25	23	27	27	19	29	11	26	21	19	26	32	36			
20	37	29	30	34	40	44	45	43	45	17	20	25	24	30	27	39	26	30	17		
21	37	31	17	35	32	26	31	36	27	41	40	36	43	18	23	35	17	36	33	41	
22	41	23	25	38	44	39	40	36	48	16	30	19	28	24	25	28	27	32	34	46	37
23	47	35	38	34	44	43	41	50	50	26	25	35	38	24	26	36	29	27	30	51	45
24	28	21	20	25	42	43	44	39	42	53	41	41	36	30	17	26	33	41	37	37	45

The factors may seem quite mysterious, but in reality they are conceptually quite simple. A factor is nothing more than a weighted linear sum of the variables; that is, each factor is a precise statistical combination of the tests used in the analysis. In a sense, a factor is produced by "adding in" carefully determined portions of some tests and perhaps "subtracting out" fractions of other tests. What makes the factors special is the elegant analytical methods used to derive them. Several different methods exist. These methods differ in subtle ways beyond the scope of this text; the reader can gather a sense of the differences by examining names of procedures: principal components factors, principal axis factors, and method of unweighted least squares, maximum likelihood method, image factoring, and alpha factoring (Tabachnick & Fidell, 1989). Most of the methods yield highly similar results.

The factor loadings depicted in Table 14.3 are nothing more than correlation coefficients between variables and factors. These correlations can be interpreted as showing the weight or loading of each factor on each variable. For example, variable 9, the test of Word Meaning, has a very strong loading (.69) on factor I, modest negative loadings (-.45 and -.29) on factors II and HI, and negligible loadings (.08 and .00) on factors IV and V.

Table 14.3 The Principal-Axes Factor Analysis for 24 Variables

		Factors				
		I	II	III	IV	V
23.	Series Completion	.71	-.11	.14	.11	.07
8.	Word Classification	.70	-.24	-.15	-.11	-.13
5.	General Information	.70	-.32	-.34	-.04	.08
9.	Word Meaning	.69	-.45	-.29	.08	.00
6.	Paragraph Comprehension	.69	-.42	-.26	.08	-.01
7.	Sentence Completion	.68	-.42	-.36	-.05	-.05
24.	Arithmetic Problems	.67	.20	-.23	-.04	-.11
20.	Deduction	.64	-.19	.13	.06	.28
22.	Problem Reasoning	.64	-.15	.11	.05	-.04
21.	Numerical Puzzles	.62	.24	.10	-.21	.16
13.	Straight and Curved Capitals	.62	.28	.02	-.36	-.07
1.	Visual Perception	.62	-.01	.42	-.21	-.01
11.	Code (Perceptual Speed)	.57	.44	-.20	.04	.01
18.	Number-Figure	.55	.39	.20	.15	-.11
16.	Figure Recognition	.53	.08	.40	.31	.19
4.	Flags	.51	-.18	.32	-.23	-.02
17.	Object-Number	.49	.27	-.03	.47	-.24
2.	Cubes	.40	-.08	.39	-.23	.34
12.	Count Groups of Dots	.48	.55	-.14	-.33	.11
10.	Add Digits	.47	.55	-.45	-.19	.07
3.	Paper Form Board	.44	-.19	.48	-.12	-.36
14.	Word Recognition	.45	.09	-.03	.55	.16
15.	Number Recognition	.42	.14	.10	.52	.31
19.	Figure-Word	.47	.14	.13	.20	-.61

Geometric Representation of Factor Loadings

It is customary to represent the first two or three factors as reference axes in two or three dimensional space. Within this framework the factor loadings for each variable can be plotted for examination. In our example, five factors were discovered, too many for simple visualization. Nonetheless, we can illustrate the value of geometric representation by oversimplifying somewhat and depicting just the first two factors (Figure 14.1). In this graph, each of the 24 tests has been plotted against the two factors that correspond to axes I and II. The reader will notice that the factor loadings on the first factor (I) are uniformly positive, whereas the factor loadings on the second factor (II) consist of a mixture of positive and negative.

The Rotated Factor Matrix

An important point in this context is that the position of the reference axes is arbitrary.

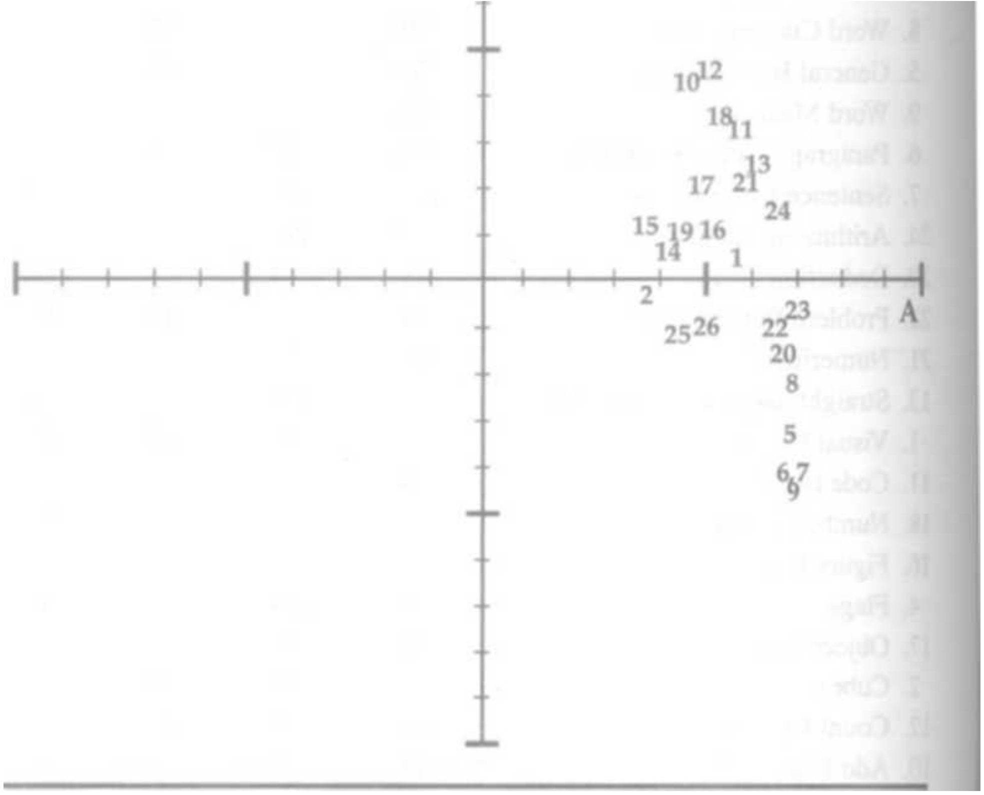


Figure 14.1 Geometric Representation of the First Two Factors from 24 Ability Tests

There is nothing to prevent the researcher from rotating the axes so that they produce a more sensible fit with the factor loadings. For example, the reader will notice in Figure 14.1 that tests 6,7, and 9 (all language tests) cluster together. It would certainly clarify the interpretation of factor I if it were to be redirected near the center of this cluster (Figure 14.2). This manipulation would also bring factor II alongside interpretable tests 10, 11, and 12 (all number tests).

Although rotation can be conducted manually by visual inspection, it is more typical for researchers to rely upon one or more objective statistical criteria to produce the final rotated factor matrix. Thurstone's (1947) criteria of positive manifold and simple structure are commonly applied. In a rotation to positive manifold, the computer program seeks to eliminate as many of the negative factor loadings as possible. Negative factor loadings make little sense in ability testing, because they imply that high scores on a factor are correlated with poor test performance. In a rotation to simple structure, the computer program seeks to simplify the factor loadings so that each test has significant loadings on as few factors as possible. The goal of both criteria is to produce a rotated factor matrix that is as straightforward and unambiguous as possible.

The rotated factor matrix for this problem is shown in Table 14.4. The particular method of rotation used here is called varimax rotation. Varimax should not be used if the theoretical expectation suggests that a general factor may occur. Should we expect a general factor in the analysis of ability tests? The answer is as much a matter of faith as of science. One researcher may conclude that a general factor is likely and therefore pursue a different type of rotation.

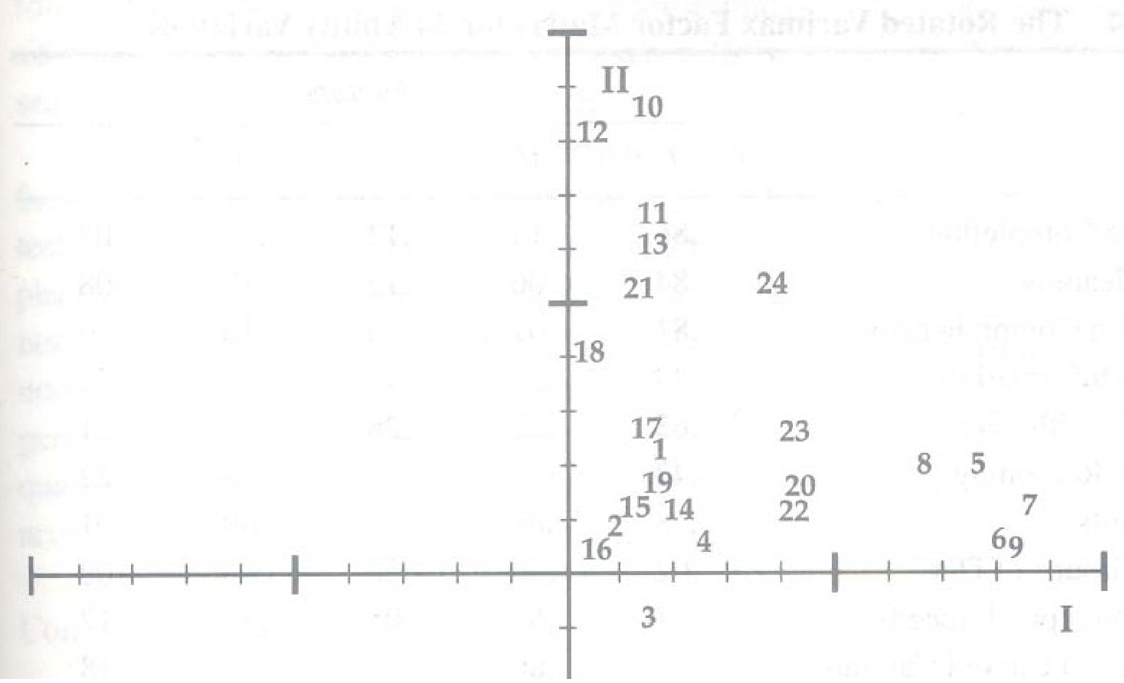


Figure 14.2 Geometric Representation of the First Two Rotated Factors from 24 Ability Tests

A second researcher may be comfortable with a Thurstonian viewpoint and seek multiple ability factors using a varimax rotation. We will explore this issue in more detail later, but it is worth pointing out here that a researcher encounters many choice points in the process of conducting a factor analysis. It is not surprising, then, that different researchers may reach different conclusions from factor analysis, even when they are analyzing the same data set.

The Interpretation of Factors

Table 14.4 indicates that five factors underlie the intercorrelations of the 24 ability tests. But what shall we call these factors? The reader may find the answer to this question disquieting, because at this juncture we leave the realm of cold, objective statistics and enter the arena of judgment, insight, and presumption. In order to interpret or name a factor, the researcher must make a reasoned judgment about the common processes and abilities shared by the tests with strong loadings on that factor. For example, in Table 14.4 it appears that factor I is verbal ability, because the variables with high loadings stress verbal skill (e.g., Sentence Completion loads .86, Word Meaning loads .84, and Paragraph Comprehension loads .81). The variables with low loadings also help sharpen the meaning of factor I. For example, factor I is not related to numerical skill (Numerical Puzzles loads .18) or spatial skill (Paper Form Board loads .16). Using a similar form of inference, it appears that factor II is mainly numerical ability (Add Digits loads .85, Count Groups of Dots loads .80). Factor III is less certain but appears to be a visual-perceptual capacity, and factor IV appears to be a measure of recognition. We would need to analyze the single test on factor V (Figure-Word) to surmise the meaning of this factor.

Table 14.4 The Rotated Varimax Factor Matrix for 24 Ability Variables

	Factors				
	I	II	III	IV	V
Sentence Completion	.86	.15	.13	.03	.07

Word Meaning	.84	.06	.15	.18	.08
Paragraph Comprehension	.81	.07	.16	.18	.10
General Information	.79	.22	.16	.12	-.02
Word Classification	.65	.22	.28	.03	.21
Problem Reasoning	.43	.12	.38	.23	.22
Add Digits	.18	.85	-.10	.09	-.01
Count Groups of Dots	.02	.80	.20	.03	.00
Code (Perceptual Speed)	.18	.64	.05	.30	.17
Straight and Curved Capitals	.19	.60	.40	-.05	.18
Arithmetic Problems	.41	.54	.12	.16	.24
Numerical Puzzles	.18	.52	.45	.16	.02
Number-Figure	.00	.40	.28	.38	.36
Visual Perception	.17	.21	.69	.10	.20
Cubes	.09	.09	.65	.12	-.18
Flags	.26	.07	.60	-.01	.15
Paper Form Board	.16	-.09	.57	-.05	.49
Series Completion	.42	.24	.52	.18	.11
Deduction	.43	.11	.47	.35	-.07
Number Recognition	.11	.09	.12	.74	-.02
Word Recognition	.23	.10	.00	.69	.10
Figure Recognition	.07	.07	.46	.59	.14
Object-Number	.15	.25	-.06	.52	.49
Figure-Word	.16	.16	.11	.14	.77

These results illustrate a major use of factor analysis, namely, the identification of a small number of marker tests from a large test battery. Rather than using a cumbersome battery of 24 tests, a researcher could gain nearly the same information by carefully selecting several tests with strong loadings on the five factors. For example, the first factor is well represented by test 7, Sentence Completion (.86) and test 9, Word Meaning (.84); the second factor is reflected in test 10, Add Digits (.85), while the third factor is best illustrated by test 1, Visual Perception (.69). The fourth factor is captured by test 15, Number Recognition (.74) and Word Recognition (.69). Of course, the last factor loads well on only test 19, Figure-Word (.77).

Issues in Factor Analysis

Unfortunately, factor analysis is frequently misunderstood and often misused. Some researchers appear to use factor analysis as a kind of divining rod, hoping to find gold hidden underneath tons of dirt. But there is nothing magical about the technique. No amount of statistical analysis can rescue data based on trivial, irrelevant, or haphazard measures. If there is no gold to be found, then none will be found; factor analysis is not alchemy. Factor analysis will yield meaningful results only when the research was meaningful to begin with.

An important point is that a particular kind of factor can emerge from factor analysis only if the tests and measures contain that factor in the first place. For example, a short-term memory factor cannot possibly emerge from a battery of ability tests if none of the tests requires short-term memory. In general, the quality of the output depends upon the quality of the input. We can restate this point as the acronym GIGO, or "garbage in, garbage out."

Sample size is crucial to a stable factor analysis. Comrey (1973) offers the following rough guide:

Sample Size	**Rating**
50	Very poor
100	Poor
200	Fair

300	Good
500	Very good
1,000	Excellent

In general, it is comforting to have at least five subjects for each test or variable (Tabachnick & Fidel 1, 1989).

Finally, we cannot overemphasize the extent to which factor analysis is guided by subjective choices and theoretical prejudices. A crucial question in this regard is the choice between orthogonal axes and oblique axes. With orthogonal axes, the factors are at right angles to one another, which means that they are uncorrelated (Figures 14.1 and 14.2 both depict orthogonal axes). In many cases the clusters of factor loadings are situated such that oblique axes provide a better fit. With oblique axes, the factors are correlated among themselves. Some researchers contend that oblique axes should always be used, whereas others take a more experimental approach. Tabachnick and Fidell (1989) recommend an exploratory strategy based on repeated factor analyses.

Their approach is unabashedly opportunistic:

> *During the next few runs, researchers experiment with different numbers of factors, different extraction techniques, and both orthogonal and oblique rotations. Some number of factors with some combination of extraction and rotation produces the solution with the greatest scientific utility, consistency, and meaning; this is the solution that is interpreted.*

With oblique rotations it is also possible to factor analyze the factors themselves. Such a procedure may yield one or more second-order factors. Second-order factors can provide support for the hierarchical organization of traits and may offer a rapprochement between ability theorists who posit a single general factor (e.g., Spearman) and those who promote several group factors (e.g., Thurstone). Perhaps both camps are correct, with the group factors sitting underneath the second-order general factor.

MULTIPLE APTITUDE TEST BATTERIES

As previously noted aptitude tests did not flourish until the prerequisite statistical tools factor analytic procedures—were available. One of the major applications of factor analysis was the development of multiple aptitude test batteries. In a multiple aptitude test battery, the examinee is tested in several separate, homogeneous aptitude areas. The development of the subtests is dictated by the findings of factor analysis. For example, Thurstone developed one of the first multiple aptitude test batteries, the Primary Mental Abilities Test, a set of seven tests chosen on the basis of factor analysis (Thurstone, 1938).

More recently, several multiple aptitude test batteries have gained favor for educational and career counseling, vocational placement, and armed services classification (Gregory, 1994a). Each year hundreds of thousands of persons are administered one of these prominent batteries: The Differential Aptitude Test (DAT), the General Aptitude Test Battery (GATB), and the Armed Services Vocational Aptitude Battery (ASVAB). These batteries either used factor analysis directly for the delineation of useful subtests or were guided in their construction by the accumulated results of other factor-analytic research. The salient characteristics of each battery are briefly reviewed in the following sections.

The Differential Aptitude Test (DAT)

The DAT was first issued in 1947 to provide a basis for the educational and vocational guidance of students in grades seven through twelve. Subsequently, examiners have found the test useful in the vocational counseling of young adults out of school and in the selection of employees. Now in its fifth edition (1992), the test has been periodically revised and stands as one of the most popular multiple aptitude test batteries of all time (Bennett, Seashore, & Wesman, 1982, 1984).

The DAT consists of eight independent tests:

1. Verbal Reasoning (VR)
2. Numerical Reasoning (NR)
3. Abstract Reasoning (AR)
4. Perceptual Speed and Accuracy (PSA)
5. Mechanical Reasoning (MR)
6. Space Relations (SR)
7. Spelling (S)
8. Language Usage (LU)

A characteristic item from each test is shown in Figure 14.3.

The authors chose the areas for the eight tests based on experimental and experiential data rather than relying upon a formal factor analysis of their own. In constructing the DAT, the authors were guided by several explicit criteria:

- Each test should be an independent test: There are situations in which only part of the battery is required or desired.

- The tests should measure power: For most vocational purposes to which test results contribute, the evaluation of power—solving difficult problems with adequate time—is of primary concern.

- The test battery should yield a profile: The eight separate scores can be converted to percentile ranks and plotted on a common profile chart.

- The norms should be adequate: In the fifth edition, the norms are derived from 100,000 students for the fall standardization, 70,000 for the spring standardization.

- The test materials should be practical: With time limits of 6 to 30 minutes per test, the entire DAT can be administered in a morning or an afternoon school session.

- The tests should be easy to administer: Each test contains excellent "warm up" examples and can be administered by persons with a minimum < special training.

- Alternate forms should be available: For purposes of retesting, the availability of alternate forms (currently forms C and D) will reduce any practice effects.

VERBAL REASONING

Choose the correct pair of words to fill in the blanks.

________ is to eye as eardrum is to _________

A. vision — sound   D. sight — cochlea
B. iris — hear   E. eyelash — earlobe
C. retina — ear

NUMERICAL ABILITY

Choose the correct answer.

4(-5) (-3) =

A. -60   B. 27   C.-27   D. 60   E. none of these

ABSTRACT REASONING

The four figures in the row to the left make a series. Find the single choice on the right that would be next in the series.

