Tests and Measurements.

3rd basic year.

Educational psychology department.

English program.

2023/ 2024 academic year.

Table of content:-

Chapter 1	What are Psychological Tests?
Chapter 2	List of standardized tests.
Chapter 3	Meaning, Purpose & Construction of Achievement Tests.
Chapter 4	A Guide for Writing and Improving Achievement Tests.

What Are Psychological Tests?

CHAPTER 1: WHAT ARE PSYCHOLOGICAL TESTS?
After completing your study of this chapter, you should be able to do the following: • Define what a psychological test is and understand that psychological tests extend beyond personality and intelligence tests. • Trace the history of psychological testing from Alfred Binet and intelligence testing to the tests of today. • Describe the ways psychological tests can be similar to and different from one another. • Describe the three characteristics that are common to all psychological tests, and understand that psychological tests can demonstrate these characteristics to various degrees. • Describe the assumptions that must be made when using psychological tests. • Describe the different ways that psychological tests can be classified. • Describe the differences among four commonly used terms that students often get confused: psychological assessment, psychological tests, psychological measurement, and surveys. • Identify and locate print and online resources that are available for locating information about psychological tests.

"When I was in the second grade, my teacher recommended that I be placed in the school's gifted program. As a result, the school psychologist interviewed me and had me take an intelligence test."

"Last semester I took a class in abnormal psychology. The professor had all of us take several personality tests, including the MMPI [Minnesota Multiphasic Personality Inventory].

It was awesome! We learned about different types of psychological disorders that the MMPI can help diagnose."

"This year I applied for a summer job with a local bank. As a part of the selection process, I had to participate in a structured interview and an assessment center."

"Yesterday I took my driving test—both the written and the road test. I couldn't believe everything they made me do. I had to parallel park, switch lanes, and make both right and left turns."

I f your instructor asked whether you have ever taken a psychological test, you would probably report the intelligence test you took as an elementary school student or the personality test you took in your abnormal psychology class. If your instructor asked what the purpose of psychological testing is, you would probably say its purpose is to determine whether someone is gifted or has a psychological disorder. Intelligence tests and personality tests are indeed psychological tests—and they are indeed used to identify giftedness and diagnose psychological disorders. However, this is only a snapshot of what psychological testing is all about. There are many types of psychological tests, and they have many different purposes.

In this chapter, we introduce you to the concept of psychological testing. We discuss what a psychological test is and introduce some tests you might never have considered to be psychological tests. Then, after exploring the history of psychological testing, we discuss the three defining characteristics of psychological tests and the assumptions that must be made when using these tests. We then turn our attention to the many ways of classifying tests. We also distinguish four concepts that students often get confused: psychological assessment, psychological tests, psychological measurement, and surveys. We conclude this chapter by sharing with you some of the resources (print and online) that are available for locating information about psychological testing and specific psychological tests.

Why Should You Care About Psychological Testing?

Before discussing what a psychological test is, we would like you to understand just how important it is for you to understand the foundations of psychological testing. Psychological testing is not just another subject that you may study in college; rather, it is a topic that personally affects many individuals. Each day, psychological tests are administered by many different professionals to many different individuals, and the results of these tests are used in ways that significantly affect you and those around you. For example, test scores are used to diagnose mental disorders, to determine whether medicines should be prescribed (and, if so, which ones), to treat mental and emotional illnesses, to select individuals for jobs, to select individuals for undergraduate and professional schools (for example, medical school, law school), and to determine grades. Good tests facilitate high-quality decisions, and bad tests facilitate low-quality decisions.

The consequences of bad decisions can be significant. For example, a poor hiring decision can dramatically affect both the person being hired and the hiring organization. From the organization's perspective, a poor hiring decision can result in increased absenteeism, reduced morale of other staff, and lost productivity and revenue. From the employee's perspective, a poor hiring decision may result in a loss of motivation, increased stress leading to depression and anxiety, and perhaps loss of opportunity to

make progress in his or her career. Although you might never administer, score, or interpret a test, it is very likely that you or someone you know may have a life-altering decision made about him or her based on test scores. Therefore, it is important that you understand the foundations of psychological testing, specifically how to tell whether a decision is a good or bad one. Being able to do this requires that you understand the foundations of psychological testing.