<   <»   «»   «»»   <>   «<»   «<»»   ««»»

A B C D

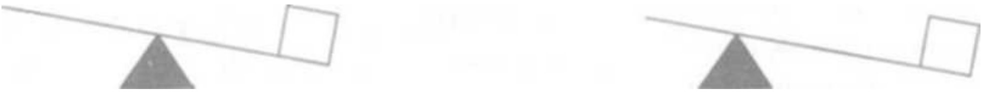
CLERICAL SPEED AND ACCURACY

In each test item, one of the combinations is underlined. Mark the same combination on the answer sheet.

1.	AB	Ab	AA	BA	Bb	2.	5m	5M	M5	Mm	m5
	Ab	Bb	AA	BA	AB		M5	m5	Mm	5m	5M
1.	O	O	O	O	O	2.	O	O	O	O	O

MECHANICAL REASONING

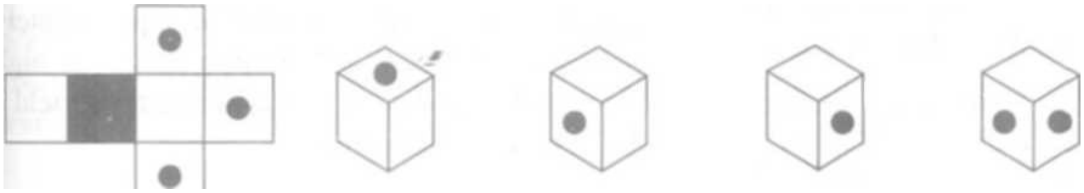
Which lever will require more force to lift an object of the same weight? If equal, mark C.



A C (equal) B

SPACE RELATIONS

Which of the figures on the right can be made by folding the pattern at the left? The pattern always displays the outside of the figure.



A B C D

SPELLING

Mark whether each word is spelled right or wrong.

1. irelevant R W
2. parsimonious R W
3. excellant R W

LANGUAGE USAGE

Decide which part of the sentence contains an error and mark the corresponding letter on the answer sheet. Mark N (None) if there is no error.

In spite of public criticism, / the researcher studied /
A B
the affects of radiation / on plant growth.
C D

Figure 14.3 Differential Aptitude Tests and Characteristic Items

The reliability of the DAT is generally quite high, with split-half coefficients largely in the .90s and alternate-forms reliabilities ranging from .73 to .90, with a median of .83. Mechanical Reasoning is an exception, with reliabilities as low as .70 for girls. The tests show a mixed pattern of intercorrelations with each other, which is optimistically interpreted by the authors as establishing the independence of the eight tests. Actually, many of the correlations are quite high and it seems likely that the eight tests reflect a smaller number of ability factors. Certainly, the Verbal Reasoning and Numerical Reasoning tests measure a healthy general factor, with correlations around .70 in various samples.

The manual presents extensive data demonstrating that the DAT tests, especially the VR + NR combination, are good predictors of other criteria such as school grades and scores on other aptitude tests (correlations in the .60s and .70s). For this reason, the combination of VR + NR often is considered an index of scholastic aptitude. Evidence for the differential validity of the other tests is rather slim. Bennett, Seashore, and Wesman (1974) do present results of several follow-up studies correlating vocational entry/success with DAT profiles, but their research methods are more impressionistic than quantitative; the independent observer will find it difficult to make use of their results. Schmitt (1995) notes that a major problem with the battery is the

> *lack of discriminant validity between the eight subtests. With the exception of the Perceptual Speed and Accuracy test, all of the subscales are highly intercorrelated (.50 to .75). If one wants only a general index of the person's academic ability, this is line; if the scores on the subtests are to be used in some diagnostic sense; this level of intercorrelation makes statements about students' relative strengths and weaknesses highly questionable.*

Even so, the revised DAT is better than previous editions. One significant improvement is the elimination of apparent sex bias on the Language Usage and Mechanical Reasoning tests—a source of criticism from earlier reviews. The DAT has been translated into several languages and is widely used in Europe for vocational guidance and research applications (e.g., Nijenhuis, Evers, & Mur, 2000; Colom, Quiroga, & Juan-Espinosa, 1999).

The General Aptitude Test Battery (GATB)

In the late 1930s, the U.S. Department of Labor developed aptitude tests to predict job performance in 100 specific occupations. In the 1940s, the department hired a panel of experts in measurement and industrial-organizational psychology to create a multiple aptitude test battery to assess the 100 occupations previously studied and many more. The outcome of this Herculean effort was the General Aptitude Test Battery (GATB), widely acknowledged as the premiere test battery for predicting job performance (Hunter, 1994).

The GATB was derived from a factor analysis of 59 tests administered to thousands of male trainees in vocational courses (United States Employment Service, 1970). The interpretive standards have been periodically revised and updated, so the GATB is a thoroughly modern instrument even though its content is little changed. One limitation is that the battery is available mainly to state employment offices, although nonprofit organizations, including high schools and certain colleges, can make special arrangements for its use.

The GATB is composed of eight paper-and-pencil tests and four apparatus measures. The entire battery can be administered in approximately two and a half hours and is appropriate for high school seniors and adults. The twelve tests yield a total of nine factor scores:

- *General Learning Ability* (intelligence) (G). This score is a composite of Vocabulary, Arithmetic Reasoning, and Three-Dimensional Space.

- *Verbal Aptitude* (V). Derived from a Vocabulary test that requires the examinee to indicate which two words in a set are either synonyms or antonyms.

- *Numerical Aptitude* (N). This score is a composite of both the Computation and Arithmetic Reasoning tests.

- *Spatial Aptitude* (S). Consists of the Three-Dimensional Space test, a measure of the ability to perceive two-dimensional representations of three-dimensional objects and to visualize movement in three dimensions.

- *Form Perception* (P). This score is a composite of Form Matching and Tool Matching, two tests in which the examinee must match identical drawings.

- *Clerical Perception* (Q). A proofreading test called Name Comparison, the examinee must match names under pressure of time.

- *Motor Coordination* (K). Measures the ability to quickly make specified pencil marks in the Mark Making test.

- *Finger Dexterity* (F). A composite of the Assemble and Disassemble tests, two measures of dexterity with rivets and washers.

- *Manual Dexterity* (M). A composite of Place and Turn, two tests requiring the examinee to transfer and reverse pegs in a board.

The nine factor scores on the GATB are expressed as standard scores with a mean of 100 and an SD of 20. These standard scores are anchored to the original normative sample of 4,000 workers obtained in the 1940s. Alternate-forms reliability coefficients for factor scores range from the .80s to the .90s. The GATB manual summarizes several studies of the validity of the test, primarily in terms of its correlation with relevant criterion measures. Hunter (1994) notes that GATB scores predict training success for all levels of job complexity. The average validity coefficient is a phenomenal .62.

The absolute scores are of less interest than their comparison to updated Occupational Aptitude Patterns (OAPs) for dozens of occupations. Based on test results for huge samples of applicants and employees in different occupations, counselors and employers now have access to a wealth of information about score patterns needed for success in a variety of jobs. Thus, one way of using the GATB is to compare an examinee's scores with OAPs believed necessary for proficiency in various occupations.

Hunter (1994) recommends an alternative strategy based on composite aptitudes (Figure 14.4). The nine specific factor scores combine nicely into three general factors: Cognitive, Perceptual, and Psychomotor. Hunter notes that different jobs require various contributions of the Cognitive, Perceptual, and Psychomotor aptitudes. For example, an assembly line worker in an automotive plant might need high scores on the Psychomotor and Perceptual composites, whereas the Cognitive score would be less important for this occupation. Hunter's research demonstrates that general factors dominate over specific factors in the prediction of job performance. Davison, Gasser, and Ding (1996) discuss additional approaches to GATB profile analysis and interpretation.

SPECIFIC FACTORS		GENERAL FACTORS
G	General Learning Ability (intelligence)	Cognitive
V	Verbal Aptitude	
N	Numerical Aptitude	
S	Spatial Aptitude	Perceptual
P	Form Perception	
Q	Clerical Perception	
K	Motor Coordination	

F	Finger Dexterity	Psychomotor
M	Manual Dexterity	

Figure 14.4 Specific and General Factors on the GATB

Van de Vijver and Harsveld (1994) investigated the equivalence of their computerized version of the GATB with the traditional paper-and-pencil version. Of course, only the cognitive and perceptual subtests were compared—tests of motor skills cannot be computerized. They found that the two versions were not equivalent. In particular, the computerized subtests produced faster and more inaccurate responses than the conventional subtests. Their research demonstrates once again that the equivalence of traditional and computerized versions of a test should not be assumed. This is an empirical question answerable only with careful research. Nijenhuis and vander Flier (1997) discuss a Dutch version of the GATB and its application in the study of cognitive differences between immigrants and majority group members in the Netherlands.

The Armed Services Vocational Aptitude Battery (ASVAB)

The ASVAB is probably the most widely used aptitude test in existence. This instrument is used by the Armed Services to screen potential recruits and to assign personnel to different jobs and training programs. The ASVAB is also available in a computerized version that is rapidly supplanting the original paper-and-pencil test (Segall & Moreno. 1999). The computerized ASVAB is discussed in more detail at the end of this section. More than 2 million examinees take the ASVAB each year. The current version consists of ten subtests, four of which produce the Armed Forces Qualification Test (AFQT), the common qualifying exam for all services (Table 14.5). Eight subtests are power tests with adequate time limits for most subjects, whereas two subtests (Numerical Operations and Coding Speed) are speeded tests that place a premium upon rapid performance. Alternate-forms reliability coefficients for ASVAB scores are in the mid-.80s to mid-.90s, and test-retest coefficients range from the mid-.70s to the mid-.80s (Larson, 1994). The one exception is Paragraph Comprehension with a reliability of only .50. The test is well normed on a representative sample of 12.000 persons between the ages of 16 and 23 years. The ASVAB manual reports a median validity coefficient of .60 with measures of training performance.

Decisions about ASVAB examinees are typically based upon composite scores not subtest scores. For example, a Clerical Composite is derived by combining Word Knowledge, Paragraph Comprehension, Numerical Operations, and Coding Speed. Subjects scoring well on this composite might be assigned to secretarial positions. Since the composite scores are empirically derived, new ones can be developed for placement decisions at any time. Composite scores are continually updated and revised. For example, Ree and Carretta (1994) advocated three composites derived from a factor analysis of more than 11,000 participants in the ASVAB testing.

Table 14.5 The Armed Services Vocational Aptitude Battery (ASVAB) Subtests

Arithmetic Reasoning	30-item test of arithmetic word problems based upon simple calculation
Mathematics Knowledge	25-item test of algebra, geometry, fractions, decimals, and exponents
Paragraph Comprehension	15-item test of reading comprehension in short paragraphs
Word Knowledge	35-item test of vocabulary knowledge and synonyms
Coding Speed	84-item speeded test of substitution of numeric for verbal codes
General Science	25-item test of general knowledge in physical and biological science
Numerical Operations	50-item speeded test of ability to add, subtract, multiply, and divide
Electronics Information	20-item test of electronics, radio, and electrical principles

Mechanical Comprehension	25-item test of mechanical and physical principles
Auto and Shop Information	25-item test of basic knowledge of autos, shop practices, and tool usage

These composites and their constitutent tests were as follows:

1. Speed: Numerical Operations and Coding Speed
2. Verbal/Math: Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, and Mathematics Knowledge
3. Technical Knowledge: General Science, Auto and Shop Information, Mechanical Comprehension, and Electronics Information

The reader will notice that the second factor is identical to the AFQT, mentioned previously.

At one point, the Armed Services relied heavily upon the seven composites in the following list (Murphy, 1984). The first three constitute academic composites, whereas the remaining are occupational composites. The reader will notice that individual subtests may appear in more than one composite:

1. Academic Ability: Word Knowledge, Paragraph Comprehension, and Arithmetic Reasoning

2. Verbal: Word Knowledge, Paragraph Comprehension, and General Science
3. Math: Mathematics Knowledge and Arithmetic Reasoning

4. Mechanical and Crafts: Arithmetic Reasoning, Mechanical Comprehension, Auto and Shop Information, and Electronics Information

5. Business and Clerical: Word Knowledge, Paragraph Comprehension, Mathematics Knowledge, and Coding Speed

6. Electronics and Electrical: Arithmetic Reasoning, Mathematics Knowledge, Electronics Information, and General Science

7. Health, Social, and Technology: Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, and Mechanical Comprehension

The problem with forming composites in this manner is that they are so highly correlated with one another as to be essentially redundant. In fact, the average intercorrelation among these seven composite scores is .86! (Murphy, 1984). Clearly, composites do not always provide differential information about specific aptitudes. Perhaps that is why recent editions of the ASVAB have steered clear of multiple, complex composites. Instead, the emphasis is on simpler composites that are composed of highly related constructs. For example, a Verbal Ability composite is derived from Word Knowledge and Paragraph Comprehension, two highly interrelated subtests. In like manner, a Math Ability composite is obtained from the combination of Arithmetic Reasoning and Mathematics Knowledge.

Some researchers have concluded that the ASVAB does not function as a multiple aptitude test battery, but achieves success in predicting diverse vocational assignments because the composites invariably tap a general factor of intelligence. For example, Dunai and Porter (2001) report favorably on the ASVAB as a predictor of entry-level success of radiography students in Air Force medical training. The ASVAB may be a good test of general intelligence, but it falls short as a multiple aptitude test battery. Another concern is that the test may possess different psychometric structures for men and women. Specifically, the Electronics Information subtest is a good measure of g (the general factor of intelligence) for men, but not women (Ree & Car-retta, 1995). The likely explanation for this is that men are about nine times more likely to enroll in high school classes in electronics and auto shop, and men therefore have the opportunity for their general ability to shape what they learn about electronics information, whereas women do not. Scores on this subtest will therefore function as a measure of achievement (what has already been learned) but not as an index of aptitude (forecasting future results).

Research on a computerized adaptive testing (CAT) version of the ASVAB has been underway since the 1980s. Computerized adaptive testing is discussed in Topic 15A, Computerized Assessment and the Future of Testing. We provide a brief overview here. In CAT, the examinee takes the test while sitting at a computer terminal. The difficulty level of the items presented on the screen is continually readjusted as a function of the examinee's ongoing performance. In general, an examinee who answers a subtest item correctly will receive a harder item, whereas an examinee who fails that item will receive an easier item. The computer uses item response theory as a basis for selecting items. Each examinee receives a unique set of test items tailored to his or her ability level.

In 1990, the CAT-ASVAB began to replace the paper-and-pencil ASVAB. Currently, more than two-thirds of all military applicants are tested with the computerized version. Larson (1994) lists the reasons for adopting the CAT-ASVAB as follows:

1. Shorten overall testing time (adaptive tests require roughly one-half the items of standard tests).
2. Increase test security by eliminating the possibility that test booklets could be stolen.
3. Increase test precision at the upper and lower ability extremes.
4. Provide a means for immediate feedback on test scores, since the computers used for testing can immediately score the tests and output the results.
5. Provide a means for flexible test start times (unlike group-administered paper-and-pencil tests, for which everyone must start and stop at the same time, computer-based testing can be tailored to the examinees' personal schedules).

Reliability and validity studies of the CAT-ASVAB provide strong support for its equivalence to the original test. In general, the computerized version of the instrument measures the same constructs as its paper-and-pencil counterpart—and does so in less time and with greater precision (Moreno & Segall, 1997; Segall, 1997). With the success of this project, the CAT-ASVAB and other tests likely will be expanded to measure new aspects of performance such as response latencies and to display unique items types such as visu-ospatial tests of objects in motion (Larson, 1994). The CAT-ASVAB has the potential to change the future of testing.

PREDICTING COLLEGE PERFORMANCE

As most every college student knows, a major use of aptitude tests is the prediction of academic performance. In most cases, applicants to college must contend with the Scholastic Assessment Tests (SAT) or the American College Test (ACT) assessment program. Institutions may set minimum standards on the SAT or ACT tests for admission, based on the knowledge that low scores foretell college failure. In this section we will explore the technical adequacy and predictive validity of the major college aptitude tests.

The Scholastic Assessment Tests (SAT)

Formerly known as the Scholastic Aptitude Tests, the Scholastic Assessment Tests, or SAT, is the oldest of the college admissions tests, dating back to 1926. The SAT is published by the College Board (formerly the College Entrance Examination Board), a group formed in 1899 to provide a national clearinghouse for admissions testing.

As noted by historian Fuess (1950), the purpose of a nationally based admissions test was "to introduce law and order into an educational anarchy which towards the close of the nineteenth century had become exasperating, indeed almost intolerable, to schoolmasters." Over the years, the test has been extensively revised, continuously updated, and repeatedly renormed. In the early 1990s, the SAT was renamed the Scholastic Assessment Tests to emphasize changes in content and format. The new SAT assesses mastery of high school subject matter to a greater extent than its predecessor, but continues to tap reasoning skills. The SAT represents state of the art for aptitude testing.

The new SAT consists of the SAT-I Reasoning Tests and the SAT-II Subject Tests. The SAT-I Verbal Reasoning Test emphasizes vocabulary in context, reading comprehension, and critical reasoning. The SAT-I Math Reasoning emphasizes the application of mathematical concepts, the interpretation of data, and the actual construction of a response, as opposed to the typical multiple-choice format. A calculator is highly recommended but not required.
The SAT-I Verbal Reasoning and Math Reasoning scores are reported on a scale that ranges from 200 to 800. Characteristic item types for the Verbal portion include the following:

- Analogies: Select a pair of words that best expresses a relationship similar to that expressed in a stimulus pair.

- Sentence Completions: For a sentence with one or two blanks, choose a word or pair of words that best fits the meaning of the sentence as a whole.

- Reading Comprehension: Read a passage and answer multiple-choice questions based on what is stated or implied in the passage.

Characteristic item types for the Math portion include the following:

- Regular Mathematics: Solve basic problems in geometry and algebra.

- Quantitative Comparisons: Choose from two quantities which is greater, or denote that they are equal, or denote that the problem is unsolvable from the information given.

A persistent misconception about SAT scores is that 500 and 100 represent the mean and standard deviation of the most recent sample of SAT test takers (Donlon, 1984). In fact, the numbers 500 and 100 refer to the mean and standard deviation of the anchor group of 10,654 students who took the Verbal portion of the SAT in April 1941. (The Mathematics portion was equated to this verbal portion the next year.) All new scores are equated to the anchor scores by linking each new form of the SAT to one or more previous forms. For example, if a new form is slightly easier than previous forms, the test taker may need a few more correct answers in order to attain an equivalent score. This procedure guarantees that current SAT scores are based on the same measurement scale used at the inception of the anchoring procedure in 1941. A rescaling and repositioning of SAT scores was scheduled for 1996. One purpose of the rescaling was to provide more reliable measurement in upper and lower score ranges by widening the item difficulty level (Johnson, 1994).

From year to year, the average score for SAT test takers may be substantially different from the original average of 500. In fact, SAT scores declined precipitously from 1963 to 1980. By 1980, the Mathematics average had declined from 500 to about 465, while the Verbal average reached a low of 420, nearly a full standard deviation below its starting point. Average scores on both scales have increased only slightly since then. This phenomenon has been the subject of intense scrutiny, and it is beyond the scope of this text to review all the explanations that have been profferred. The following findings are generally accepted:

- The decline was not an artifact of SAT difficulty or scaling; it was a real phenomenon that affected other major testing programs.

- The decline was significant, representing a sizeable shift in test performance; the change represented a "serious deterioration of the learning process in America" (Wirtz, 1977).

- The decline did not lessen the predictive validity of the SAT; the test continued to correlate well with college performance.

- Population shifts such as increases in family size may explain part of the decline; if average family size continues to decrease, SAT scores are predicted to increase (Zajonc, Markus, & Markus, 1979).
- Social changes such as the expansion of television may have contributed to the decline; however, such hypotheses are difficult to prove (Donlon, 1984).

Great care is taken in the construction of new forms of the SAT Verbal and Math tests because unfailing reliability and a high degree of parallelism are essential to the mission of this testing program. The internal consistency reliability of both forms is repeatedly in the range of .91 to .93; with only a few exceptions, test-retest correlations vary between .87 and .89. The standard error of measurement is 30 to 35 points.