What Are Psychological Tests?

Each anecdote at the beginning of this chapter involves the use of a psychological test. Intelligence tests, personality tests, interest and vocational inventories, college entrance exams, classroom tests, structured interviews, assessment centers, and driving tests all are psychological tests. Even the self-scored tests that you find in magazines such as *Glamour* and *Seventeen* (tests that supposedly tell you how you feel about your friends, stress, love, and more) can be considered psychological tests. Although some are more typical, all meet the definition of a psychological test. Together, they convey the very different purposes of psychological tests. For a continuum of some of the most and least commonly recognized types of psychological tests, see Figure 1.1.

Similarities Among Psychological Tests

Psychological testing is best defined as "the process of administering, scoring, and interpreting psychological tests" (Maloney & Ward, 1976, p. 9). But what exactly is a psychological test? We can easily answer this question by considering what all psychological tests do.

First, all **psychological tests** require a person to perform some **behavior**—an observable and measurable action. For example, when students take a multiple-choice midterm exam, they must read the various answers for each item and identify the best one. When individuals take an intelligence test, they may be asked to define words or solve math problems. When participating in a structured job interview, individuals must respond to questions from the interviewer—questions such as "Tell me about a time when you had to deal with an upset customer. What was the situation, what did you do, and what was the outcome?" In each of these cases, individuals are performing some observable and measurable behavior.

Figure 1.1	A Continuum of Psychological Tests

More Typical	←——————————————→		Less Typical
Personality tests Intelligence tests	Vocational tests Interest inventories Achievement tests Ability tests	Self-scored magazine tests Classroom quizzes and exams	Road portion of driving test Structured employment interviews Assessment centers

Second, the behavior an individual performs is used to measure some personal attribute, trait, or characteristic that is thought to be important in describing or understanding human behavior. For example, the questions on a multiple-choice exam might measure your knowledge of a particular subject area such as psychological testing. The words you defined or the math problems you solved might measure your verbal ability or quantitative reasoning. It is also important to note that sometimes the behavior an individual performs is also used to make a prediction about some outcome. For example, the questions you answered during a structured job interview may be used to predict your success in a management position.

So, what is a psychological test? It is something that requires you to perform a behavior to measure some personal attribute, trait, or characteristic or to predict an outcome.

Differences Among Psychological Tests

Although all psychological tests require that you perform some behavior to measure personal attributes, traits, or characteristics or to predict outcomes, these tests can differ in various ways. For example, they can differ in terms of the behavior they require you to perform, what they measure, their content, how they are administered and formatted, how they are scored and interpreted, and their psychometric quality (**psychometrics** is the quantitative and technical aspect of mental measurement).

Behavior Performed

The behaviors a test taker must perform vary by test. For example, a popular intelligence test, the Wechsler Adult Intelligence Scale–fourth edition (WAIS-IV), a general test of adult intelligence, requires test takers to (among other things) define words, repeat lists of digits, explain what is missing from pictures, and arrange blocks to duplicate geometric card designs. The Thematic Apperception Test (TAT), a widely used and researched projective personality test designed at Harvard University in the 1930s, requires test takers to look at ambiguous pictures showing a variety of social and interpersonal situations and to tell stories about each picture. The Graduate Record Examinations (GRE) General Test, a graduate school admissions test that measures verbal and quantitative reasoning, critical thinking, and analytical writing skills, requires test takers to answer multiple-choice questions and respond to two analytical writing tasks. The road portion of an auto driving test typically requires test takers to do things such as start a car, change lanes, make right and left turns, use turn signals properly, and parallel park. Assessment centers require job applicants to participate in simulated job-related activities (that mimic the activities they would perform in the job) such as engaging in confrontational meetings with disgruntled employees, processing e-mail and paperwork, and conducting manager briefings.

Attribute Measured and Outcome Predicted

What a test measures or predicts can vary. For example, the WAIS-IV asks individuals to explain what is missing from pictures to measure verbal intelligence. The TAT requires individuals to tell stories about pictures to identify conscious and unconscious drives, emotions, conflicts, and so on in order to ultimately measure personality. The road portion of a driving test requires individuals to perform various driving behaviors to measure driving ability. The GRE requires students to answer different types of questions to predict success in graduate school.

Some of the characteristics, attributes, and traits commonly measured by psychological tests include personality, intelligence, motivation, mechanical ability, vocational preference, spatial ability, and anxiety. Some of the outcomes that tests typically predict include success in college, worker productivity, and who will benefit from specialized services such as clinical treatment programs.

> More detail about the WAIS-IV can be found in Test Spotlight 1.1 in Appendix A. More detail about the GRE can be found in Test Spotlight 13.1 in Appendix A.

Content

Two tests that measure the same characteristic, attribute, or trait can require individuals to perform significantly different behaviors or to answer significantly different questions. Sometimes how the test developers define the particular characteristic, attribute, or trait affects how the test is structured. For example, the questions on two intelligence tests may differ because one author may define intelligence as the ability to reason and another author may define it in terms of **emotional intelligence**—one's ability to understand one's own feelings and the feelings of others and to manage one's emotions (Gibbs, 1995).

The difference in content may also be due to the theoretical orientation of the test. (We talk more about theoretical orientation and its relation to test content in Chapter 8.)

Administration and Format

Psychological tests can differ in terms of how they are administered and their format. A test can be administered in paper-and-pencil format (individually or in a group setting), on a computer, or verbally. Similarly, a psychological test may consist of multiple-choice items, agree/disagree items, true/false items, open-ended questions, or some mix of these. There are also tests that ask respondents to perform some behavior such as sorting cards, playing a role, or writing an essay.

Scoring and Interpretation

Psychological tests can differ in terms of how they are scored and interpreted. Some tests are completed on scannable sheets and are computer scored. Some are hand-scored by the person administering the test. Others are scored by the test takers themselves. In terms of interpretation, some tests generate results that can be interpreted easily by the test taker, and others require a knowledgeable professional to explain the results to the test taker.

Psychometric Quality

Last, but extremely important, psychological tests can differ in terms of their psychometric quality. For now, let us just say that there are a lot of really good tests out there that measure what they say they measure and do so consistently, but there are also a lot of really poor tests out there that do not measure what they say they measure. Good tests measure what they claim to measure, and any conclusions that are drawn from the test scores about the person taking the test are appropriate (they are what we call *valid*). Good tests also measure whatever they measure consistently (they are what we call *reliable*). The concepts of reliability and validity are central to determining

whether a test is "good" or "bad" and are covered in detail later in this textbook. These concepts are so important that four chapters are devoted to them (Chapter 6 covers reliability, and Chapters 7–9 discuss validity).

Because tests can differ in so many ways, to make informed decisions about tests, you must know how to properly critique a test. A critique of a test is an analysis of the test. A good critique answers many of the questions in Table 1.1. (These questions are also in Appendix B.) Your instructor may have additional ideas about what constitutes a good critique.

**INTERIM SUMMARY 1.1
SIMILARITIES AND DIFFERENCES AMONG PSYCHOLOGICAL TESTS**

Similarities

- All psychological tests require an individual to perform a behavior.
- The behavior performed is used to measure some personal attribute, trait, or characteristic.
- This personal attribute, trait, or characteristic is thought to be important in describing or understanding behavior.
- The behavior performed may also be used to predict outcomes.

Differences

Psychological tests can differ in terms of the following:

- The behavior they require the test taker to perform
- The attribute they measure
- Their content
- How they are administered and formatted
- How they are scored and interpreted
- Their psychometric quality

The History of Psychological Testing

Some scholars believe that the use of psychological tests can be traced to 2200 BCE in ancient China. For a summary of this history, see For Your Information Box 1.1. Most scholars agree that serious research efforts on the use and usefulness of psychological tests did not begin until the 20th century with the advent of intelligence testing.