The primary evidence for SAT validity is criterion-related, in this case, the ability to predict first-year college grades. Donlon (1984, chap. VIII) reports a wealth of information on this point; we can only summarize trends here. In 685 studies, the combined SAT Verbal and Math scores correlated .42, on average, with college first-year grade point average. Interestingly, high school record (e.g., rank or grade point average) fares better than the SAT in predicting college grades (r = .48). But the combination of SAT and high school record proves even more predictive; these variables correlated .55, on average, with college first-year grade point average. Of course, these findings reflect a substantial restriction of range: low SAT-scoring high school students tend not to attend college. Donlon (1984) estimates that the real correlation without restriction of range (SAT + high school record) would be in the neighborhood of .65.

One issue of great practical concern is the effect of special study and coaching on SAT scores. Does it help to receive special coaching on vocabulary and mathematics or to read the numerous preparation guides available in most any bookstore? Messick and Jungeblut (1981) reviewed the available studies that employed an experimental versus con¬trol group format. They concluded that coaching boosts the combined Verbal and Math scores about 28 to 30 points, not a substantial increase compared to no coaching/preparation. However, for highly motivated students who seek out coaching and receive a rigorous, structured program, coaching effects are much larger, 45 to 110 points on the combined Verbal and Math scores (Johnson, 1994). A related issue pertains to the sizeable proportion of students who take the SAT more than once. Do scores tend to rise with repeated testing? In cases in which the retesting occurs within five to eight months, the average increase is about 12 points each for the Verbal and Mathematics scores (Donlon, 1984). The increase reflects, in part, familiarity with the test, but a factor overlooked by many is the active learning that might take place in the interim.

The American College Test (ACT)

The American College Test (ACT) assessment program is a recent program of testing and reporting designed for college-bound students. In addition to traditional test scores, the ACT assessment program includes a brief 90-item interest inventory (based on Holland's typology) and a student profile section (in which the student may list subjects studied, notable accomplishments, work experience, and community service). We will not discuss these ancillary measures here, except to note that they are useful in generating the Student Profile Report, which is sent to the examinee and the colleges listed on the registration folder.

Initiated in 1959, the ACT is based on the philosophy that direct tests of the skills needed in college courses provide the most efficient basis for predicting college performance. In terms of the number of students who take it, the ACT occupies second place behind the SAT as a college admissions test. The four ACT tests require knowledge of a subject area, but emphasize the use of that knowledge:

- English (75 questions, 45 minutes). The examinee is presented with several prose passages excerpted from published writings. Certain portions of the text are underlined and numbered, and possible revisions for the underlined sections are presented; in addition, "no change" is one choice. The examinee must choose the best option.

- Mathematics (60 questions, 60 minutes). Here the examinee is asked to solve the kinds of mathematics problems likely to be encountered in basic college mathematics courses. The test emphasizes concepts rather than formulas and uses a multiple-choice format.

- Reading (40 questions, 35 minutes). This subtest is designed to assess the examinee's level of reading comprehension; subscores are reported for social studies/sciences and arts/literature reading skills.

- Science Reasoning (40 questions, 35 minutes). This test assesses the ability to read and understand material in the natural sciences. The questions are drawn from data representations, research summaries, and conflicting viewpoints.

In addition to the area scores listed previously, ACT results are also reported as an overall Composite score, which is the average of the four tests.

ACT scores are reported on a standard score 36-point scale. In 2002, the average ACT Composite score of high school graduates was 20.8, with a standard deviation of about 5 points (Maxey, 1994). However, like the SAT, scores on the ACT are not fixed in any given year. ACT scores showed the same decline in the 1960s and 1970s as observed on the SAT.

Critics of the ACT program have pointed to the heavy emphasis upon reading comprehension that saturates all four tests. The average intercorrelation of the tests is typically around .60. These data suggest that a general achievement/ability factor pervades all four tests; results for any one test should not be overinterpreted. Fortunately, college admission officers probably place the greatest emphasis upon the Composite score, which is the average of the four separate tests. The ACT test appears to measure much the same thing as the SAT; the correlation between these two tests approaches .90. It is not surprising, then, that the predictive validity of the ACT Composite score rivals the SAT combined score, with correlations in the vicinity of .40 to .50 with college first-year grade point average. The predictive validity coefficients are virtually identical for advantaged and disadvantaged students, indicating that the ACT tests are not biased.

Kifer (1985) does not question the technical adequacy of the ACT and similar testing programs, but does protest the enormous symbolic power these tests have accrued. The heavy emphasis upon test scores for college admissions is not a technical issue, but a social, moral, and political concern:

Selective admissions means simply that an institution cannot or will not admit each person who completes an application. Choices of who will or will not be admitted should be, first of all, a matter of what the institution believes is desirable and may or may not include the use of prediction equations. It is just as defensible to select on talent broadly construed as it is to use test scores however high. There are talented students in many areas— leaders, organizers, doers, musicians, athletes, science award winners, opera buffs —who may have moderate or low ACT scores but whose presence on a campus would change it.

The reader may wish to review Topic 7B, Test Bias and Other Controversies, for further discussion of this point.

POSTGRADUATE SELECTION TESTS

Graduate and professional programs also rely heavily upon aptitude tests for admission decisions. Of course, many other factors are considered when selecting students for advanced training, but there is no denying the centrality of aptitude test results in the selection decision. For example, Figure 14.5 depicts a fairly typical quantitative weighting system used in evaluating applicants for graduate training in psychology. The reader will notice that an overall score on the Graduate Record Exam (GRE) receives the single highest weighting in the selection process. We review the GRE in the following sections, as well as admission tests used by medical schools and law schools.

Graduate Record Exam (GRE)

The GRE is a multiple-choice and essay test widely used by graduate programs in many fields as one component in the selection of candidates for advanced training. The GRE offers subject examinations in many fields (e.g., Biology, Computer Science, History, Mathematics, Political Science, Psychology), but the heart of the test is the general test designed to measure verbal, quantitative, and analytical writing aptitudes. The verbal section (GRE-V) includes verbal items such as analogies, sentence completion, antonyms, and reading comprehension. The quantitative section (GRE-Q) consists of problems in algebra, geometry, reasoning, and the interpretation of data, graphs, and diagrams. The analytical writing section (GRE-AW) was added in October 2002 as a measure of higher-level critical thinking and analytical writing skills. It consists of two writing tasks: a 45-minute essay in which the applicant takes a position on an issue, and a 30-minute essay in which the applicant analyzes an argument. This new addition to the GRE replaced a multiple-choice test of analytical thinking that is no longer used.

GRE Scores	**0**	**6**	**12**	**18**	**24**	**30**
GRE-V + GRE-Q total:		1000	1100	1200	1300	1400
Undergraduate GPA	**0**	**5**	**10**	**15**	**20**	**25**
		3.0	3.2	3.4	3.6	3.8
Psychology GPA	**0**	**1**	**2**	**3**	**4**	**5**
		3.0	3.2	3.4	3.6	3.9
Background in Statistics/Experimental	**0**	**1**	**2**	**3**	**4**	**5**
Background in Biology/Chemistry	**0**	**1**	**2**	**3**	**4**	**5**
Background in Math/Computer Science	**0**	**1**	**2**	**3**	**4**	**5**
Research Experience	**0**	**1**	**2**	**3**	**4**	**5**
Positive Interpersonal Skills	**0**	**2**	**4**	**6**	**8**	**10**
Ethnic/Linguistic/Cultural Diversity	**0**	**2**	**4**	**6**	**8**	**10**
				Maximum Total:		**100**

Figure 14.5 Representative Weighting Scheme Used by Graduate Program Admission Committees in Psychology

The first two scores (GRE-V and GRE-Q) are reported as standard scores with an approximate mean of 500 and standard deviation of 100. Actually, the mean score may differ from year to year because all test results are anchored to a standard reference group of 2,095 college seniors tested in 1952 on the verbal and quantitative portions of the test. Historically, graduate programs have tended to pay attention to the combination of scores on the first two parts (GRE-V + GRE-Q), where combined scores above 1,000 would be considered above average. Recently, graduate programs have paid more attention to writing skills in their applicants, which explains the addition of the analytical writing section (GRE-AW) to the test.

Scoring of the analytical writing section is based on 6-point holistic ratings provided independently by two trained raters. If the two scores differ by more than one point on the scale, the discrepancy is adjudicated by a third GRE-AW reader. According to the GRE Board (www.gre.org) the GRE-AW test reveals smaller ethnic group differences than found in the multiple-choice sections. For example, the differences between African American and Caucasian examinees and between Hispanic and Caucasian examinees are smaller on the GRE-AW than on the GRE-V or GRE-Q. This suggests that the new test does not unduly penalize ethnic groups traditionally underrepresented in graduate programs.

The reliability of the GRE is strong, with internal consistency reliability coefficients typically around .90 for the three components. The validity of the GRE commonly has been examined in relation to the ability of the test to predict performance in graduate school. Performance has been operationalized mainly as grade point average, although faculty ratings of student aptitude also have been used. For example, based upon a meta-analytic review of 22 studies with a total of 5,186 students, Morrison and Morrison (1995) concluded that GRE-V correlated .28 and GRE-Q correlated .22 with graduate grade point average. Thus, on average, GRE scores accounted for only 6.3 percent of the variance in graduate-level academic performance. In a recent study of 170 graduate students in psychology at Yale University, Sternberg and Williams (1997) also found minimal correlations between GRE scores and graduate grades. When GRE scores were correlated with faculty ratings on five variables (analytical, creative, practical, research, and teaching abilities), the correlations were even lower, for the most part hovering right around zero. The single exception was the GRE analytical thinking score, which correlated modestly with almost all of the faculty ratings. However, this correlation was observed only for men (on the order of $r = .3$), whereas for women it was almost exactly zero in every case! Based upon these and similar studies, the consensus would appear to be that excessive reliance on the GRE for graduate school selection may overlook a talented pool of promising graduate students.

However, other researchers are more supportive in their evaluation of the GRE, noting that the correlation of GRE scores and graduate grades is not a good index of validity because of the restriction of range problem (Kuncel, Campbell, & Ones, 1998). Specifically, applicants with low GRE scores are unlikely to be accepted for graduate training in the first place and thus relatively little information is available with respect to whether low scores predict poor academic performance. Put simply, the correlation of GRE scores with graduate academic performance is based mainly upon persons with middle to high levels of GRE scores, that is, GRE-V + GRE-Q totals of 1,000 and up. As such, the correlation will be attenuated precisely because those with low GREs are not included in the sample. Another problem with validating the GRE against grades in graduate school is the unreliability of the criterion (grades). Based upon the expectation that graduate students will perform at high levels, some professors may give blanket A's such that grades do not reflect real differences in student aptitudes. This would lower the correlation between the predictor (GRE scores) and the criterion (graduate grades). When these factors are accounted for, many researchers find reason to believe the GRE is still a valid tool for graduate school selection (Melchert, 1998; Ruscio, 1998).

Medical College Admission Test (MCAT)

The MCAT is required of applicants to almost all medical schools in the United States. The test is designed to assess achievement of the basic skills and concepts that are prerequisites for successful completion of medical school. There are three multiple-choice sections (Verbal Reasoning, Physical Sciences, and Biological Sciences) and one essay section (Writing Sample). The Verbal Reasoning section is designed to evaluate the ability to understand and apply information and arguments presented in written form. Specifically, the test consists of several passages of about 500 to 600 words each, taken from humanities, social sciences, and natural sciences. Each passage is followed by several questions based on information included in the passage. The Physical Sciences section is designed to evaluate reasoning in general chemistry and physics. The Biological Sciences is designed to evaluate reasoning in biology and organic chemistry. These physical and biological science sections contain 10 to 11 problem sets described in about 250 words each, with several questions following.

The Writing Sample Test consists of two 30-minute essays. This test is designed to assess basic writing skills such as developing a central idea, synthesizing concepts and ideas, writing logically, and following accepted practices of grammar, syntax, and punctuation. The writing sample essays begin with a prompt, which consists of a topic statement (printed in boldface) followed by instructions for interpretation and response. The writing sample prompts resemble the following (www.aamc.org):

Scientists should seek to confirm theories or hypotheses rather than to refute them.
Describe a specific situation in which a scientist might seek to refute a theory or hypothesis rather than to confirm it. Discuss what you think determines when scientists should seek to confirm theories or hypotheses and when they should seek to refute them.

The writing samples are scored on a 6-point scale by independent raters. The basis for the inclusion of writing samples in the MCAT is that physicians are expected to communicate clearly with patients, write lucid and effective medical notes, and contribute persuasively to local and national debates about health care policy.

Each of the MC AT scores (except Writing Samples) is reported on a scale from 1 to 15 (means of about 8.0 and standard deviations of about 2.5). The reliability of the test is lower than that of other aptitude tests used for selection, with internal consistency and split-half coefficients mainly in the low .80s (Gregory, 1994a). MCAT scores are mildly predictive of success in medical school, but once again the restriction of range conundrum (previously discussed in relation to the GRE) is at play. In particular, examinees with low MCAT scores who would presumably confirm the validity of the test by performing poorly in medical school are rarely admitted, which reduces the apparent validity of the test.

Law School Admission Test (LSAT)

The LSAT is a half-day standardized test required of applicants to virtually every law school in the United States. The test is designed to measure skills considered essential for success in law school, including the reading and understanding of complex material, the organization and management of information, and the ability to reason critically and draw correct inferences. The LSAT consists of multiple-choice questions in four areas: reading comprehension, analytical reasoning, and two logical reasoning sections. An additional section is used to pretest new test items and to preequate new test forms, but this section does not contribute to the LSAT score. The score scale for the LSAT extends from a low of 120 to a high of 180. In addition to the objective portions, a 30-minute writing sample is administered at the end of the test. The section is not scored, but copies of the writing sample are sent to all law schools to which the examinee applies.

The LSAT has acceptable reliability (internal consistency coefficients in the .90s) and is regarded as a moderately valid predictor of law school grades. Yet, in one fascinating study, LSAT scores correlated more strongly with state bar test results than with law school grades (Melton, 1985). This speaks well for the validity of the test, insofar as it links LSAT scores with an important, real-world criterion.

Group Tests of Achievement

In this topic, we continue the discussion of group tests by surveying their use within educational systems. Beginning in the elementary grades, school districts use group achievement tests to track the progress of individual students and to gauge the success of educational programs. The ubiquitous practice of group achievement testing within U.S. schools is largely a positive affair because it provides an objective basis for evaluation. However, there is on occasion a dark side as well, insofar as testing can become the tail that wags the dog. The negative impact falls into three general categories. First, teachers may teach to the tests rather than trying to impart genuine knowledge. Second, in a quest to obtain high scores for their school systems, administrators may foster an environment that encourages liberal, nonstandard testing. Worse yet, school personnel may engage in outright fraud such as "correcting" answer sheets. The third consequence is that individual examinees will find ingenious ways of cheating on nationally normed tests. We review a few of these disquieting trends at the end of this topic.

ESSENTIAL CONCEPTS I IN ACHIEVEMENT TESTS

Achievement tests, known as attainment tests in the United Kingdom, are the most widely used of all types of tests. Although precise figures on usage do not exist, virtually every school-aged child in the United States encounters group standardized achievement testing on a yearly or biyearly basis. One estimate is that public schools administer an average of two and one-half tests per student per year (Medina & Neill, 1990). Beyond a doubt, the number of achievement tests administered surpasses all other forms of psychological and educational testing.

Achievement tests are designed to measure the attainment of skills taught within schools or training programs. These tests can be quite narrowly defined such as a test of punctuation skills, or more broadly conceived such as a test of reading comprehension. Even though achievement tests differ in their specificity, they all serve a related function: to measure current skill level in a well-defined domain.

As catalogued in the Mental Measurements Yearbook series (e.g., Plake & Impara, 2001)), literally hundreds of achievement tests have been published. It is not feasible to survey the vast panorama of these instruments. Instead, we review representative achievement tests and focus upon the issues raised by their use. We begin with a primer on essential concepts in achievement testing.

Group and Individual Achievement Tests

A fundamental distinction is drawn between achievement tests and individual achievement Group achievement tests are used mainly in classroom, whereas individual achievement tests employed one on one in clinical or educational tings. Group achievement tests might also called educational achievement tests, since these instruments are commonly administered to entire school systems at the behest of state school superintendents or other administrators. Of course, group tests are given simultaneously to dozens or hundreds of students at the same time, with all the advantages and pitfalls attendant to this approach (see Topic 2B, The Testing Process).

Individual achievement tests play an essential role in the diagnosis of a learning disability (LD). Not only do these tests provide documentation of impaired performance in such crucial academic areas as reading, writing, and calculation, some achievement tests can help identify the particular skill deficits that underlie learning disabilities. Individual achievement tests are used in conjunction with other instruments, especially intelligence tests, as discussed in Topic 10A, School-Based Assessment.

Norm-Referenced and Criterion-Referenced Tests

In addition to the fundamental dichotomy that separates group from individual achievement tests, another important distinction is between norm-referenced and criterion-referenced achievement tests. The reader will recall from Topic 2A (The Nature and Uses of Psychological Tests) that norm-referenced tests allow for interpretation in reference to a large standardization sample. Norm-referenced tests facilitate the reporting of scores as percentile ranks, standard scores, and the like. In contrast, criterion-referenced tests allow for interpretation in reference to the specific content mastered by the individual examinee. For example, a criterion-referenced test might determine that an examinee knows how to spell correctly 94 out of 100 items from a designated list of essential words. Of course, these two approaches are not necessarily incompatible. In fact, most major achievement test batteries provide both norm-referenced and criterion-referenced interpretations.

Ability, Aptitude, and Achievement Tests

The distinction between ability, aptitude, and achievement tests merits brief review in this context.

Ability tests sample a broad assortment of skills in order to estimate general intellectual level. In contrast, aptitude tests usually measure homogeneous segments of ability and are often used to predict future performance. The exceptions here include multiple aptitude test batteries that sample abilities broadly; these instruments are very similar to ability tests. Finally, as noted, achievement tests measure current skill attainment, particularly in relation to school and training programs.

In the real world, the distinction between these three types of tests is often quite fuzzy (Gregory, 1994a). It has been known for some time that the correlation between an achievement test and an ability test may be nearly as high as that between any two ability tests. In many cases, achievement and ability tests tap similar underlying cognitive factors. However, the assumptions that underly these two forms of testing differ widely. Achievement tests are generally designed to measure the effects of relatively standardized educational experiences, whereas aptitude tests typically make fewer assumptions about specific prior learning experiences.

The applications of aptitude and achievement tests also differ widely. Aptitude tests are designed primarily to predict future performance in schools or training programs. For example, a scholastic aptitude test might be used to predict future academic performance in college; a clerical aptitude test might be used to predict future performance in the role of secretary. In contrast, achievement tests are used to gauge a student's current level of attainment in a given subject matter. In other words, aptitude tests are oriented to the future, whereas achievement tests are oriented to the present. The assessment of current skill level with achievement tests can serve several purposes, as outlined in the following section.

The Functions of Achievement Testing

Achievement tests permit a wide range of potential uses. Practical applications of individual and group achievement tests include the following:

- To identify children and adults with specific achievement deficits who might need more detailed assessment for learning disabilities
- To help parents recognize the academic strengths and weaknesses of their children and thereby foster individual remedial efforts at home
- To identify classwide or schoolwide achievement deficiencies as a basis for redirection of instructional efforts
- To appraise the success of educational programs by measuring the subsequent skill attainment of students
- To group students according to similar skill level in specific academic domains
- To identify the level of instruction that is appropriate for individual students

Thus, achievement tests serve institutional goals such as monitoring schoolwide achievement levels, but also play an important role in the assessment of individual learning difficulties. As previously noted, different kinds of achievement tests are used to pursue these two fundamental applications (institutional and individual). Institutional goals are best served by group achievement test batteries, whereas individual assessment is commonly pursued with individual achievement tests (even though group tests may play a role here, too). In this topic we focus on group educational achievement tests.