Intelligence Tests

Alfred Binet and the Binet–Simon Scale

Late in the 19th century, Alfred Binet founded the first experimental psychology research laboratory in France. In his lab, Binet attempted to develop experimental techniques to measure intelligence and reasoning ability. He believed that intelligence was a complex characteristic that could be determined by evaluating a person's reasoning, judgment, and problem-solving abilities. Binet tried a variety of tasks to measure reasoning, judgment, and problem solving on his own children as well as on other children in the French school system.

Binet was successful in measuring intelligence, and in 1905 he and Théodore Simon published the first test of mental ability, the Binet–Simon Scale. Parisian school officials used this scale to decide which children, no matter how hard they tried, were unable to profit from regular school programs (Binet & Simon, 1905).

Table 1.1	Guidelines for Critiquing a Psychological Test

General descriptive information

- What is the title of the test?
- Who is the author of the test?
- Who publishes the test, and when was it published? (Include dates of manuals, norms, and supplementary materials)
- How long does it take to administer the test?
- How much does it cost to purchase the test? (Include the cost of the test, answer sheets, manual, scoring services, and so on)
- Is the test proprietary or nonproprietary?

Purpose and nature of the test

- What does the test measure? (Include scales)
- What does the test predict?
- What behavior does the test require the test taker to perform?
- What population was the test designed for (for example, age, type of person)?
- What is the nature of the test (for example, maximal performance, behavior observation, self-report, standardized or nonstandardized, objective or subjective)?
- What is the format of the test (for example, paper-and-pencil or computer, multiple choice or true/false)?

Practical evaluation

- Is the test manual comprehensive (does it include information on how the test was constructed, its reliability and validity, composition of norm groups, whether it is easy to read)?
- Is the test easy or difficult to administer?
- How clear are the administration directions?
- How clear are the scoring procedures?
- What qualifications and training does a test administrator need to have?
- Does the test have face validity?

Technical evaluation

- Is there a norm group?
- Who comprises the norm group?
- What types of norms are there (for example, percentiles, standard scores)?
- How was the norm group selected?
- Are there subgroup norms (for example, by age, gender, region, occupation, and so on)?
- What is the estimate of the test's reliability?
- How was reliability determined?
- What is the evidence for the validity of the test?
- How was the evidence for validity gathered?
- What is the standard error of measurement?
- What are the confidence intervals?

Test reviews

- What do reviewers say are the strengths and weaknesses of the test?
- What studies that use the test as a measurement instrument have been published in peer-reviewed journals?
- How did the test perform when researchers or test users, other than the test developer or publisher, used it?

Summary

- Overall, what do you see as being the strengths and weaknesses of the test?

FYI

FOR YOUR INFORMATION BOX 1.1

Psychological Tests: From Ancient China to the 20th Century

2200 BCE: Xia Dynasty

Some scholars believe that the use of psychological tests dates back approximately 4,000 years to 2200 BCE, when the Chinese emperor Yushun examined officials every third year to determine whether they were suitable to continue in office (DuBois, 1970; Martin, 1870). However, modern scholars of ancient China say that there is little archaeological evidence to support these claims. Reliable writing systems were developed by the Chinese somewhere between 1766 and 1122 BCE (Shang dynasty; Bowman, 1989). Nowhere in these writings were there any hints suggesting that leaders were examined as just described. Even in 1115 BCE, with the advent of more elaborate writing systems, there were no inscriptions or writings to suggest the existence of such an examination process (Martin, 1870).

200–100 BCE: Late Qin, Early Han Dynasty

Most modern scholars of ancient China agree that royal examinations began around 200 to 100 BCE, in the late Qin (Ch'in) or early Han dynasty (Eberhard, 1977; Franke, 1960; Pirazzoli-t'Serstevens, 1982; Rodzinski, 1979). Hucker (1978) believes that the first written examinations in world history began in 165 BCE, when the emperor administered written examinations to all nominees. Pirazzoli-t'Serstevens also believes that this was the beginning of all examination systems. Eberhard, on the other hand, admits that there may have been some assessment procedures before 165 BCE for selecting officials, who were probably tested more for literacy than for knowledge.

618–907 CE: T'ang Dynasty

Such examination systems seem to have been discontinued until the T'ang dynasty, when their use increased significantly (Bowman, 1989).