EDUCATIONAL ACHIEVEMENT TESTS

Virtually every school system in the nation uses at least one educational achievement test, so it is not surprising that test publishers have responded to the widespread need by developing a panoply of excellent instruments. In the following section, we describe several of the most widely used group standardized achievement tests. The tests to be described share several characteristics in common.

First, these instruments are multilevel batteries that contain comparable subtests for students in the different grades of primary and/or secondary school. Some of the batteries span kindergarten (K) through grade 12, whereas others are designed for elementary grades (K through 8) or secondary grades (9 through 12) only. In a multilevel battery, test booklets contain overlapping sections, and students at different grade levels enter and exit the test materials at grade-appropriate positions.

A second feature common to many educational test batteries is concurrent norming with an ability test. For example, the achievement battery known as the Sequential Tests of Educational Progress (STEP-III) is concurrently normed with the ability battery known as the School and College Ability Test (SCAT-UI). Tests that are concurrently normed share the same standardization sample. As a result, average performance on one test can be directly equated with average performance on the other. Concurrent norming is helpful because it allows parents, teachers, and counselors to make precise, direct, and meaningful comparisons between achievement and ability. After all, the implications of an achievement score are moderated by knowledge of the student's ability. A student with high ability scores but low achievement scores might be a good candidate for educational intervention, including a more detailed assessment for learning disability (as discussed in Topic 10A, School-Based Assessment). In contrast, a student with low ability scores and low achievement scores might be working at full potential; specialized interventions may not be warranted.

The third commonality in group achievement tests is that they measure similar educational skills. Educational achievement tests tend to emphasize these skill areas:

1. Reading, including comprehension and vocabulary

2. Written language, including spelling, punctuation, and capitalization
3. Mathematics, including computation and application

In addition, tests at the elementary grade levels often assess listening skills, including oral comprehension. Some test batteries also assess knowledge of basic concepts in science, social studies, and humanities.

Finally, the educational achievement tests discussed here possess generally excellent psychometric characteristics. Test contents are relevant and appropriate, that is, the instruments show good content validity; subscales possess excellent internal and alternate-forms reliability; standardization samples are invariably large and representative; and overt gender and race bias are nonexistent. The psychometric quality of the widely used educational achievement tests is typically respectable, if not extraordinary.

We survey several widely used tests of educational achievement subsequently. The reader will discover that a detailed analysis of psychometric properties—reliability, validity, norming and the like—is encountered only for the first instrument reviewed, the Iowa Tests of Basic Skills. In general, the psychometric quality of the other tests is equally laudable, so for these test batteries we focus upon functions, applications, special features, and an occasional shortcoming or two. Readers who desire more information on these instruments should consult reviews in the Mental Measurements Yearbook (Conoley & Impara, 1995:Cono-ley & Kramer, 1989, 1992; Impara & Plake, 1998: Plake & Impara, 2001; Mitchell, 1985).

Iowa Tests of Basic Skills (ITBS)

First published in 1935, the Iowa Tests of Basic Skills (ITBS) were most recently revised and re-standardized in 1992. The ITBS is a multilevel battery of achievement tests that covers grades K through 8. A companion test, the Tests of Achievement and Proficiency (TAP), covers grades 9 through 12. In order to expedite direct and accurate comparisons of achievement and ability, the ITBS and the TAP were both concurrently normed with the Cognitive Abilities Test (CogAT), a respected group test of general intellectual ability.

The ITBS is available in several levels that correspond roughly with the ages of the potential examinees: levels 5 and 6 (grades K-l), levels 7-8 (grades 2-3), and levels 9-14 (grades 3-8). The basic subtests for the older levels measure vocabulary, reading, language, mathematics, social studies, science, and sources of information (e.g.. uses of maps and diagrams).

From the first edition onward, the ITBS has been guided by a pragmatic philosophy of educational measurement. The manual states the purpose of testing as follows:

> *The purpose of measurement is to provide information which can be used in improving instruction. Measurement has value to the extent that it results in better decisions which directly affect pupils.*

To this end the ITBS incorporates a criterion-referenced skills analysis to supplement the usual array of norm-referenced scores. For example, one feature available from the publisher's scoring service is item-level information. This information indicates topic areas, items sampling the topic, and correct or wrong response for each item. Teachers therefore have access to a wealth of diagnostic-instructional information for each student. Whether this information translates to better instruction—as the test authors desire—is very difficult to quantify. As Linn (1989) notes, "We must rely mostly on logic, anecdotes, and opinions when it comes to answering such questions."

The technical properties of the ITBS are beyond reproach. Internal consistency and equivalent-form reliability coefficients are mostly in the mid-.80s to low .90s. Stability coefficients for a one-year interval are almost all in the .70 to .90 range. The test is free from overt racial and gender bias, as determined by content evaluation and item bias studies. The year 2000 norms for the test were empirically developed from large, representative national probability samples.

Standardization of a previous form in 1988 revealed an intriguing trend in comparison to results for versions that were standardized several years earlier. The 1988 sample demonstrated higher achievement levels, on the order of 1 to 3 months of grade equivalent. This pattern of slowly rising test performance emphasizes the need for annual or biannual restandardization of achievement test batteries. What has happened in the absence of timely restandardization of major achievement tests is that all 50 states can report honestly that they exceed the national average on group standardized tests (Cannell. 1988).

Item content of the ITBS is judged relevant by curriculum experts and reviewers, which speaks to the content validity of the test (Lane, 1992: Linn, 1989; Raju. 1992; Willson, 1989). Although the predictive validity of the latest ITBS has not been studied extensively, evidence from prior editions is very encouraging. For example, ITBS scores correlate moderately with high school grades (rs around .60). The ITBS is not a perfect instrument, but it represents the best that modern test development methods can produce.

Metropolitan Achievement Test (MAT)

The Metropolitan Achievement Test dates back to 1930 when the test was designed to meet the curriculum assessment needs of New York City. The stated purpose of the MAT is "to measure the achievement of students in the major skill and content areas of the school curriculum." The MAT is concurrently normed with the Otis-Lennon School Ability Test (OLSAT).

Now in its eighth edition, the MAT is a multilevel battery designed for grades K through 12 and was most recently normed in 2000. The areas tested by the MAT include the traditional school-related skills:
Reading
Mathematics
Language
Writing
Science

Social Studies

An attractive feature of the MAT is that student reading scores are reported as Lexile measures, a new and practical indicator of reading level. Lexile measures are likely to become a standard feature in most group achievement tests in the years ahead, so it is worth a brief detour to explain their nature and significance.

Lexile Measures

The Lexile approach is a major new improvement in the assessment of reading skill. It was developed over a span of more than twelve years using millions of dollars in grant funds from the National Institute of Child Health and Human Development (NICHD) (www.lexile.com). The Lexile approach is based upon two simple, commonsense assumptions, namely (1) reading materials can be placed on a continuum as to difficulty level (comprehensibility), and (2) readers can be ordered on a continuum as I reading ability. The Lexile framework provides common metric for matching readers and text which, in turn, permits parents and educators choose appropriate reading materials for children.

The Lexile scale is a true interval scale. Lexile measure for a reading selection is a specific number indicating the reading demand of the text based on the semantic difficulty (vocabulary) syntactic complexity (sentence length), measures for reading selections typically rang from 200L to 1700L (Lexiles). The Lexile score for a student, obtained from the Reading Comprehension test of the MAT or other achievement tests, a precise index of the student's reading ability, calibrated on the same scale as the Lexile measure 1 text. The value of the Lexile approach is that student comprehension can be predicted as a function of the disparity between the demands of the text and the student's ability. For example, when readers are well targeted (the difference between text and reader is close to 0 Lexiles), research indicates that reader comprehension will be about 75 percent. When the text difficulty exceeds the reader's ability by 250L, comprehension drops to about: percent. When the skill of the reader exceeds demands of the text by 250L, comprehension about 90 percent (www.lexile.com).

The Lexile approach has a number of potential benefits and applications for teachers and parent Teachers can look up Lexile measures for specific books (the Lexile corporation has evaluated over 30,000 titles to date) as a way of building a library of titles at varying levels. Also, they can produce individualized reading lists suitable for each student. Likewise, parents can select well-matched books to read to their children. Stenner (2001) captures the allure of the Lexile approach as follows:

> *One of the great strengths of the Lexile Framework is the way it encourages thought about what forecasted comprehension rate would be optimal for different instructional contexts. Harry Potter and the Goblet of Fire is a 910L text. Readers at 400L to 500L can nonetheless enjoy listening to this story read aloud. A 700L reader could read the text in a one-on-one tutoring context. A 900L reader will disappear for an hour or two, fully capable of self-engaging with the text, and a 1600L adult reader can become so engrossed that a two-hour plane ride flies by.*

The Lexile approach is not a panacea, but it is a major improvement in the assessment of reading skill.

The Iowa Tests of Educational Development (ITED)

The widely used Iowa Tests of Educational Development were first released in 1942, then revised and restandardized every few years. The purpose of the ITED is: "To assess intellectual skills that are important in adult life and provide the basis for continued learning." Unlike many other achievement tests which emphasize skills linked to specific curricular goals, the intention of the ITED is to measure the fundamental goals or generalized skills of education that are independent of the curriculum. For this reason, the ITED items emphasize higher-order thinking skills. Rather than testing isolated bits of knowledge, questions on the ITED feature problems which require the synthesis of knowledge or a multiple-step solution (Figure 14.6).

The ITED is designed for high school students in grades 9 through 12. The test yields nine basic scores plus a composite:

Vocabulary
Reading Comprehension
Language: Revising Written Measures
Spelling
Mathematics: Concepts and Problem Solving
Computation
Analysis of Social Studies Materials
Analysis of Science Materials
Sources of Information
Total Battery Score

Social Studies

Advertisement

Four out of five doctors surveyed favored
BALM SOAP
Tests show that Balm Soap clears up complexion problems
faster than any other product!

On the basis of this advertisement, which of the following conclusions, if any, is valid?
A. It has been scientifically demonstrated that the quickest way to get rid of any complexion problem is to use Balm Soap.
B. Of the five leading brands of complexion soaps, only one is better than Balm Soap from a medical point of view.
C. Of all the doctors who recommended skin care products, four out of five recommended Balm Soap.
D. None of these conclusions is valid.

Natural Sciences

Soon after being bitten by a mosquito, a person became ill with yellow fever. Which conclusion, if any, is justified solely from these observations?
A. There is insufficient evidence to draw any of the conclusions that follow.
B. Mosquitoes are the direct cause of yellow fever.
C. The mosquito introduced a microorganism into the person's bloodstream.
D. The mosquito carried an organism that caused yellow fever.

Figure 14.6 Representative Items from the Iowa Tests of Educational Development

The core battery consists of the first six tests listed. The ITED is anchored to previous editions so that regardless of form, level, or edition, a given score represents the same level of accomplishment. The test was renormed in 2000 with a large national sample of high school students.

The Tests of Achievement and Proficiency (TAP)

The Tests of Achievement and Proficiency (TAP) are designed to provide a comprehensive appraisal of student progress toward traditional academic goals in grades 9 through 12. The TAP is the second component in the Riverside Basic Skills Assessment Program; the first component is the Iowa Tests of Basic Skills (ITBS), used in grades K through 8 (previously discussed). Like the ITBS, the TAP is concurrently normed with the CogAT, an ability test that measures verbal, quantitative, and nonverbal reasoning abilities. The subtests from the TAP measure achievement in reading comprehension, mathematics, written expression, using sources of information, social studies, and science. A total or composite score is also provided.

The TAP also yields an applied proficiency score that assesses the examinee's capacity to handle real-life situations. This score reflects student competence in applying mathematics and communication skills to solving problems of daily living. The items emphasize communication of ideas in writing, mathematical solution of problems, use of reference materials, and the interpretation of tabular and graphic material. The TAP was conormed with the ITED and the CogAT and restandardized in 1996.

Tests of General Educational Development (GED)

Another widely used achievement test battery is the Tests of General Educational Development (GED), developed by the American Council on Education and administered nationwide for high school equivalency certification (www.acenet.edu). The GED consists of multiple-choice examinations in five educational areas:

Language Arts—Writing
Language Arts—Reading
Mathematics
Science
Social Studies

The Language Arts—Writing section also contains an essay question that examinees must answer in writing. The essay question is scored independently by two trained readers according to a 6-point holistic scoring method. The readers make a judgment about the essay based upon its overall effectiveness in comparison to the effectiveness of other essays.

The GED comes in numerous alternate forms. Typically, internal consistency reliabilities for the subscales are above .90. However, the interrater reliability of scoring on the writing samples is more modest, typically between .6 and .7. These findings indicate that a liberal criterion for passing this subtest is appropriate so as to reduce decision errors. Regarding validity, the GED correlates very strongly (r = .77) with the graduation reading test used in New York (Whitney, Malizio, & Patience, 1985). Furthermore, the standards for passing the GED are more stringent than those employed by most high schools: Currently, individuals who receive a passing score for a GED credential out form at least 40 percent of graduating high school seniors (www.acenet.edu).

The GED emphasizes broad concepts rather than specific facts and details. In general, the; pose of the GED is to allow adults who did graduate from high school to prove that they obtained an equivalent level of knowledge from life experiences or independent study. Employers regard the GED as equivalent (if not superior) to earning a high school diploma. Successful performance on the GED enables individuals to apply to colleges, seek jobs, and request promotions that require a high school diploma as a prerequisite Rogers (1992) and Trevisan (1992) provide unusually ally thorough reviews of the GED.

Additional Group Standardized Achievement Tests

In addition to the previously listed batteries, a other widely used group standardized achiever tests deserve brief mention. Because these strongly resemble the instruments discussed previously, we provide only the barest listing here. The Sequential Tests of Educational Progress (STEP-III) are organized into two batteries, one used for grades K through 3, the other used for grades 3 through 12. The basic STEP-III battery assesses the following educational skills: reading, writing skills, vocabulary, mathematics computation, and mathematics concepts. Additional tests measure attainment in social studies, science, study skills, and oral comprehension (listening). The STEP-III is a companion test to the School and College Ability Tests (SCAT-III).

The widely used Stanford Achievement Series is one of the oldest and most prestigious testing programs in the United States. The series consists of three related test batteries covering grades K through 13: the Stanford Early School Achievement Test (SESAT) for kindergarteners and first graders; the Stanford Achievement Test (SAchT) for grades 1 through 9; and the Stanford Test of Academic Skills (TASK) for grades 8 through 13 (grade 13 refers to the first year of college). Reviewers are cautious about the SESAT because the value of the test is predicated solely upon content validity. Little is known about test-retest reliability, criterion-related validity, and construct validity (Ackerman. 1992; Carpenter, 1992). The SAchT is lauded because of its excellent norm-referenced coverage of a representative and balanced national consensus curriculum (Brown, 1992; Stoker, 1992). The TASK has excellent psychometric characteristics, but in attempting to span high school and college achievement, this test undertakes a difficult assignment. After all, there is modest agreement about the curricular intentions of grade school and high school, but what are the educational goals of the first year in college?

SPECIAL-PURPOSE ACHIEVEMENT TESTS

Achievement tests can be used for many important applied purposes, including the appraisal of knowledge in advanced fields and the evaluation of professional competency. In this final section we will examine two special-purpose achievement tests.

The College-Level Examination Program (CLEP) is a widely used program by which students can demonstrate college-level achievement and receive advance credit or exemption from certain college courses. The National Teacher Examination (NTE) is a controversial test required by many states for teacher certification.

College-Level Examination Program (CLEP)

CLEP is one of two national testing programs through which students can receive college credit by examination without enrolling in the courses. The other program is the ACt Proficiency Examination Program, which we do not discuss here. CLEP is administered by the College Board with financial support from the Carnegie Corporation of New York. The original purpose of the program was to support nontraditional students such as returning veterans and older adults who had obtained valuable learning experiences outside of the classroom. However, it is mainly ambitious high school students enrolled in advanced classes who now register to take the CLEP examinations. Some students begin college with nearly a full semester of course credits obtained through CLEP and similar programs.

CLEP examinations cover material taught in basic first- and second-year courses and colleges usually grant the same amount of credit as would be earned in the corresponding courses. Except for English Composition with Essay, each exam is 90 minutes long and composed primarily of multiple-choice questions; a few exams have a fill-in format. For the English Composition with Essay test, students also write an essay responding to a specific question. Each essay is read by two or more faculty consultants, and this grade is combined with the multiple choice score and reported as a scaled score. Areas tested include American literature, English literature, foreign languages (French, German. Spanish), American government, United States history, principles of macroeconomics, introductory psychology, college algebra, biology, chemistry, and principles of accounting.

Scores on the CLEP tests are reported on a scale from 20 to 80, with an average of 50 and a standard deviation of 10. The reference groups for these scores consisted of volunteer students completing courses in each of the specified areas. These students were recruited from a nonrandom but presumably representative selection of U.S. colleges and universities. The CLEP scores are, in general, highly reliable, with split-half coefficients mainly in the .90s. The validity of the Subject Examinations has been evaluated by means of correlating CLEP scores with final grades in the relevant courses. Most of these correlation coefficients are in the .40s and .50s, which supports the concurrent validity of these tests.

The CLEP program has received high marks from reviewers, but there is a potential negative side as well. In particular, some students might "test out" of college courses that would have proved enriching, inspiring, even life-changing. For example, it is possible to have factual knowledge about art, music, or drama and therefore pass a CLEP test in one or more of these areas. However, in the ambitious quest to finish college quickly, students could overlook important experiences for personal growth.

National Teacher Examination (NTE)

The National Teacher Examination is actually a series of tests published by the Educational Testing Service and known more formally as the Praxis Series. Of the 43 states that require testing as part of the licensure process, 35 use the Praxis Series, which explains why the test is known informally as the National Teacher Examination, or NTE. The Praxis Series is nationally administered and continually updated and improved. The three categories of assessment correspond to major milestones in teacher development:

- Praxis I (Academic Skills Assessments): Entering a teacher training program
- Praxis II (Subject Assessments): Graduating from college and entering the profession
- Praxis III (Classroom Performance Assessments): The first year of teaching

The initial test, Praxis I, is taken early in the student's college career to evaluate reading, writing and math skills essential for the success of any teacher. A passing score on this multiple-choice test is required before the student can continue his or her major in education. These tests can be taken in the traditional paper-and-pencil format or as a computer-based test that is tailored to each candidate's ongoing performance. One advantage to the computer-based testing is year-round availability, whereas the traditional version is given only six times a year. Praxis II assesses knowledge of the subjects a candidate will teach, as well as how much he or she knows about teaching the subject. More than 120 content tests (all multiple choice) are available. Praxis III is an in-class evaluation by trained local assessors who use structured criteria that have been nationally validated.

The reliability of the Praxis I and Praxis II tests is beyond reproach. Similarly, the content validity of these tests is outstanding because they were carefully constructed and refined with the help of many experts and test consumers. What is less clear is the predictive validity of the Praxis Series insofar as little information exists to show that good scores or Praxis evaluations predict good teaching and vice versa. Of course, part of the difficulty here is finding a suitable definition and measure of "good teaching." The National Teaching Examination probably serves a useful purpose by requiring that prospective teachers possess minimum levels of knowledge in their disciplines, but the test also raises difficult questions with regard to how our society identifies promising teachers. Is factual knowledge enough? Should we not also insist that our teachers possess enthusiasm for their material and the capacity to inspire children? These are features not easily captured by objective tests.

CHEATING: THE DARK SIDE OF ACHIEVEMENT TESTING

The prevailing view in the general public is that cheating rarely or never occurs in nationally administered testing programs. We tend to think that the risks are too high and the opportunities too limited for cheaters to prevail. Therefore, we rest assured that test fraud must be a rare event. Unfortunately, this view is probably naive. After all, a growing number of people must pass a test to gain college entry, get a job, or obtain a promotion. Furthermore, school officials increasingly are evaluated on the basis of average test scores in their district. Precisely because the stakes are so high, unscrupulous individuals will try to beat the system.

Consider the case of superior test scores at an acclaimed elementary school in Connecticut (Associated Press, May 4, 1996, and March 15, 1997). The stellar reputation of the school was based upon high exam scores on the Iowa Tests of Basic Skills given to first, third, and fifth graders. The school had won blue ribbons from the U.S. Education Department and was featured as one of the nation's best elementary schools in a prominent magazine. However, in a fluke discovery, school district personnel noted a high number of erasures on the tests from this school and notified the test publisher. On close inspection, the publishers found an exceedingly high number of erasures—9 percent—which was three to five times higher than two nearby schools. Even more suspicious was the fact that 89 percent of the erasures were changed from the wrong answer to the correct answer. Based upon retesting under close supervision, the test publisher found "clearly and conclusively" that tampering occurred. The principal resigned amid allegations that he was responsible for the tampering.

Widespread cheating in public school systems is sporadically reported in many large cities across the United States. In most cases, the cheating is motivated by the desire of teachers and principals to further their own careers by creating the illusion of educational excellence. For example, in 1999, dozens of teachers and two principals in the New York City public school system were charged with helping students cheat on the standardized reading and math tests used to rank schools and determine whether students move on to the next grade (New York Times, December 12, 1999). The cheating scheme was described as "one of the largest in the recent history of American public schools." In 2000, an entire eighth-grade class in a Chicago elementary school was required to retake the Iowa Tests of Basic Skills because a school administrator allegedly filled in incomplete tests and changed incorrect answers to correct ones (Chicago Tribune, June 2, 2000). Officials were tipped off to the fraud because the test scores were simply too good to be true—the average score for the class was two years above their standing. In 2002, Chicago was back in the news again, when sophisticated software detected skills-test cheating at seven schools (Chicago Tribune, October 2, 2002). In this case, the school chief sought to fire six teachers and an aide, remarking, "We need to stand for something, to teach values to our students." Of course, we only read about the cases of cheating that are detected. The number of undetected cases is simply unknown, although probably larger than the public would like to believe.

An especially flagrant instance of cheating on national tests was uncovered in Louisiana in 1997. This case involved wholesale circulation of the Educational Testing Service (ETS) exam administered to teachers who want to be school principals. As reported in the New York Times (September 28, 1997), copies of the 145-item test, along with correct answers, had circulated among teachers throughout southern Louisiana, most likely for several years. In a state ranked at or near the bottom on nearly every educational index, it appears that many potentially unqualified persons cheated their way into running the schools. ETS handled this case quietly by asking more than 200 teachers to retake the test so as to "confirm" their initial scores. Unfortunately, the Louisiana case was not an isolated instance. The New York Times article includes this disquieting conclusion:

> *In numerous instances across the country, E.T.S. has confronted case after case of cheating but withheld information from the public and failed to take aggressive steps in time to insure the integrity of its tests, according to internal documents and interviews with current and former officials there.*

Among the examples cited, ETS allegedly failed to monitor its handling of the federal government's test for immigrants who want to become citizens, with the likely result that test supervisors accepted bribes. English-proficiency tests for foreign students also were vulnerable to cheating. In 1994, ETS canceled the scores of 30,000 students from China after discovering a ring that was selling the examinations abroad. In another case, federal prosecutors uncovered a nationwide cheating ring involving hundreds of Students who paid thousands of dollars each for answers to the GRE and similar exams. This scheme involved a well-known time-zone scam in which experienced test takers took the exams in New York City and then relayed the answers to paying customers taking the tests in the later time zones. Cizek (1999) catalogues literally dozens of ingenious ways that students have developed for cheating on tests: writing information on the floor, in tissues, on the back of a bottled water label; using an ultraviolet pen to write information on "blank" paper; and using a video transmitter (e.g., hidden in an eyeglass case) to send pictures of the test to an outside accomplice who then coaches the student by means of an audio receiver (e.g., hidden in the ear).

Dishonest and inappropriate practices by school officials are implicated in the recent inflation of scores on nationally normed group tests of achievement. By definition, for a norm-referenced test, 50 percent of the examinees should score above the 50th percentile, 50 percent below. If the same test is used in a large sample of typical and representative school systems, average scores for the school systems should be split evenly—about half above the nationally normed 50th percentile, half below.
According to a recent survey reported in the news media (Foster, 1990) virtually all states of the union claim that average achievement scores for their school systems exceed the 50th percentile. The resulting overly optimistic picture of student achievement is labeled the Lake Wobegon Effect, in reference to humorist Garrison Keillor's mythical Minnesota town where "all the children are above average."

How does inflation of achievement test scores arise? According to Cannell (1988), the major cause is educational administrators who are desperate to demonstrate the excellence of their school systems. Precisely because our society attaches so much importance to achievement test results, some educators apparently help students cheat on standardized tests. The alleged cheating includes the following:

- Teachers and principals coach students on test answers.
- Examiners give more than the allotted time to take tests.
- Administrators alter answer sheets.
- Teachers teach directly to the specific test items.
- Teachers make copies of the tests to give to their students.

Cannell notes that over 300 teachers and school administrators answered his trade journal advertisement, admitting that they or colleagues had tampered with tests or helped students improperly. These improprieties constitute a quiet crisis that continues unabated. Another consequence of Lake Wobegon Effect is that test publishers and federal reviewers will likely increase their efforts to monitor test security. In sum the importance that our society attaches to achievement test scores has caused a number of unappealing side effects that undermine the very foundations of nationally normed group-testing programs.

Moore (1994) reports on a special case in educational testing, namely, the districtwide consequences of court-ordered achievement testing. He surveyed 79 teachers from third- through fifth-grade level in a midwestern town in which the court required the use of a standardized test to determine the effectiveness of a desegregation effort. The test in question, the Iowa Tests of Basic Skills (ITBS) is a well-respected group achievement test that requires strict adherence to instructions and time limits for obtaining valid results (discussed earlier), the teachers found little value in the testing program, complaining that its benefits did not offset the time and costs involved. As a consequence of their devaluing the effort, nonstandard testing was practically the rule rather than the exception. The teachers engaged in several nonstandard practices, most of which tended to inflate the test scores. Inappropriate testing practices included praising students who answered a question correctly during the test (67 percent), using last year's test questions for practice (44 percent), recoding a student's answer sheet because he or she just "miscoded" the answer (26 percent), giving students as much time as they needed (24 percent), giving students items that were directly off the test (24 percent), and giving hints or clues during the test (23 percent). In general, Moore (1994) notes that teachers modified their instructional efforts and curriculum in anticipation of having their students take the test. More than 90 percent of the teachers added test-related lessons to the curriculum, and more than 70 percent eliminated topics so that they could spend more time on test-related skills. Whether these are desirable changes is surely open to debate. Moore (1994) concludes:

> *Standardized testing has held a central role in education for many years. What studies of testing program impact have most recently demonstrated is the growing reliance on test scores for decision making and the increasing potential for misuse of test scores. Educational and political policymakers need to address the important link between instruction and testing and ensure mat teachers are integrated into, not isolated from, the intent of testing, (p. 365)*

In sum, what this study demonstrates is that mandated educational testing can have the unanticipated consequence of polluting the validity of a worthy test—especially when crucial stakeholders have no voice in the process.

We cannot survey here all the unintended side effects of educational achievement tests, because the possibilities are nearly endless. For example, what about the warping effects of achievement-testing programs upon school curricula? As we have seen in Moore's (1994) study, teachers do modify their classroom practices with the intention of helping students score well on the tests. However, in teaching to the tests, educators may emphasize bits and pieces of factual knowledge rather than imparting a general ability to think clearly and solve problems. In conclusion, it appears that an excessive emphasis upon nationally normed achievement tests for selection and evaluation promotes inappropriate behavior, including outright fraud and cheating on the part of students and school officials. Just how widespread is the problem? Although we live with the optimistic assumption that fraud in nationally normed testing programs is rare, the disturbing truth is that we really don't know how often.

Lesson 15

TESTING IN CLINICAL AND COUNSELING SETTINGS

Uses of Tests in Clinical Settings

The professionals who identify themselves as clinicians work in such settings as hospitals, community clinics, mental health centers, counseling centers and private practice. Not all these professionals use psychological tests—after all, "a radiologist who treats patients is a clinician. The clinicians most likely to make use of psychological tests are psychologists, psychiatrists, and clinical or psychiatric social workers. The tests are used in the process of diagnosis in the planning of treatment, and in evaluation of the course of treatment.

Clinical Psychology versus Other Fields

The type of clinician best prepared to select, administer and interpret psychological tests is a clinical psychologist. A clinical psychologist typically has completed a master's or doctoral degree and has been trained extensively in psychological theory and practice and in the design and use of psychological tests.

Students are often confused about the difference between clinical psychologists and psychiatrists. A chapter like this is a good place to emphasize one basic difference between them: A psychiatrist is a physician who has completed the same medical school program as all other physicians. It is the subsequent residency in psychiatry that distinguishes this physician from a surgeon or an ophthamologist. However, that residency is quite different from graduate training in clinical psychology. Of necessity, psychiatric residencies emphasize medical procedures for the diagnosis and treatment of psychological problems. Significantly less time is devoted to psychological testing. Although clinical psychology has a historical link to psychiatry, the training and expertise of professionals in these two disciplines can differ radically.

Likewise clinical or psychiatric social workers receive less extensive training in psychological theory and testing. In fact, much of the testing that clinical psychologists do is with clients referred for assessment by psychiatrists and social workers (Lubin et al., 1986b). In a survey of clinical psychologists working in mental health centers, up to two-thirds of their assessments were requested either by psychiatrists or by social workers (Lubin et al., 1986a). Because psychological testing is primarily the provence of psychologists, our discussion of testing in clinical settings will focus on the use of tests by clinical psychologists. When we use the word clinician, we are referring to clinical psychologists.

Testing by Clinical Psychologists

We often think of clinical psychologists as therapists. However, without diagnostic tools, it would be difficult to develop a reasonable treatment plan.

According to surveys of members of the American Psychological Association's Division 12 (Clinical Psychology), clinicians report spending about one-third of their lime administering and evaluating the results of psychological tests (e.g., Wade & Baker, 1977).

Table 15.1 Tests Used Most Frequently by Clinical Psychologists

Rank	Test
1	Wechsler Adult Intelligence Scale
2	Minnesota Multiphasic Personality Inventory
3	Bender Visual Motor Gestalt Test
4	Rorschach Inkblot Test
5	Thematic Apperception Test
6	Wechsler Intelligence Scale for Children—Revised
7.5	Peabody Picture Vocabulary Test
7.5	Sentence Completion Tests (all kinds)
9	House-Tree-Person Test
10	Draw-a-Person Test
11	Wechsler Memory Scale
12	Rotter Sentence Completion Test
13	Memory for Design's
14	Vineland Social Maturity Scale
15	Stanford-Binet Intelligence Scale
16	Strong Vocational Interest Blank—Men
17.5	Bender Visual Retention Test
17.5	Edward Personal Preference Schedule
19	Strong Vocational Interest Blank—Women
20.5	Children's Apperception Test
20.5	Progressive Matrices
22	Kuder Preference Record
23	Porteus Mazes
24	Full Range Picture Vocabulary Test
25	Differential Aptitude Tests
26	Gray Oral Reading Test
27	Wechsler-Bellevue Intelligence Scale
28	Cattell Infant Intelligence Scale
29	Goldstein-Scheerer Tests of Abstract and Concrete Thinking.
30	Blacky Pictures

Table 15.1 lists the 30 tests used most frequently by Division 12 members (Lubin, Larsen, & Matarazzo, 1984). The top five tests include an intelligence test (the Wechsler), three personality tests (two of which are projective tests), and a test of visual-motor integration that is useful for identifying brain damage (the Bender). Table 15.2 categorizes the more familiar tests from this list in terms of purpose and design.

There is some variation in the rankings of tests according to the setting in which-a clinician works. For example, clinical psychologists in veterans administration hospitals use projective tests less frequently .than clinicians in private practice or mental health centers (Lubin et al., 1986a, 1986b). However, the top three tests across all three settings are the MMPI, the Wechsler, and the Bender.

The pattern makes a lot of sense, given what clinical psychologists do. Clinical psychologists typically work with individuals who are, experiencing psychological and/or neuropsychological (brain behavior) problems. Psychological tests are a central component to the process of identifying the nature and extent of these problems and planning an effective treatment program.

You may also have heard of clinical psychologists conducting Mental Status exams for the courts to determine if defendants (1) understand why they fare involved in a court proceeding and (2) can participate effectively in then-own defense.(The administration of psychological tests, including intelligence and personality tests is a key component of mental status evaluations

Table 15.2 Categorizing the Design of Tests Frequently Used in Clinical Settings

Purpose/Format	Examples
Personality Objective Norm referenced Normative	Minnesota Multiphasic Personality Inventory Sixteen Personality Factor Questionnaire California Psychological Inventory
Personality Objective Norm referenced Ipsative	Edwards Personal Preference Schedule
Personality Projective Norm referenced Normative	Rorschach Inkblot Test Thematic Apperception Test House-Tree-Person Test Draw-a-Person Test Sentence-completion tests
Brain function Free response Norm referenced Normative	Bender Visual-Motor Gestalt Wechsler Memory Scale

Other Components of Clinical Assessment

Testing is not the only data-gathering process used by clinical psychologists. In fact, the decision to use tests and the selection of tests to administer usually follow some initial appraisal of an individual's functioning, In addition to administering tests, clinicians assess client characteristics through observations and interviews. In other words, there are other more subjective elements to clinical assessment, and decisions are made on the basis of both objective and subjective data. The fact that clinical judgments are based on both types of data underscores the importance of ensuring that clinicians be well trained and experienced.

The stereotype of the clinical psychologist watching and analyzing you, like all stereotypes, contains an element of truth. .Clinicians tend to observe behavior during all client contacts, including interviews and the administration of tests. Clinicians may use a variety of interview techniques, each with its own particular purpose. Three common types are the life history interview, the diagnostic interview, and the stress interview (e.g. Aiken, 1991).

The goal of a life history interview is to obtain background information about the client, the client's family, and various events in the client's life (e.g., education and employment). These interviews are fairly structured, covering a specific set of topics. The goal of a diagnostic or clinical interview is to obtain information about the client's problems, the effects of these problems on the client's life, and the events that might be influencing these problems (e.g., interpersonal relationships or coping strategies). These interviews are less structured, with successive questions determined by the client's previous responses.

In a stress interview, the clinician seeks information about the impact of stress on client mood, thought, and behavior. The clinician purposely confronts or challenges the client by asking more sensitive questions and or questioning the client's answers. The success of stress interviews depends greatly on the clinician's interviewing skills.

Although assessment through testing, observation, and interview is an important feature of clinical settings, there is a lot of variability in assessment activities across different clinical settings. Psychologists at mental health centers and in private practice typically engage in less assessment than hospital-based psychologists (Lubin et al., 1986a, 1986b). Furthermore, almost half the psychologists at mental health centers reported spending less time in all forms of assessment—and more time in treatment—than they did 10 years ago.

PSYCHOLOGICAL ASSESSMENT

Testing by clinicians occurs most often in the context of psychological assessment—an evaluation of an individual's cognitive and emotional functioning. These assessments primarily use intelligence and personality tests. Then; goal is to generate an accurate Picture of an individual's problems to identify the interpersonal and environmental factors contributing to these problems and to decide on an effective course of treatment? Tests are important because they Provide a more standardized procedure for gathering and interpreting relevant data, compared to techniques like observation and interview .In some cases the problems identified are serious psychological disturbances referred to as forms of psychopathology or clinical syndromes. In other cases, the problems identified may be less serious problems of adjustment, but they are no less important.

Intelligence Tests

Perhaps you were surprised to see in Table 15.1 that the test administered most often by clinical psychologists was an intelligence test. We already know that intelligence tests provide objective, norm-referenced information about level of intelligence (sec Chapter 9). In this section, we will discuss how intelligence tests in addition can be used diagnostically to suggest hypotheses about emotional and personality characteristics. We will not repeat a discussion of the design and scoring of the Wechsler tests and the Stanford-Binet; if you 'need to refresh your memory, reread the appropriate sections of Chapter 9.

Interpreting Intelligence Test Scores

When clinical psychologists administer intelligence tests, they choose individual tests like the Wechsler tests or the Stanford-Binet. Individual tests provide more comprehensive and sensitive measures) remember that they also are the formal of choice' for diagnostic evaluations in school settings,! In clinical settings, however, we focus on slightly different aspects of intelligence tests. Although we are interested in clients' level of intellectual functioning, we are also interested in analyzing the patterns of test scores and the answers to specific test items.

Pattern Analysis, bittern analysis is a procedure in which we compare the scores on different subscales of an intelligence test and try to interpret the meaning of score differences. The process is based on two assumptions:

(1) Psychological disturbance does not necessarily affect all intellectual functions equally, and (2) different disorders are characterized by different score patterns. Actually, we have already discussed one specific pattern analysis. Chapter 9 noted that we expect a learning disabled individual to show lower scores on verbal subtests than on nonverbal subtests (see p. 317).

The concept of pattern was introduced in early discussions of the WA1S (e.g., Wechsler, 1958). One reason why Wechsler tests are so popular among clinicians is that their scoring system was designed to simplify direct comparison of subtest scores. All Wechsler subtests are converted 10 the same standard score scale. To give you an idea of what pattern analysis involves, some common, interpretations of Wechsler score patterns are presented in Table 15.3. For example, individuals with thought disorders, such as schizophrenia, might show low scores on the comprehension subtest. Why? Because the subtest assesses understanding of cultural standards and individuals with schizophrenia may not have logical ideas or may be unable to explain their ideas logically.

There are many other types of pattern analysis that clinicians may consider. For example, in scatter analysis, we examine the degree to which subtest scores differ. We may even calculate a scatter index representing the variability of subtest scores (Anastasi, 1988). The underlying assumption is that normal individuals show less scatter than individuals with psychopathology. Scatter analysis can focus on comparison of subtest scores or only on scores within a subtest (Groth-Marnat, 1990). In deterioration analysis, we compare performance on two types of subtests: those that theoretically are resistant to the effects of pathology and those that theoretically decline in the presence of pathology (Anastasi, 1988). We can also conduct an intrasubtest analysis, examining the patterns of correct responses and errors within a subtest (Groth-Marnat, 1990). Since items are arranged in order of increasing difficulty, we would expect test takers to pass initial items and begin slowly to fail more difficult items. An inconsistent pattern of passing and failing items might suggest the presence of pathology.

Table 15.3 Sample Interpretations of WAIS—R Subtest Patterns

Subtest Patterns	**Possible Personality Characteristics**
Low digit span, arithmetic, and coding	Anxiety (the "anxiety triad")
High information, vocabulary	Compulsivity, intellectualization
Low comprehension	Antisocial tendencies, thought disorder
Low picture completion	Impulsiveness
Low picture arrangement	Poor rapport, interpersonal problems

Although pattern analysis is very popular, research has produced inconsistent and contradictory findings on its diagnostic value (e.g.. Frank, 1970; Matarazzo, 1972). For example, although a pair of scores may differ by a significant amount, the difference may occur frequently (e.g., Field, 1960). A verbal performance IQ difference of at least 15 points is statistically significant, but in fact occurred in 25% of the WISC—R and 20% of the WAIS—R standardization samples (F. M. Grossman, 1983; Kaufman, 1976). Are all these people suffering from some psychological disturbance?

To some extent, the problem is due to the fact that the same score pattern may be associated with some very different characteristics. For example, according to Table 15.3 a low score on the comprehension subtest could signal the presence of a psychotic disorder like schizophrenia or a personality disorder like antisocial personality. Although clinicians may use pattern analysis to formulate and evaluate hypotheses, they tend to avoid drawing specific conclusions from this single procedure.