1368–1644: Ming Dynasty

During the Ming dynasty, the examinations became more formal. There were different levels of examinations (municipal, county, provincial, and national), and the results of examinations became associated with granting formal titles, similar to today's university degrees. On passing each level of examination, people received more titles and increasingly more power in the civil service (Bowman, 1989). These examinations were distressful, and this distress became a part of Chinese culture and also a part of folk stories and the literature (poems, comedies, and tragedies). Nonetheless, this examination system seemed to work well. Today, many scholars believe that this examination system kept talented men in the national government (Kracke, 1963) and kept members of the national government from becoming nobility because of their descent.

Seeing the value of these examinations for making important decisions, European governments, and eventually the governments of the United Kingdom, the United States, Canada, and other countries, adopted the use of such examination systems.

1791: France and Britain

France initially began using this kind of examination system in 1791. However, soon after, Napoleon temporarily abolished them. The system adopted by France served as a model for a British system started in 1833 to select trainees for the Indian civil service—the beginning of the British civil service.

1860s: United States

Due to the success of the British system, Senator Charles Sumner and Representative Thomas Jenckes proposed to Congress in 1860 that the United States use a similar system. Jenckes's report, *Civil Service in the United States,* described the British and Chinese systems in detail. This report laid the foundation for the establishment of the Civil Service Act Health and Psychosocial Instruments (HAPI), passed in January 1883.

20th Century: Western Europe and the United States

In 1879, Wilhelm Wundt introduced the first psychological laboratory, in Leipzig, Germany. At this time, psychology was the study of the similarities among people. For example, physiological psychologists studied how the brain and the nervous system function, and experimental psychologists conducted research to discover how people learn and remember. Strongly influenced by James McKeen Cattell, an American researcher in Wundt's laboratory, psychologists turned their attention to exploring individual differences. Cattell and others realized that learning about the differences among people was just as important as learning about the similarities among people. They believed that developing formal psychological tests to measure individual differences could help solve many social problems, such as who should be placed in remedial programs, who should be sent to battlefields, and who should be hired for particular jobs. At this time, scientists were particularly interested in finding a quantitative way of measuring general intelligence.

During the early 20th century, serious research efforts began on the use and usefulness of various testing procedures. Research conducted by scholars in the United States and Germany eventually led to Alfred Binet's research on intelligence in children.

Lewis Terman and the Stanford–Binet

Binet's work influenced psychologists across the globe. Psychological testing became a popular method of evaluation, and the Binet–Simon Scale was adapted for use in many countries. In 1916, Lewis Terman, an American psychologist, produced the

> More detail about the Stanford-Binet Intelligence Scales can be found in Test Spotlight 1.2 in Appendix A.

Stanford–Binet Intelligence Scales, an adaptation of Binet's original test. This test, developed for use with Americans ages 3 years to adulthood, was used for many years. A revised edition of the Stanford–Binet remains one of the most widely used intelligence tests today.

The Wechsler–Bellevue Intelligence Scale and the Wechsler Adult Intelligence Scale

By the 1930s, thousands of psychological tests were available, and psychologists and others were debating the nature of intelligence (what intelligence was all about). This dispute over defining intelligence prompted the development in 1939 of the original Wechsler–Bellevue Intelligence Scale (WBIS) for adults, which provided an index of general mental ability (as did the Binet–Simon Scale) and revealed patterns of a person's intellectual strengths and weaknesses. David Wechsler, the chief psychologist at Bellevue Hospital in New York City, constructed the WBIS believing that intelligence is demonstrated based on an individual's ability to act purposefully, think logically, and interact/cope successfully with the environment (Hess, 2001; Rogers, 2001; Thorne & Henley, 2001). Wechsler published the second edition, the WBIS-II, in 1946.

> More detail about the fourth edition of the WAIS-IV can be found in Test Spotlight 1.1 in Appendix A.

In 1955, Wechsler revised the WBIS-II and renamed it the Wechsler Adult Intelligence Scale (WAIS). In 1981 and 1991 the WAIS was updated and published as the WAIS-R and WAIS-III, respectively. In a

continuing effort to improve the measurement of intelligence, as well as the clinical utility and user-friendliness of the test, the fourth edition was published in 2008 (Pearson Education, 2009).