Content Analysis. Clinicians may also examine the content of a client's answers to intelligence test items for clues to possible psychological problems. Many individual intelligence tests use free response items that require—test takers to generate rather than select an answer. These responses may suggest the presence of particular thought or personality characteristics.

Consider Wechsler's (1958) example of possible definitions for the word sentence in a vocabulary subtest. An individual defining a sentence as a group of words is likely to differ in experiences or personality from an individual defining sentence as a penalty imposed by a judge. Logical but unusual associations might indicate certain personality characteristics; bizarre associations to vocabulary items might indicate a thought disorder such as schizophrenia.

The form of responses may also have diagnostic implications. Drawing .lines through the walls of mazes might imply impulsivity. Generating overly long, elaborate responses might suggest compulsivity, whereas very short, cautious responses might suggest paranoia (Groth-Marnat, 1990).

As was the case with pattern analysis, content analysis is not a well-validated technique. It is even more difficult to study empirically because the responses to be analyzed are idiosyncratic—particular to the individuals being tested (Anastasi, 1988). Content analysis is best used to suggest characteristics to be investigated by additional assessments.

Observing Behavior during Intelligence Testing

An individual intelligence test is an intimate and challenging .situation. The intimacy result from the one-on-one nature of individual testing; the challenge results from the fact that intelligence tests are power tests (sec p. 34) including at least some items that are too difficult for the average person. The administration of an individual intelligence test, therefore, provides an opportunity to observe how clients react to both cognitive and interpersonal demands.

Because intelligence tests assess reasoning and problem-solving skill, clinicians can draw inferences about cognitive characteristics such as attention, memory, judgment. Tests including verbal tasks, such as the Wechsler tests, provide additional information about communication skills and style of speech? Responses can also reveal cleiycnjLs_j2LcogiiiiJve style, such as being reflective or impulsive in problem solving.

Aspects of personality and emotional functioning can also be observed. Clinicians may note a client's reaction to being tested, including such features as cooperativeness, motivation to perform well and anxiety .about making mistakes.]Clinicians can make inferences about reactions to frustration, such as when a client cannot answer a question or is confronted with a timed task.

The observational data gathered during intelligence testing can be useful for identifying characteristics that should be explored in more depth, confirming initial impressions or the results of other assessments, and raising questions about inconsistencies in performance across different assessments. Like the data gathered from pattern and content analysis, observational data become part of a larger pool of data from which clinicians can develop their hypotheses.

Objective Personality Tests

A key element of psychological assessment is the administration of objective personality tests. Note in Table 15.1 that alter the Wechsler adult test, the most frequently used test is the MMPI. The popularity of objective tests is a direct result of their design. Compared to other personality assessments, objective personality tests have the highest degree of standardization of administration and scoring—two important characteristics of a good test. As a result, these tests tend to be more reliable and valid than other personality assessments.

The two most popular tests in this category are the MMPI and the Edwards Personal Preference Schedule. We already have discussed some of the design features of these two tests in Chapter 4: in fact, these tests were used specifically to contrast different approaches to the design of personality tests. In this section, we will focus on the use of these tests in diagnosis and treatment planning.

Minnesota Multiphasic Personality Inventory

The MMPI is one of the first tests designed to discriminate between normal and pathological individuals. Both the original and revised versions (MMPI and MMPI—2) were developed through empirical scale construction using criterion groups of normal and previously diagnosed individuals (Hathaway & McKinley, 1940, .1943, 1989). The current 567 true/false statements are used to generate scores on four validity scales and ten clinical scales, each tied to a particular personality attribute. The various validity scales were described in detail in Chapter 4 (see p. 102). Table 15.4 presents a description of each clinical scale.

Raw scores on each scale are converted to T-scores with means of 50 and standard deviations of 10. In general, scores of 45 to 57 are considered within normal limits; scores of 65 or more are considered high scores since they are 1.5 standard deviations above the mean, and they are likely to suggest serious problems (Meyer, 1993). -When clinicians use the MMPI—2, they construct profile codes based on the high scores individuals earn on the various clinical scales. The profile codes are written numerically, using the numbers assigned to each scale, with the scales listed in order from highest to lowest score. Codes may include one, two, or three elevated scales.

Table 15.4 Clinical Scales of the MMPI—2

Scale Number/Name	Symbol	Meaning of High Score
I Hypochondriasis	Hs	Many physical concerns/complaints
2 Depression	D	Distress, depression, withdrawal
3 Hysteria	Hy	Naivete, use of neurotic defenses
4 Psychpathic deviate	Pd	Antisocial, rebellious, hostile
5 Masculinity-femininity	Mf[a]	Nontraditional sex-role interests
6 Paranoia	Pa	Suspicious, hostile, externalizing
7 Psychasthenia	Pt	Anxiety, inferiority, obsessiveness
8 Schizophrenia	Sc	Confusion, bizarre ideas, alienation
9 Mania	Ma	High energy, distractible, grandiose
0 Social introversion	Si[a]	Shyness, social discomfort

Considerable research has been devoted to linking profile codes with different types of psychological problems. For example, scales I, 2, and 3 have been referred to as the "neurotic triad," and individuals with high score's on these scales typically show some type of anxiety (neurotic) disorder. Individuals with high scores on the 6 and 8 scales are likely to exhibit features of paranoid schizophrenia (Meyers, 1993).

The MMPI—2 has added a number of supplemental scales (Hathaway & McKinley, '1989). Additional content scales are designed to measure such attributes as health concerns, Type A personality, cynicism, low self-esteem, family problems, and attitudes predictive of work performance. Three additional validity scales are available: a supplemental F-scale, including items not keyed to the original scale, and two scales to identify possible response styles—a scale to check on random responding and a scale to measure acquiescence. As mentioned in Chapter 3, the original MMPI only had validity scales for identification or response sets.

From a psychometric point of view, the MMPI—2, like its predecessor, is a good instrument. Split half coefficients run in the .70s, and median test-retest coefficients are in the .80s. Although it has only recently been published, the validity of MMPI—2 scores is already documented in several hundred studies.

There are, however, several criticisms of the MMPI—2. First, many of the items are scored on more than one scale; the answer to a single item may add points to several different scales. The resulting scales, therefore, are not statistically independent. This raises certain problems for analyzing the scales and also leads some people to question their diagnostic value. Second, critics are concerned about the design of two of the validity scales. The L-scale, designed to identify people presenting themselves in an overly identify people presenting themselves in an overly isealistic/perfectionistic way, is composed of 15 items. For all 15 items, points are added to the scale only if test takers mark them "false." A similar problem exists with the K-scale, designed to identify people who might be "faking good." Of the 30 items, 29 add points to the K-scale only if they are marked "false." From a test design standpoint, it would be preferable to construct these scales using both true and false responses (Kaplan & Saccuzzo, 1993). The addition of the new response style scales provides a mechanism to somewhat address this problem.

A final concern raised about the MMPI—2 is the influence of demographic characteristics-(e.g., Butcher et al., 1990). Research indicates that elevated scores on some scales may reflect characteristics such as age, race, and socioeconomic status, rather than the presence of a psychological disorder. For example, elderly individuals often score higher on the social introversion scale (Meyer, 1993). It is necessary, therefore, that clinicians consider demographic variables when constructing score profiles.

Edwards Personal Preference Schedule

As described in Chapter 4, the EPPS represents a very different approach to personality assessment. First, the EPPS uses forced-choice items that generate ipsative, rather than normative, scores. Second, the statements used in EPPS items were selected using a theoretical 'strategy, rather than the empirical process used in the MMPI. Specifically, statements were written to tap 15 basic needs drawn from Murray's (1938) model of human needs. Third, the tendency of test takers to 'Make good" was presumably controlled by equating each pair of statements for social desirability (Edwards, 1959)

The EPPS is a 225-item test composed of 210 paired statements, 15 of which are repeated at random locations to check for response bias) Why 210 pairs? The test is constructed using statements to measure 15 different needs. A statement for each need is paired twice with one statement for each of the other 14 needs. This requires a total of [15(15- 1)] pairs, or 210 statements.

The 15 needs measured by the EPPS are described in Table 15.5. Because the items are forced choice, each need score is ipsative, expressing the strength of that need relative to the other 14. This means that two individuals who have the same ipsative score on nurturance in fact could differ in the absolute strength of that need. Raw scores are converted into percentile ranks based on the performance of a standardization sample. Separate norms are available for high school and college/adult groups. However, because the percentile ranks are .derived from ipsative scores, it is difficult to use them to compare different people. Differences in percentile ranks actually refer to differences within each test taker in need strength.

Split-half and test-retest coefficients for the need scales range from the .50s to the .80s. Specific need scales generally produce moderate but statistically significant correlations with similar scales in other tests) (Drummond, 1984). Criticisms of the EPPS focus on such factors as the lack of large-scale validity studies and the conversion of ipsative raw scores to normative percentiles (Kaplan & Saccuzzo, 1993). In addition, several studies question whether the forced-choice format really controls the social desirability response set (Feldman & Corah. I960; Wiggins, 1966).

Clearly, the EPPS differs from the MMPI in more than just item format and scoring. The personality dimensions assessed by the EPPS are more like the dimensions we expect lo find in normal individuals. In fact, the EPPS is used more often to explore an individual's personality structure than to diagnosis psychopathology. It typically is used by clinical psychologists who are working with less severely disturbed clients. Other frequent users of the EPPS are counselors, who also use it to explore personality structure, and psychology professors, who often use it as a teaching tool in courses on personality theory or measurement (Drummond, 1984).

Table 15.5 Needs Measured by the Edwards Personal Preference Schedule

Need	**Description**
Achievement	Need to excel, to accomplish something difficult, to rival or surpass others
Deference	Need to support, admire, or emulate a superior, to conform to what is expected
Order	Need to achieve organization, balance, and tidiness, precision
Exhibition	Need to make an impression, to be seen and heard by others
Autonomy	Need to be independent, to break confinement, to resist coercion/restriction
Affiliation	Need to form friendships and join groups, to cooperate, to love
Intraception	Need to analyze feelings and motives of oneself and others
Succorance	Need to be dependent, to seek aid, protection or help

Dominance	Need to influence or control others, to be regarded as a leader
Abasement	Need to submit passively to injury, blame, or criticism, to surrender, to be resigned
Nurturance	Need to help, aid, or protect others, to express sympathy
Change	Need to experience variety, to avoid routine and sameness
Endurance	Need to persevere, to persist, to be strong, to have stamina
Heterosexuality	Need to form erotic relationships, to engage in sexual activity
Aggression	Need to overcome opposition forcefully, to oppose, to light

Projective Personality Tests

Projective personality tests play a different but no less important role in clinical assessment. In Table 15.1, note that two of the lop five tests are projective instruments, the Rorschach Inkblot Test and the Thematic Apperception Test. Projective tests are among the most controversial assessment procedures. Their focus, purpose, and design are very different from other types of tests.

Projective tests were developed specifically for use in psychological assessment by clinicians. We rarely find them used by other types of professionals or in other testing scenarios. They are typically individually administered tests, and considerable time is taken in graduate programs to cover their administration. Projective tests as a group have a number of distinctive features. All projective tests use a relatively unstructured free-response task in which test I takers can produce an almost infinite variety of possible answers. The choice of an unstructured task is deliberate. As discussed in Chapter 4 (see p. 107), /projective tests are designed to reveal unconscious or hidden aspects of personality and thought This locus is designed to complement the design of objective tests, in which test takers directly self-report on their personal characteristics (e.g. Anastasi. 1988).

This underlying assumption about projective tests is frequently challenged by critics. Because projective tests are designed to elicit information below the threshold of conscious awareness, it is difficult to evaluate the validity of these tests. Validity studies often end up comparing the results of one projective with the results of another (e.g. Little & Shneidman, 1959) when the validity of the second procedure itself may be questionable.

However, projective tests remain popular. In part, this reflects the fact that projective tests take a more global approach to personality assessment. Clinicians see this, too, as complementing objective tests, which tend to break personality down into a set of dimensions that are measured separately. The goal of projective testing is to produce a more integrated picture of test-taker personality, rather than to assess individual differences on specific dimensions. Perhaps now you can see why clinicians are not particularly concerned about the variety of answers test takers can produce to projective test items. If these tests were designed to compare individuals on specific dimensions, we would need items for which the answers can be neatly categorized. In projective testing, we are more concerned about developing a unique profile of each person tested.

None of the preceding discussion is meant to imply that projective tests cannot or are not scored. However, instead of the objective scoring rules used on tests like the MMPI, projective tests are likely to be scored using subjective rules—guidelines for the interpretation and classification of answers. This is the second controversial aspect of projective testing. Critics charge that it is difficult to evaluate the adequacy of projective tests because they do not necessarily score test-taker responses numerically. Without a set of scores to analyze, it is difficult to determine the reliability of the scoring procedure or the validity of the scores it produces. There are quantitative scoring systems for some projective tests, although the use of these systems varies widely among clinicians. Studies of reliability and validity have been conducted when these systems are used, but it is difficult to generalize from these studies to the typical, more qualitative analysis of projective tests favored by clinicians.

Rorschach Inkblot Test

Like the MMPI in its category, the Rorschach Inkblot Test is a prototype for projective tests (Rorschach, 1921%The Rorschach is composed of 10 symmetric inkblots: 5 black, while, and gray; 2 that also include touches of red; and 3 that include several other colors. Each is printed on a separate card and handed one at a lime lo the test taker. Test takers are simply asked to describe what the image on the card might be. No other instructions are given. Alter completing all the cards, the examiner readministers each one, systematically questioning test takers about their responses. During his second phase, test takers can clarify or expand on their previous descriptions.

In addition to recording the description of each card verbatim, the examiner may note the way each card is held, the time taken to respond to each card, facial expressions and other nonverbal behaviors—basically anything that might reveal some characteristic of the test taker. Although these additional data are not specifically scored, they become part of the pool of projective information considered during diagnosis.

You may recall from Chapter 4 that there are a variety of procedures for scoring Rorschach responses (sec p. I 13). The scoring dimensions of the system taught in most clinical psychology graduate programs, the Exner system_ (e.g., Exner, 1974, 1978), are described in Table 15.6. Note that we consider much more than just what the test taker sees in the inkblot. Our focus is on the way the test taker organizes the information in the inkblot to construct a response.

Table 15.6 Exner Dimensions for Scoring Rorschach Responses

Dimension	Description
Location	The part of the inkblot associated with each element of the test taker's description
Determinant	How the test taker responds lo the form, color, and shading of the inkblot, including whether the test taker perceives movement within the pattern
Form quality	Extent to which the test taker's description is a reasonable match to the actual features of the inkblot
Content	What the test taker actually perceives when describing the inkblot
Popularity	Extent lo which the responses given are common among people in general or original

The Exner system has been referred to as a psychometric or "signs" system of Rorschach scoring (e.g., Lerner & Lerner, 1986). The task of describing an inkblot becomes a problem-solving task, and dimensions of personality are revealed in the way the individual copes with that task. The link between responses and personality was empirically determined during the development of the scoring system. Studies indicated that certain responses were signs of particular personality characteristics because they occurred more often in certain subgroups of people. Thus, the scoring system uses a comparison of answers in different criterion groups to link Rorschach responses to diagnostic categories. To illustrate in an overly simplified way, if only depressed individuals see a dead turtle in a particular card, the presence of that description is linked to the presence of depression.

When we use this system to score Rorschach responses, our key concerns are the reliability and criterion validity of the test. In other words, scores on the test must be reliable predictors of membership in different criterion groups. In fact, scores generated by the Exner system are (1) reliable over time, (2) good at identifying a number of specific personality characteristics, such as emotional immaturity and intense anger, and (3) good at differentiating between a variety of disorders, such as borderline personality and schizophrenia (e.g., Exner. 1978).

Although this approach lo scoring is similar to what Rorschach himself envisioned (see p. 112). some psychologists see it as omitting an important aspect of projective testing—the development of an integrated picture of each test taker's unique personality. They contrast the Exner approach with a more "conceptual" approach that links Rorschach responses to underlying personality processes (e.g., Lerner iv. Lerner, 1986). In essence, these clinicians advocate interpreting, rather than scoring, Rorschach responses. Their "goal in studying an individual's answers is to reconstruct the processes used to generate-the answers and then to use that information in light of personality theory to draw diagnostic inferences. In this approach, we are more concerned about the construct validity of the test—its ability to identify key aspects of personality and thinking.

There are even proposals for systems to score the Rorschach using this more conceptual approach (e.g., Blatt & Berman, 1984). The focus is on creating composite variables, such as "thought organization," that integrate scores generated by the more traditional scoring system. Rather than replacing the traditional system, this alternative approach to scoring is proposed as a second stage or "higher-order" analysis.

At this point, you might well ask. "What's the point of giving this test?" True, there is still extensive controversy about how to use the Rorschach, but administration and scoring are much more standardized today than ever before. And in spite of all the controversy about how to score and use the test(the Rorschach is emerging as a reliable and valid instrument. A recent analysis of the combined statistical data in 39 Rorschach papers, published during the 1970s, indicated an overall internal consistency coefficient of .83 (Parker, 1983). Other studies using the Kuder-Richardson procedure generate an average coefficient of .77 (e.g., Wagner et al., I986> In light of these data, the Rorschach likely will remain one of the most popular clinical tests.

Thematic Apperception Test

The contrast between the Thematic Apperception Test (TAT) and the Rorschach parallels the contrast between the Edwards and the MMPI. Both the TAT and the Edwards reflect a theoretical approach lo scale construction, whereas the Rorschach and the MMPI use an empirical approach based on the contrasted performance of criterion groups. In fact the TAT and the Edwards are designed to tap dimensions of personality drawn from the same theory—Murray's (1938) theory of needs. Like the Edwards the TAT was not developed specifically as a diagnostic instrument: it was designed to explore personality issues in both clinical and nonclinical groups. In contrast, the MMPI and the Rorschach were designed to identify clinical disorders.

The TAT is designed to assess the strength of different needs as they are reflected in stories (Murray, 1943). The TAT stimuli consist of 19 cards with drawings of people in ambiguous situations, plus a blank white card. Some cards are designed to be used only with particular types of people, such as adult males, adult females, boys, or girls. Test takers are asked lo tell a story about each picture, including the events leading up lo the scene on the card and the outcome of the scene on the card. For the blank card, test takers are asked to imagine a picture, describe it and then tell a story about it. In addition lo recording the story verbatim, the examiner records the lime taken to respond to each card. The test is projective in that each card can elicit a wide variety of stories, and we assume that the story told by a test taker reflects issues that are important to that person. Furthermore, the lime taken to respond lo a card may imply something about areas or issues that are particularly sensitive.

Rather than being scored in the traditional psychometric sense, TAT stories are analyzed on several dimensions. First, the examiner identifies the hero of the story, the protagonist or the central figure. It is assumed that the test taker identifies with this character and projects onto the hero many personal needs and problems. Second, the actual content of the story is analyzed in terms of the needs expressed and the environmental forces that affect satisfaction of those needs—referred to as *press*. The needs examined are similar lo those assessed by the Edwards, which also uses Murray's (1938) model. A list of Murray's complete set of needs is given in Table 15.7.

When analyzing the content of TAT stories, clinicians pay attention to the intensity, duration, and frequency of each need and press.

Table 15.7 Needs Explored by the TAT

Needs Also Assessed in the Edwards

Achievement	Autonomy	Abasement
Deference	Affiliation	Nurturance
Order	Succorance	Heterosexuality
Exhibition	Dominance	Aggression

Additional Needs	**Description**
Blame avoidance	To be well behaved, to obey, to avoid blame or punishment
Counteraction	To defend one's honor by taking action, to refuse to admit defeat, to retaliate
Defendance	To defend oneself against blame, to offer explanations or excuses, to justify actions
Harm avoidance	To avoid pain or injury, to take precautions, to escape from dangerous situations
Inferiority avoidance	To avoid humiliation or embarrassment, to fear failure
Play	To enjoy oneself, to relax, to have fun, to avoid tension
Rejection	To be aloof and indifferent, to ignore or snub others
Sentience	To seek out sensuous experience, to enjoy sensation
Understanding	To analyze experience, to synthesize ideas, to engage in abstraction

They attempt to create a picture of the relative importance of each element in the test taker's life. For example, the presence of several stories revolving around achievement or accomplishment implies that this particular need is strong. Frequent descriptions of how authority figures react to the hero's actions imply that the judgments of others are a significant press in the test taker's life.

Unfortunately, administration and interpretation of the TAT are not well standardized. In practice, most clinicians administer 6 to 10 cards; bun there is considerable variability in the cards used and their order of presentation/ (Ryan, 1985). Over the years, a large amount of normative information has been compiled on the frequent types of responses to each card (e.g., Henry, 1956). However, many clinicians prefer to use their own experiences with TAT stories when analyzing their content. The same is true for scoring systems. A variety of scoring systems has been proposed, including systems focusing on achievement issues (McClelland el al., 1953), on gender identity (May, 1966), on assessment of defenses (Cramer, 1983) and on ego development (Sutton & Swensen, 1983)7 However, most of these systems are complex and difficult to learn and do not provide the type of whole-person picture clinicians seek from TAT responses. In fact, qualitative methods are used much more frequently than quantitative methods (e.g., Westen. 1991).

Because the TAT is rarely scored numerically, it is difficult lo draw conclusions about its reliability. Agreement among clinicians in TAT interpretations varies widely (e.g., Harrison & Roller, 1945), and the median test-retest correlation when stories are scored quantitatively is only about .30 (e.g., Kraiger, Hakel, & Cornelius, 1984). Given the preference for qualitative scoring, it is also difficult lo study its criterion validity. Most clinicians are more concerned about the construct and content validity of the TAT, noting that it does elicit much data relevant to needs, conflicts, and environmental influences. When used as part of a larger battery of tests and accompanied by detailed interviews and observations, clinicians view the TAT as useful for formulating hypotheses about the structure and dynamics of personality (Ryan, 1985y

There are several different special adaptations of the TAT. In addition to Murray's original test, there are two forms for children, one using animal characters and one using human characters (Bellak, 1954; Bellak & Hurvich, 1966). There are also two forms designed for older adults (Bellak & Bellak, 1973; Wolk & Wolk, 1971). The goal in developing these adaptations was to create stimulus cards likely to elicit identification with the characters by different groups of people, thereby facilitating projective responses. However, none of these alternative TATs appears lo improve significantly on the information obtained through the traditional test (e.g., Anastasi, 1988), and Murray's original test is still the preferred technique.

Other Types of Projective Tasks

In addition to using complex projective tests, clinicians may choose to include other types of projective tasks in their assessment batteries. They are often used as screening tests to aid clinicians in developing initial hypotheses about client characteristics and problems. These other tasks are smaller in scope and easier to administer than tests like the Rorschach or the TAT. The popularity of these techniques is evident from examining Table 15.1. Although none are included in the top five assessments, we find three of these other techniques in the next set of five: sentence-completion tests and two types of drawing tasks.

Sentence-completion and drawing tasks provide a good contrast with other projective assessments. Sentence completion uses verbal stimuli to elicit verbal responses, in contrast to the picture-stimulus/verbal-response design of the Rorschach and the TAT. Drawing tasks combine verbal instructions with a psychomotor task. By using a variety of projective techniques, clinicians provide individuals with many different ways to express themselves. Some of us may respond more to verbal tasks, cithers of us to pictures, and still others of us to the opportunity to draw.

Sentence-completion Tasks

Sentence-completion tasks have become very popular in clinical assessment. They are simple to administer and take little time. Test takers are presented with a series of sentence steins, such as "My mother..." or "I hate...," which they complete in their own words. Although some people view these as self-report tasks, most clinicians see sentence completion as a projective technique. Test takers are confronted with ambiguous stimuli, and the way they respond to these stimuli is likely to reveal their personal issues.

There are two approaches to using sentence-completion tasks. In one approach, clinicians write their own sentence stems to tap specific aspects of personality or functioning. The goal is to elicit responses that are relevant to issues that have been identified as important through other assessments. This ability to tailor the task to each individual is one reason why sentence-completion tasks are so popular. However the responses to clinician-constructed stems must be evaluated qualitatively within the context of the clinician's prior experiences and other data on the test taker.

The second approach makes use of standardized sentence-completion tests, such as the Rotter Incomplete Sentences Blank (Rotter & Rafferly, 1950) or the Incomplete Sentences Task (Lanyon & Lanyon, 1980). These instruments are designed to be scored, and test-taker scores are compared to normative data collected during their standardization. The Rotter test consists of 40 sentence stems selected for their relevance to clinical issues, a theoretical approach to test construction. Test takers are instructed to make a complete sentence from each stem that expresses their true feelings.

Answers are compared to sample answers in the manual and scored on a 7-point scale reflecting emotional adjustment (7) to maladjustment (1). Scoring is based on both the length and content of answers. For example, short, concrete answers with neutral to positive content receive high scores. Long answers with depressed or anxious content receive low scores. The scores on the 40 stems can be summed to provide an overall index of adjustment. Because the manual includes a good sample of completion answers, scoring is relatively objective and straightforward. However, little reliability or validity data are available.

The Incomplete Sentences Task (1ST) reflects a more empirical approach to test construction. The three dimensions assessed—hostility, anxiety, and dependence—were selected for their special relevance to the adjustment of secondary school and college students (Lanyon & Lanyon, 1980). The stems were selected by a contrasted-groups approach comparing the answers of students rated high or low on each dimension by their teachers. Two forms a| available, one for grades 7 lo 12, the other for college years. Each completer is scored on a 3-point scale (0, I, 2) by comparison to sample answers in the manual. Answers are grouped by personality dimension and totaled; total raw scores are converted to percentile ranks based on performance of students in the standardization samples.

In contrast to the Roller, several studies have been undertaken lo evaluate the reliability and validity of the 1ST. Initial data indicate that the scoring of the 1ST is reliable and that it is useful in assessing individual differences or the three personality dimensions. (Cundick. 1985; Dush. 1985).

Drawing Tasks

Drawing tasks have been used in a variety of ways in psychological assessment. One early application of drawing tasks was the assessment of intelligence (e.g., Goodenough, 1926). There are predictable developmental changes in the detail of children's human figure drawings.')As children get older, they incorporate more detail into their drawings and represent the arrangement and proportion of body parts in more realistic ways. A scoring system was developed to translate the level of detail and accuracy in a drawing into an 1Q score.

Today, drawing tasks are used primarily as projective techniques. These tasks are sometimes referred to as *expressive techniques* because they provide test takers with relatively unstructured opportunities for self expression (e.g., Anastasi, 1988). From a diagnostic standpoint, the construction of the drawings—the elements included, their relative locations and proportions—are believed to reflect emotional conflicts and needs.

The two most popular tests as indicated in Table 15.1 are the Draw-a-Person Test (Machover, 1949, 1971) and the House-Tree-Person Technique (Buck, 1948, 1981). In the Draw-a-Person Test (DAP), the test taker is given a blank, unlined sheet of paper and a pencil and instructed to draw a person. After completing the drawing, the test taker is given new materials and asked lo draw a figure of the opposite sex from the first one. During the drawing process, the examiner takes notes on the order of drawing elements, the comments made by the test taker, and the test taker's nonverbal behavior during the task. The examiner may follow the drawings with a series of questions designed lo elicit information about the figures, such as whom they represent and their ages and genders (Machover, 1949, 1971).

DAP scoring is qualitative. Attention is paid to the size of the figures, the level of detail in each, their body proportions, the sequence of drawing body elements, the inclusion of clothing, and the position of each figure and body part. The goal is to construct a primarily psychoanalytic description of personality Based on the analysis of the figure.. Although an interpretive guide is presented in the manual, there are no norms and no reported reliability and validity data. Attempts by other psychologists to validate the test have been unsuccessful (e.g., Klopler & Taulbee. 1976).

The House-Tree-Person Technique (HTP) is a somewhat more complex drawing task. The tree objects drawn are assumed to be symbolic of aspects of the test taker's life (Buck, 1948). The house is symbolic of home and family life, the tree is symbolic of the test taker's relationship to the environment, and the person is symbolic of the test taker's interpersonal relationships. Test takers are assumed to project their feelings about these elements into their drawings. In its revised form (Buck, 1981), the drawing task has two phases, each including a drawing and a structured interview stage. In the first phase, the test taker is given blank, unlined paper and a pencil and asked to draw a house. This is followed by two other sheets of paper with instructions to draw (1) a tree and (2) a person. The drawing stage is followed by a series of 60 questions designed to elicit information about the meaning of each figure. In the second phase, test takers draw the same three figures again on separate pages using a set of eight or more crayons. This second drawing stage is also followed by a series of questions.

During the drawing tasks, the examiner notes the time taken to respond, the order in which parts are sketched, and the test taker's comments and nonverbal behavior. During .the postdrawing interviews, examiners may deviate from the question list to follow up on points that appear diagnostically relevant.

Drawings are quantitatively and qualitatively scored. The quantitative analysis compares test takers' drawings to those produced by a norm group of 140 adults representing a variety of levels of intelligence. Points are earned for different features of each figure and are converted to an overall IQ score. Qualitative analysis is based on comparison to a similarly small norm group (150 adults) representing a variety of clinical syndromes.

Although the HTP technique was developed through more empirical procedures, it has not emerged as a psychometrically sound test. The norm samples are extremely small, and the sample for qualitative analysis does not include a group of normal individuals (Killian, 1984). Few reliability and validity data are provided, with the exception of demonstrations that performance on the task does reflect both intellectual and personality factors (Buck, 1981). For example, the IQ scores generated by the HTP do show moderate correlations with other measures of intelligence, such as the Wechsler tests and the Stanford-Bi net.

Although drawing tasks remain very popular, they are the psychometrically weakest type of projective technique. Many psychologists caution against overinterpreting figure drawings and advise using them instead as part of a diagnostic interview during the initial screening of a client (e.g., Anastasi. 1988; Kaplan & Saccuzzo, 1993). Until more reliability and validity data are available, these techniques are better viewed as investigative tools than as psychological tests.

Issues in Psychological Assessment

As you read through the preceding sections, you may have noticed less concern with traditional psychometric indexes of test adequacy, such as test-retest reliability and criterion validity, than we see in other testing scenarios. For example, in our discussion of testing in educational settings, we referred lo specific cutoff points needed on certain tests to justify educational placements. We discussed how the courts have restricted the use of certain tests because of their failure to demonstrate statistically significant relationships with important criteria. You may also have noted that clinical assessment relics more on subjective information and interpretation of responses than does testing in other settings.

Clinical psychologists themselves have different perspectives on the purpose of testing and therefore on the need for tests lo demonstrate the statistical properties usually used to identify a test as a "good" test. For example, testing in clinical settings can be viewed from a psychometric or a clinical perspective (Lerner & Lerner7 1986). The psychometric perspective is more concerned about testing as a measurement process. The emphasis is on creating standardizes administration and scoring procedures and demonstrating that-tests are reliable and valid measures. The examiner is seen as a potential source of measurement error, as someone who can introduce bias and subjectivity into the measurement process, and efforts are made to reduce the role of the examiner to that of a recorder of data. In essence, the data collected by an examiner during testing should be equivalent to the data collected by a machine such as a computer.

The clinical perspective is very different. From this perspective, the test administrator is a critical element of the data gathering process (Lerner & Lerner, 1986). It is the examiner's ski11, judgment and intuition that enable us to understand the person being tested. The test itself is simply a tool that, when used by a skillful psychologist, can provide useful information for diagnosis and treatment planning. In this perspective, it is difficult to talk meaningfully about the general statistical properties of a test. When used by different clinicians, who differ in their training and expertise, tests are likely to produce different results.

Probably the most realistic position on clinical testing is somewhere between these two extremes. The issues being addressed in clinical assessment are complex and elusive. Examiner skill is likely to be of more importance in clinical assessment than in the assessment of math skills. We might expect more variation in administration and more interpretation of responses. However, the quality of a decision is influenced by the accuracy of the data being considered. Although tests for clinical assessment may not need to be held to the rigorous psychometric standards used in educational decisions, some demonstrations of reliability and validity are necessary. No test will be useful if it cannot accurately represent what it is designed to measure.

NEUROPSYCHOLOGICAL ASSESSMENT

Neuropsychology is a specialization within psychology and medicine that studies. The relationship between behavior and the brain. In clinical neuropsychology, we examine the impact of brain injury, brain disease, or abnormal brain development on cognitive and behavioral functioning. As a multidisciplinary area, clinical neuropsychology involves both medical and psychological testing. Psychological tests typically are administered by clinical psychologists with additional training in neuropsychology.

The majority of neuropsychological testing occurs in medical settings, such as veteran's administration hospitals and medical centers. However, neuropsychological tests may also be given to individuals seeking outpatient treatment when patterns of behavior suggest a problem in brain function. Although you may think of neuropsychological testing as important primarily in cases of stroke or Alzheimer's disease, neuropsychological tests are also important in the diagnosis of problems such as learning disability, mental retardation, and attention deficit disorder, a syndrome sometimes associated with hyperactive behavior.

Intelligence Tests

Once again, intelligence tests like the Wechslers can provide useful information. In the simplest scenario, intelligence tests can be part of the diagnostic battery used lo identify individuals with specific cognitive deficits. This procedure was discussed in Chapter 9 relative to identifying learning disability and mental retardation (see p. 315). I f particular score patterns are observed, test takers can be referred for additional evaluation to determine whether these patterns reflect neurological problems or learning failures that might be tied to experience.

Similarly, current intelligence test scores can be compared to earlier scores lo sec if changes in ability have occurred. For example, evaluation of an individual with Alzheimer's disease—a progressive degeneration of brain functions—might involve repeated administration of an intelligence test lo track the nature and extent of cognitive decline.

Considerable research has been conducted on using Wechsler tests to identify different cognitive deficits. For example, some psychologists suggest that individuals with a learning disability show low scores on the arithmetic, coding, information, and digit span subtests (e.g., Kaufman, 1979). Other psychologists propose that learning disability also is characterized by higher scores on performance subtests that do not require sequential actions, such as object assembly, block design, and picture completion. Spatial subtests that require sequencing, such as digit span, digit symbol, and picture arrangement, are usually more difficult (Bannatyne, 1974).

However, as often happens with pattern analysis, other studies challenge these proposals. For example, the difference between scores on scquencing/nonsequencing performance subtests is not present for all learning disabled individuals (Kavale & Forness, 1984). In addition, the pattern is not unique to learning disability. A similar pattern has been observed in the test scores of delinquents (Groff & Hubble, 1981) and emotionally handicapped children (Thompson, 1981). Once again, it is clear that diagnostic decisions must be based on data from a variety of assessments.

Neuropsychological Tests

As £he field of neuropsychology grew, so did the collection of tests designed specifically to assess brain damage. The majority of these tests focus on spatial perception' and memory, two cognitive abilities believed to be greatly affected by brain functioning (Anastasi, 1988). Neuropsychological tests can be grouped into two broad categories: tests of specific tasks and neuropsychological test batteries. All tests in both categories require individual administration.

Tests of Specific Tasks

Tests of specific tasks focus on a single cognitive function that is likely lo be affected by brain impairment. The development of these tests dales to the post-World War I era, when observations of injured soldiers (e.g., Goldstein & Scheerer, 1941) and children who had experienced birth traumas or serious infectious diseases (e.g., Werner & Strauss, 1941) identified a number of specific cognitive problems. The most common characteristics were impairments in perception, abstract thinking, and memory.

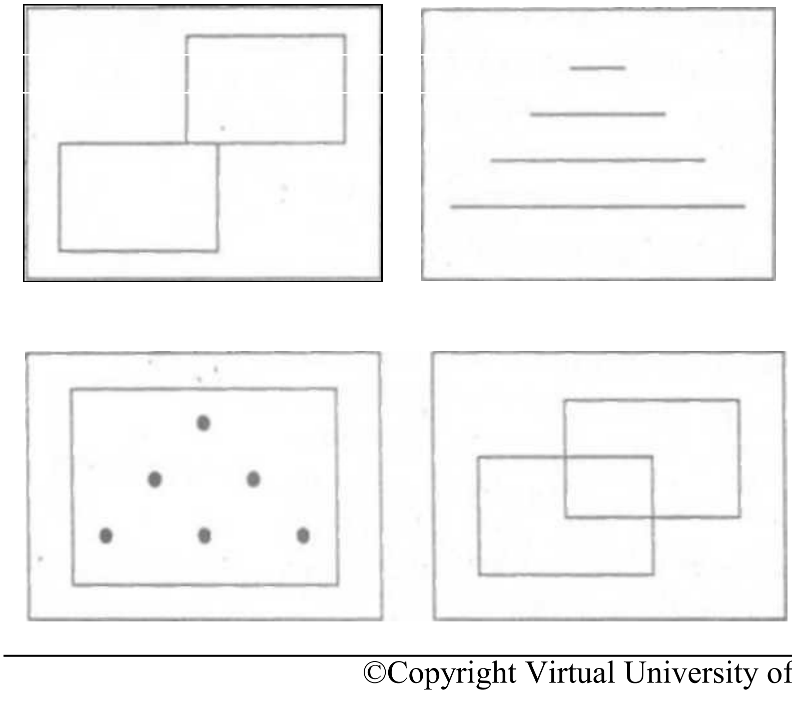
According to Table 15.1, the two most popular tests in this category are the Bender Visual Motor Gestalt and the Wechsler Memory Scale. They provide an excellent contrast between tests focusing on perceptual issues and tests focusing on memory abilities.

Bender Visual Motor Gestalt. The Bender was discussed briefly in Chapter 9 as a test useful in assessing learning disability. '.What we did not specify is that its use in these assessments derives from proposals that learning disability reflects impairment in central nervous-system functions (e.g., Sattler, 1988).

The Bender consists of nine simple line drawings presented one at a time on separate cards. The drawings include a variety of elements, such as dots, circles, and angles that are combined into patterns or figures. The designs were selected specifically to tap different Gestalt laws of perception—processes for integrating elements into a whole unit or Gestalt (Wertheimer, 1923). Simulated Bender drawings are presented in Figure 15.1. Note that each design contains elements in particular orientations relative to each other. Some touch, others overlap, and still others are contained within a larger shape.

When developed in 1938, the Bender was proposed as an index of perceptual-motor maturation (Bender, 1938). However, for many years, there was no systematic scoring procedure for performance on these designs. Even today, there are a variety of scoring procedures, including procedures for using performance to diagnose emotional and psychological disorders (e.g., Hutt, 1985).

Figure 15.1 Simulated Bender drawings



In this aspect, the Bender is similar lo the Rorschach and the TAT two other tests that can be scored and interpreted using .several different systems.

In general, the most widely accepted use of the Bender is for diagnosis of brain dysfunction (Groth-Marnat, 1990). Norms and scoring systems are available for children (e.g., Koppit, 1975) and for adults (e.g., Lacks, 1984). Clinicians look for such features as simplifying a figure by omitting elements rotating a figure or the elements within il. and replacing a straight or curved line with the other type.

Test-retest reliabilities vary greatly according to scoring system, age of test taker, and lime between testing; coefficients for the Koppit and Lacks systems typically are in the .90s. Although the results of validity studies are mixed, the Bender appears to be able to discriminate brain-damaged from nonbrain-damaged individuals and lo provide a good index of perceptual-motor development (e.g., Koppitz, 1975; Lacks, 1984). However, correlations with scores on intelligence and achievement tests are low to moderate, indicating that the Bender should not be used as an overall measure of either aptitude or achievement.

Wechsler Memory Scale. As its name implies, the Wechsler Memory Scale CWMS) is a test of short- and long-term memory deficits (Wechsler, 1945). Wechsler's goal in developing the test was to produce a measure of memory ability that was not highly correlated with intelligence (lleiby. 1984). A second form of the test was developed for use in retesting clients (Stone & -Wechsler. 1946), and a revised form tied lo current information-processing concepts is in development (sec Russell. 1975).