Personality Tests

In addition to intelligence testing, the early 1900s brought about an interest in measuring personality.

The Personal Data Sheet

During World War I, the U.S. military wanted a test to help detect soldiers who would not be able to handle the stress associated with combat. To meet this need, the American Psychological Association (APA) commissioned an American psychologist, Robert Woodworth, to design such a test, which came to be known as the Personal Data Sheet (PDS). The PDS was a paper-and-pencil psychiatric interview that required military recruits to respond *yes* or *no* to a series of 200 questions (eventually reduced to 116 questions) that searched for mental disorders. The questions covered topics such as excessive anxiety, depression, abnormal fears, impulse problems, sleepwalking, nightmares, and memory problems (Segal & Coolidge, 2004). One question asked, "Are you troubled with the idea that people are watching you on the street?" (cited in Cohen, Swerdlik, & Phillips, 1996). During a pilot study of the test, new recruits on average showed 10 positive psychoneurotic symptoms; recruits who were deemed unfit for service generally showed 30 to 40 positive psychoneurotic symptoms (Segal & Coolidge, 2004). Unfortunately, because Woodworth did not complete the final design of this test until too late in the war, the PDS was never implemented or used to screen new recruits.

After World War I, Woodworth developed the Woodworth Psychoneurotic Inventory, a version of the PDS. Unlike the PDS, the Woodworth Psychoneurotic Inventory was designed for use with civilians and was the first self-report test. It was also the first widely used personality inventory.

The Rorschach Inkblot Test and the TAT

During the 1930s, interest also grew in measuring personality by exploring the unconscious. With this interest came the development of two important projective tests: the Rorschach Inkblot Test and the TAT. The Rorschach, a projective personality test (described further in Chapters 2 and 14), was developed by Swiss psychiatrist Hermann Rorschach. The TAT, also a projective personality test, was developed by two American psychologists, Henry A. Murray and C. D. Morgan. Both tests are based on the personality theories of Carl Jung and continue to be widely used today for personality assessment.

Vocational Tests

During the 1940s, a need developed for **vocational tests** to help predict how successful an applicant would be in specific occupations. The Public Employment Services needed such tests because thousands of people had lost their jobs due to the Great Depression and thousands more were coming out of school and seeking work. Because there were not enough jobs, people were forced to look for new lines of work. As a result, psychologists developed large-scale programs to design vocational aptitude tests that would predict how successful a person would be at an occupation before entering it. In 1947, the Department

of Labor developed the General Aptitude Test Battery (GATB) to meet this need. The GATB was used for a variety of purposes, including vocational counseling and occupational selection.

By the mid-20th century, numerous tests were available and they were used by many to make important decisions about individuals. (We talk more about these decisions in Chapters 2 and 8.) Because of the increased use of psychological tests, to help protect the rights of the test taker, the APA (1953) published *Ethical Standards of Psychologists.* (We discuss these ethical standards in more detail in Chapter 3.)

Testing Today

In the 21st century, psychological testing is a big business. There are thousands of commercially available, standardized psychological tests as well as thousands of unpublished tests. Tests are published by hundreds of test publishing companies that market their tests very proactively—on the web and in catalogs. Before the turn of this century, these publishers were earning close to $200 million per year (Educational Testing Service, 1996), and approximately 20 million Americans per year were taking psychological tests (Hunt, 1993). For the names and web addresses of some of the most well-known test publishers, as well as some of the most popular tests they publish, see On the Web Box 1.1. Publishing and marketing companies are capitalizing on the testing trend, creating and marketing a bonanza of new products and study aids. To read about some of these products and study aids, see In the News Box 1.1.

Today, psychological testing is a part of the American culture. Psychological tests are in use everywhere. For example, let us take a look at Sylvan Learning Center (SLC), a provider of personal instructional services to children from kindergarten through 12th grade that has more than 1,100 centers worldwide. You might be familiar with SLC because of the test preparation programs they offer (for example, preparation for the SAT). However, did you know that much of SLC's business is focused on personalized programs to help children develop skills in areas such as reading, math, and writing? These personalized programs are created by administering and combining the results of standardized tests to capture a student's academic strengths and weaknesses and to identify skill gaps (Sylvan Learning, 2010). SLC uses identified skill gaps, often the reason for underperformance in school, to create a blueprint for an individual child's unique tutoring program. SLC also administers learning style inventories to help instructors understand how each child learns best. Trained and certified instructors integrate these learning styles into their tutoring sessions to promote individual student learning (SLC, 2010).

Now let us take a look at the Society for Human Resources Management (SHRM). As the world's largest association devoted to human resources management, SHRM provides human resources professionals with essential information and resources (SHRM, 2010a). One of these resources is an online testing center, which provides SHRM members who are qualified testing professionals with electronic access to more than 400 tests, from over 50 test publishers, in areas such as personality and skills assessment, coaching and leadership, mechanical and technical skills, information technology skills, pre-employment screening, and career exploration (SHRM, 2010b). The testing center allows qualified testing professionals to purchase individual tests, administer the tests online, and receive electronic reports.

ON THE WEB BOX 1.1

Names and Web Addresses of Test Publishers

Open your web browser, go to your favorite search engine, and conduct a search for "test publishers" or "psychological test publishers." You will find pages and pages of websites dedicated to psychological testing and publishing. You will also find the websites of hundreds of test publishers. Although there are many different publishers, some of the most well-known, including some of the widely known tests they publish, are listed here:

Publisher	Website	Popular Published Tests
Educational Testing Service	www.ets.org	• Advanced Placement (AP) Program Tests • Graduate Management Admission Test (GMAT) • Graduate Record Examinations (GRE) • Scholastic Assessment Test (SAT) • Test of English as a Foreign Language (TOEFL)
Pearson	www.pearsonassessments.com	• BarOn Emotional Quotient Inventory • Bayley Scales of Infant and Toddler Development—III • Bender Visual-Motor Gestalt Test—II • Watson–Glaser Critical Thinking Appraisal
Hogan Assessment Systems	www.hoganassessments.com	• Hogan Personality Inventory (HPI) • Hogan Development Survey (HDS) • Hogan Business Reasoning Inventory (HBRI) • Motives, Values, Preferences Inventory (MVPI)
IPAT	www.ipat.com	• 16 Personality Factors (16PF)
PAR	www3.parinc.com	• Self-Directed Search • NEO Personality Inventory • Personality Assessment Inventory • Slosson Intelligence Test—Revised for Children and Adults
Psytech International	www.psytech.co.uk	• Occupational Interest Profile • Clerical Test Battery • Values and Motives Inventory
PSI	www.psionline.com	• Customer Service Battery • Firefighter Selection Test • Police Selection Test
Hogrefe	www.testagency.com	• Rorschach Inkblot Test • Trauma Symptom Inventory (TSI) • WPQ Emotional Intelligence Questionnaire
University of Minnesota Press Test Division	www.upress.umn.edu/tests/default.html	• Minnesota Multiphasic Personality Inventory (MMPI)
Wonderlic	www.wonderlic.com	• Wonderlic Personnel Test

IN THE NEWS

Box 1.1   SAT Prep Tools: From Cellphones to Handhelds to CDs

Early in 2005, the College Board introduced thousands of high school juniors to the new SAT. No longer containing the much-dreaded analogy questions, the new SAT is longer and more difficult and, for the first time, contains a writing section (College Board, 2010). The writing section contains multiple-choice questions that assess how well test takers use standard written English language and a handwritten essay to assess how well they can develop a point of view on a topic.

   Not wanting to miss an opportunity, publishers capitalized on the updated SAT by creating and marketing a number of new and innovative products—products promising to appeal to today's technology-savvy, music-hungry, multitasking teens. In 2005, *The Wall Street Journal* published an article introducing some of these unique products. In 2009, the products are still being marketed to students preparing to take the SAT.