Each form of the WMS consists of seven subtests, described in 'fable 15.8. The test is extremely easy for normal people and docs not capture individual differences in normal memory skills. Several of the subtests were borrowed fro4ft other tests. For example, the digit strings in the memory span subtest for Form I were taken from the original WAIS digit span; the Form II digit strings were taken from an early version of the Army Alpha, the first group intelligence test for adults (lleiby, 1984). Although this may seem odd if the goal is to produce a test that is not correlated with IQ, this particular subtest taps basic processing skills rather than level of intelligence. Scores on the digit span subtest are not significant predictors of IQ scores for normal individuals.

The WMS is designed for ages 20 to 64. The test is scored by summing performance on the subtests, weighting the sum by adding a constant based oil test-taker age, and converting the weighted sum to a memory quotient CMQ) with a mean of 100. There are no norms for performance on individual subtests, no norms for determining how much below the mean an MQ must be lo suggest impairment in brain functions, and no internal consistency data. In light of these statements, you might wonder why clinicians bother to administer this test! In fact, the popularity of the WMS is due to the absence of other tests specifically designed to measure memory functions and to several significant validity studies.

Table 15.8 Subtests of the Wechsler Memory Scale

Subtest	**Description**
Personal and current information	Questions on name, date of birth, important recent events (e.g., the current president)
Orientation	Questions about time and place
Mental control	Tracking questions (e.g., counting by 3's)
Logical memory	Oral questions on two paragraphs read by the examiner
Memory span	Reciting strings of digits forward and backward
Visual reproduction	Drawing simple geometric figures from memory after studying each for 10 seconds

Associate learning	Paired words (i.e., paired associates) to be learned in three trials

For example, scores on the WMS do decline in normal individuals with increasing age, and scores on the WMS do correlate with IQ scores for individuals with mental retardation and other forms of brain damage (Prigatano, 1C)7S). It appears, therefore, that the WMS can be useful in identifying individuals with declining and impaired memory functions.

However, the technical data on the WMS raise an important point relative lo the evaluation of neuropsychological tests. In many cases, the reliability of these tests is difficult lo determine. We cannot use normal individuals in reliability studies because the tests are too easy for them. On me other hand, it is difficult to conduct reliability studies with brain-damaged individuals. People who have actual neurological problems are likely to deteriorate further over time—or to show improvement if they are involved in a productive type of treatment. As a result, many psychologists infer about reliability from validity data. Since a test cannot be valid unless it is reliable, they conclude that a test producing a valid measure in fact must have adequate reliability.

Test Batteries

The trend in neuropsychological assessment has mirrored the trend in other areas. Initially, assessment plans focused on the use of a separate series of tests, each measuring a different aspect of brain function. Over time, test list of frequently used tests in Table 15.1 probably reflects the small number of clinical psychologists who are trained to use these complicated tests.

Unlike the test batteries used in educational settings, neuropsychological test batteries are not always normed on a single group of people. In some cases, the batteries have evolved over time, with additional tests added at later dates. However, the batteries still have an important advantage over the use of individual tests: Bach battery represents a particular point of view about brain functions, and the tests within a battery are organized lo assess functions corresponding to those views.

Halstead-Reitan. The Halstead-Reitan Neuropsychological Battery is a series of tests designed to use performance on tasks lo identify deficits within different brain areas (Reitan, 1969). You probably have heard about studies of brain localization and lateralization in other psychology classes. The Halstcad-Reitan is designed to test abilities that are believed to be (1) localized within certain brain lobes or (2) lateralized within certain brain hemispheres.

Development of the battery began in the late 1930s, and additional assessments were added over a period of years. Drawing on research about brain-behavior relationships, the battery was designed lo include tasks that are known to involve certain brain areas (Reitan & Wolfson, I985). For example, we know that control of motor movements is localized in the frontal lobe. We also know that motor control of the body is lateralized; therefore, the right frontal lobe of the brain controls motor movements on the left side of the body. An individual who cannot perform motor tasks with the left hand is likely to have damage to the right frontal lobe.

Different versions are available for testing young children, teens, and adults. The basic battery consists of five tests presented in 'table 15.9. Other physical and sensory tests are sometimes added, such as a strength-of-grip test and a test of response lo sensory informal ion that is presented only lo one side of the body (e.g., the left ear). In addition, test takers may also be asked lo complete, an intelligence test, such as the Wechsler, and a personality test, such as the MMPI. Total testing lime, therefore, can be up lo 12 hours, which is divided over a number of sessions.
Scores on each test are used to determine an impairment index, and the total number of tests for which impairment is seen determines an overall impairment index. Validity studies indicate that the battery is able

lo identify conditions such as tumors within different hemispheres of the brain and within different brain lobes (e.g., Reitan, 1967).

Luria-Nebraska. In contrast to the views of Halstead and Reitan, Luna viewed the brain as an integrated system in which many areas together controlled patterns of behavior (e.g., Luria, 1973). Luria's original assessment method was very subjective and did not use a standardized procedure.

Table 15.9 Basic Tests in the Halstead-Reitan Adult Battery

Test	**Purpose**
Category test	Assesses learning and concept formation skills through a series of discrimination learning tasks in which test takers must figure out or "abstract" the principle used to group geometric shapes into categories.
Tactual- performance test	Assesses psychomotor and memory abilities by requiring test takers to put differently shaped blocks into holes on a board. Three trials: using preferred hand only, nonpreferred hand only, and both hands. Followed by request for test taker to draw a diagram from memory of the board with the blocks in their proper positions.
Speech sounds perception test	Assesses auditory-visual coordination and language skill by requiring test takers to identify the nonsense words presented to them on tape from written multiple-choice lists. All words use “ee” as the vowel sound but vary in first or last consonant.
Rhythm test	Assesses nonlanguage auditory perception by requiring test takers to discriminate between same and different pairs of rhythmic beats presented on tape.
Finger-tapping test	Assesses motor speed and hand preference by requiring a test taker to tap each index finger as fast as possible for a series of I0-second trials, alternating hands on successive trials.

It was not, therefore, well received in this country. However, a psychologist at the University of Nebraska, C. J. Golden, subsequently standardized his procedures, producing a test known as the Luria-Nebraska Neuropsychological Battery2 (Golden, Purisch, & Hammeke, 1985).

Designed for ages 15 through adult, the Luria-Nebraska includes 11 clinical scales, described in Table 15.10. Although the Luria contains more scales than the Halslead testing time typically is only about l hours. Each item is scored 0 to 2, where 0 equals normal performance. Raw scores on each test are converted to T-scores, which are evaluated relative to expected performance based on age and educational level to determine the extent of impairment.

Table 15.10 Subtests of the Luria-Nebraska Battery

Test	**Purpose**
Motor scale	Assesses performance of motor movements with the hands, arms, face, and mouth. Some are imitative, others are verbal instructions.

Rhythm scale	Assesses production of sound patterns, such as short melodies, and perception of sound qualities and patterns.
Tactile scale	Assesses processing 61' tactile information when blind-folded, such as letters traced on the back of the hand.
Visual processes scale	Assesses visual object recognition skill with line/ perspective drawings, etc.
Receptive speech scale	Assesses understanding of spoken language, from speech sounds to complex sentences.
Expressive speech scale	Assesses ability lo speak, using speech sounds, words, phrases, etc.
Writing scale	Assesses writing skills, including copying, spelling, and spontaneous writing.
Reading scale	Assesses reading skills, including letters, symbols. words, sentences, and stories.
Arithmetic scale	Assesses reading/willing numerals and simple/complex arithmetic operations.
Memory scale	Assesses immediate and delayed memory for pictures, rhythms, words, word-picture pairs.
Intellectual processes scale	Assesses general intellectual skills using tasks similar to traditional IQ tests.
Intermediate memory, scale	Assesses recent and incidental memory by asking for descriptions of earlier activities during the assessment.

Although Luria had a more holistic view of brain activity, the Luria-Nebraska is designed to identify areas within the brain that are not functioning properly. The process for doing this, however, differs from the Halstead procedure. Since Luria saw many areas as influencing each task, the key to identifying possible areas of brain damage is comparison of the various subtest scores. Scores on tests designed to lap cither right or left hemisphere activities ate compared to identify damage within hemispheres. Scores on tests designed to lap activity in different lobes are compared in a parallel way.

As a relatively new test, the psychometric properties of the Luria-Nebraska are still under evaluation. Reviews of research indicate that reliability coefficients range from .54 to 1.0, depending on the type of coefficient computed, and that the summary scores on the Luria correlate at the .70 range or better with scores on the Halstead-Reitan (e.g., Franzen, 1985).

Issues in Neuropsychological Assessment

Neuropsychological assessment has become an important specialty area, due in part to the explosion of information about brain-behavior relationships in the last two decades. Even today, data from these assessments are often used as part of larger research projects on neurological functioning. Two points about these assessments are important to note.

First, neuropsychological assessment is a truly interdisciplinary enterprise. In this section we have focused on how tests of behavior can be used to make inferences about brain functions. However, a variety of other techniques is available to study brain activity, and these are often used along with tests of behavior. The most common of these are imaging procedures—techniques that generate pictures of brain structure or brain function (e.g., Andreasen, 1988). For example, an individual with a possible neurological deficit may receive a computerized tomography (CT) scan in which multiple x-rays are used to create pictures of brain sections. A more sophisticated technique magnetic resonance imaging (MRI) uses magnetic fields and radio waves to produce computer-enhanced images of brain structure. Brain function can be mapped using a PET or positron emission tomography scan in which radioactively tagged sugars are tracked as they are absorbed by active brain areas.

Second, neuropsychological assessments are complex and specialized procedures. The typical clinical psychology graduate program does not include extensive training in neuropsychology. Clinicians interested in neuropsychology often receive additional training at the postgraduate level. In fact, the field has grown so rapidly that psychologists may now specialize within neuropsychology in such areas as the evaluation of children, the evaluation of adults, or the geriatric evaluation. Students interested in more information about this area should consult the following references: Albert, 1981; Heaton & Pendleton, 1981; Satz & Fletcher, 1981.

Testing in Counseling Settings

Why a separate chapter on the use of tests in counseling settings? Don't counselors use the same tests as clinicians? Not necessarily. Many students find it difficult to distinguish between clinical and counseling settings because clinicians and counselors do engage in some similar practices; both types of people may be trained to use some of (he same therapeutic approaches, psychological tests, and other assessment techniques. However, since they differ in the focus of their training and their assessment and treatment goals, we often find different tests used in clinical and counseling settings. To underscore the importance of distinguishing between these settings, we will discuss the nature of counseling activities as well as the psychological tests used.

USES OF TESTS IN COUNSELING SETTINGS

Counseling is found within schools, colleges, and universities, in government and social service agencies, and in prisons, hospitals, and private practice offices (Wise, 1989). The majority of counseling occurs in educational settings and in private practice (Watkins et al., 1986). The testing component of counseling may involve the administration of interest tests, ability tests, and personality tests.

Nature of Counseling Activities

The focus of counseling is to identify people's abilities, personality characteristics, and patterns of interests and lo assist people in making choices and changes to improve their sense of well-being and life-styles (Pietrofesa, Hoffman, & Splete, 1984). One way to understand better the specific uses and contributions of counseling is to contrast the fields of counseling and clinical psychology.

Counseling versus Clinical Psychology

Although these two fields do overlap, there are several important distinctions between them. Table 15.11 presents some of these contrasts based on the writings of Anastasi (1979) and Wise (1989). Unlike clinical psychology, with its historical roots in the medicine model, counseling psychology evolved from the vocational guidance movement. It is more focused, therefore, on assisting people with making decisions about their futures. The decisions may involve vocational (career) issues or personal issues such as substance abuse or marital/family discord. Although counseling psychologists are often involved in therapeutic activities, they are less likely to treat individuals with serious psychopathology.

Comparison of the employment of counseling and clinical psychologists identifies additional differences between the fields (Watkins et al., 1986). 31.1% of clinical psychologists were employed in private practice whereas 31.7% of counseling psychologists worked in university settings.

Table 15.11 Distinctions between Clinical and Counseling Psychology

Clinical Psychology	Counseling Psychology
Evolved from medical model with an emphasis on research and treatment (scientist-practitioner model).	Evolved from the vocational guidance movement with an emphasis on assessment and placement decisions.
More concerned with "identifying problems in personality functioning.	More concerned with identifying patterns of personality and interest more emphasis on strengths).
Seeks to identify causes of problems; interest in exploring the past.	More concerned with present and future functioning.
Seeks to remedy problems by changing aspects of personality and behavior.	Seeks to solve problems using individual's strengths without making major personality changes.
Likely to work with people whose problems are disabling and prevent them from living normal lives.	Likely to work with people whose problems are disruptive but who function adequately overall.

Unlike the counseling psychologists at universities, who worked primarily in counseling centers, the clinical psychologists at universities typically held academic positions in psychology departments. Furthermore, clinical psychologists were almost three times more likely to be employed in hospitals (15% versus 5.4%).

Other People Who Engage in Counseling

A wide variety of people identify themselves as "counselors" or their activities as "counseling." There also are a number of different counseling specialties, including substance abuse counseling, marriage and family counseling, and vocational counseling. But in many stales, the terms "counselor" and "counseling" are not legally regulated terms, and people do not need certain specific credentials to use these labels.

A counseling psychologist typically has a master's or doctoral degree in counseling psychology, which may be offered either through a psychology or an education department. Other individuals who may use the term counseling include social workers who have specialized in psychiatric social work, people with master's degrees in education who specialized in either school or nonschool counseling, and people with bachelor's degrees who have received additional training in substance abuse, marriage, or family counseling. To keep things simple, our discussion will focus on counseling psychologists—the group of people with the most extensive training in psychological assessment and treatment.

Tests Used by Counseling Psychologists

Because of their focus on vocational and personal problems, counseling psychologists tend to use psychological tests that assess people's abilities, personalities, .and interests. Tables 1 1.2 and 15.13 present the tests used most often by two groups of counseling psychologists. The data in Table I 1.2 are based on the responses of 700 members of APA's Division 17 (Counseling). The majority worked in university academic departments or counseling centers (56.3%); approximately 20% were in private practice, with the remainder employed in mental health clinics, medical settings, public schools, and consultation firms.

The top three tests are an interest inventory, a personality inventory, and an intelligence test, underscoring the previous point about the focus of counseling activities. A comparison of these ranks to the data from clinical psychologists in Table 15.1 reinforces the distinction between the two fields. In terms of personality testing, counseling psychologists are less likely lo use projective tests such as the Rorschach and the TAT. Furthermore, counseling psychologists are more likely lo use objective tests such as the Sixteen Personality Factor Questionnaire and the California Psychological Inventory.

Table 15.12 Tests Used Most Often by Counseling Psychologists in University and Private Practice Settings

Rank	Test
I	Strong-Campbell Interest Inventory
2	Minnesota Multiphasic Personality Inventory
3	Wechsler Adult Intelligence Scale
4	Sentence-completion blanks
5	Bender-Gestalt
6	Thematic Apperception Test
7	Sixteen Personality Factor Questionnaire
8	California Psychological Inventory
9	Wechsler Intelligence Scale for Children
10	Edwards Personal Preference Schedule

These are tests that focus on the structure of personality, rather than the identification of clinical syndromes.

The data in Table II.3 were generated by counselors in college and university counseling services and private counseling agencies. The individuals responding to the survey included not only counseling psychologists but also other types of professionals involved in counseling activities. Note that this group was even less likely to use personality tests, whether objective or projective, and focused primarily on assessment of interests and abilities.

A final point about the tests used in counseling settings: Most of these tests are paper-and-pencil, self-report tests that do not require administration by an examiner. Table 15.14 summarizes the design and administration requirements for the most popular tests.

How Tests Are Used in Counseling

Tests are used in counseling settings primarily to help psychologists better understand each client, to promote the client's self-awareness, as an aid to planning treatment, and to help client's make choices and solve problems (Goldman, 1971). The information that counseling psychologists most often seek through testing includes vocational and career-related information, evidence of possible psychopathology, and measures of clients' intellectual functioning (Fee, Elkins, & Boyd, 1982).

Table 15.13 Tests Used Most Often by Counselors in Counseling Services and Agencies

Rank	Test
l	Strong-Campbell Interest Inventory
2	Kuder Occupational Interest Survey
3	Edwards Personal Preference Schedule
4	Nelson-Denny Reading Test
5	Sixteen Personality Factor Questionnaire
6	Holland Self-Directed Search
7	Wechsler Adult Intelligence Scale
8	American College Testing Assessment
9	Survey of Habits and Attitudes
10	Minnesota Multiphasic Personality Inventory

Tests also can be used for research, such as comparison of the effectiveness of different therapeutic approaches.

One significant difference between the use of tests in counseling and clinical settings is the role of the client. In clinical settings, test results are not necessarily shared directly with clients. Instead, clinicians use these results to make decisions about the course of treatment. In counseling settings, test results are likely to be discussed with clients. Since a primary goal of counseling is to aid clients in developing self-awareness and making choices, the client is viewed as the primary user of test results (AERA/APA, 1985). The clinical psychologist can be seen as a gatekeeper, whereas the counseling psychologist is more like a facilitator (Wise, 1989).

Counseling psychologists also frequently use clinical interview and behavioral observation to learn more about their clients (Fee, Elkins, & Boyd, 1982). However, as in many other scenarios, tests provide a mechanism to obtain objective data using standardized procedures and to provide the individual difference information needed to assist client's with decision making. Counseling psychologists may also use a testing session as a standardized observational setting in which the behavior of clients can be noted and compared (Goodman, 1971). For example, the behavior of clients during an intelligence test can provide indications of their levels of self-confidence or anxiety or the extent to which they are reflective or impulsive.

Table 15.14 Categorizing the Design of Tests Frequently-Used in Counseling Settings

Purpose/Format	Examples
Aptitude Free response Norm referenced	Wechsler Adult Intelligence Scale Wechsler Intelligence Scale for Children
Personality Objective Norm referenced Normative	Minnesota Multiphasic Personality Inventory Sixteen Personality Factor Questionnaire California Psychological Inventor)'
Personality Objective Norm referenced Ipsative	Edwards Personal Preference Schedule
Personality Projective Norm referenced	Thematic Apperception Test Sentence-completion tests

Normative	
Interest Objective Norm referenced Normative	Strong-Campbell Interest Inventory3
Interest Objective Norm referenced Ipsative	Kuder Occupational Interest Survey

In contrast to clinical settings, professionals in counseling settings focus on the identification of abilities, personality characteristics, and interests to help people improve their sense of well-being and make life-style decisions. Although many types of professionals engage in counseling, a counseling psychologist is specifically trained both in psychological assessment and in a variety of treatment approaches.

The two primary activities in counseling settings are vocational or career counseling and personal counseling. Interest tests are an important component of career counseling; there are good tests available for both normative and ipsative assessment, for suggesting possible careers and college majors, and for individuals interested in technical or skilled jobs. Because success in a career involves both interest in that type of work and particular abilities, vocational counseling may also involve administration of ability tests. The ability tests most useful in career counseling include intelligence tests, achievement tests, entrance exams, and specific ability tests, the most popular of which are multiple aptitude test batteries.

In personal counseling, tests are used to explore the structure of personality both for increasing client self-awareness and for planning the course of treatment. In contrast to the tests used in clinical settings, these tests focus on typical personality dimensions rather than clinical syndromes. Some tests, such as the CPI and the 16PF, focus on the analysis of personality traits— particular patterns of responding to environmental events. For example, the 16PF identifies where individuals fall on a series of dimensions called source traits. Other tests, such as the Myers-Briggs classify individuals into groups known as personality types.

Scales that measure individual altitudes, personal qualities, or life events can complement or expand on the informal ion provided by personality tests. However, in some cases these scales are neither as reliable nor as valid as the tests used for overall personality assessment.