Princeton Review: This company offers private instruction and tutoring for standardized achievement tests. In partnership with Cocel, Princeton Review has developed a new software program called Prep for the SAT, which beams SAT practice questions, including reading passages, to cell phones so that students can prepare for the SAT at their convenience. Answers are quickly graded, and parents can even receive electronic reports. In 2005, Princeton Review also released Pocket Prep, an interactive, portable, handheld SAT prep device designed to help 21st-century high school students prepare for the SAT using a format and technology that suits their lifestyles and preferences. Pocket Prep features information about the new SAT; comprehensive verbal, math, and essay preparation; full-length timed practice exams; instant scoring; and personal diagnostic reports. It also includes practice drills, flash cards, and an extensive verbal and essay reference suite to help students maximize their grammar and essay scores.

   Kaplan: Another test preparation company, Kaplan has designed software for cell phones and handheld devices and is publishing books, such as *Frankenstein* and *Wuthering Heights,* that contain SAT vocabulary words in bold print as well as their definitions. An example of a sentence containing an SAT vocabulary word (*desolation*) might be "Mr. Heathcliff and I are such a suitable pair to divide the desolation between us."

   SparkNotes: This Internet-based, youth-oriented education product (owned by Barnes & Noble) has published several Spuzzles books containing crossword puzzles in which the answers are commonly occurring SAT vocabulary words. For example, in U.S. History Spuzzle No. 56, the clue to 8 Across is "English Quaker who founded Pennsylvania in 1681."

   Wiley Publishing: A well-known publisher of print and electronic products, Wiley has published a teen novel, *The Marino Mission: One Girl, One Mission, One Thousand Words,* that contains 1,000 need-to-know SAT vocabulary words. Not only are vocabulary words defined at the bottom of each page, but there also are self-tests at the end of the novel to help readers retain what they have learned.

   Defined Mind—These independent recording artists, along with Kaplan, have produced *Vocabulary Accelerator,* a 12-track CD full of rock, folk-funk, and techno beats. What is unique is that the lyrics are studded with SAT vocabulary.

SOURCE: Kronhold, J. (2005, March 8). To tackle the new SAT, perhaps you need a new study device. *The Wall Street Journal.* Retrieved May 20, 2010, from http://www.tilcoweb.com/wallstreetjournal01.htm

One of the most significant and controversial uses of psychological testing in the 21st century has been a result of the No Child Left Behind Act of 2001 (NCLB Act). The NCLB Act, which President George W. Bush signed into law on January 8, 2002, was intended to improve the performance of America's primary and secondary schools. The NCLB Act contains the following four basic strategies for improving the performance of schools—strategies that were intended to change the culture of America's schools by defining a school's success in terms of the achievement of its students (U.S. Department of Education, 2004):

1. Increase the accountability that states, school districts, and schools have for educating America's children by requiring that all states implement statewide systems that (a) set challenging standards for what children in Grades 3 to 8 should know and learn in reading and math, (b) test students in Grades 3 to 8 on a yearly basis to determine the extent to which they know and have learned what they should have according to state standards, and (c) include annual statewide progress objectives to ensure that all students are proficient by the 12th grade.

2. Ensure that all children have access to a quality education by allowing parents to send their children to better schools if their schools do not meet state standards.

3. Increase the amount of flexibility that high-performing states and school districts have for spending federal education dollars.

4. Place more emphasis on developing children's reading skills by making grants available to states to administer screening and diagnostic assessments to identify children who may be at risk for reading failure and by providing teachers with professional development and resources to help young children attain the knowledge and skills they need to be readers.

Years after the implementation of the NCLB Act, there remains significant controversy, some of which focuses on the overreliance on test scores that may "distort teaching and learning in unproductive ways" (Center for Public Education, 2006, para. 6). While tests have always played a critical role in the assessment of student achievement, the NCLB Act requires that students be tested more often and relies on test scores to make more important decisions than in the past. In Chapter 13, we talk more about how one state, Florida, has responded to the NCLB Act, focusing primarily on the role that psychological tests have played in assessing the extent to which children and schools measure up to state standards.

The Defining Characteristics of Psychological Tests

As we have already discussed, a psychological test is anything that requires an individual to perform a behavior for the purpose of measuring some attribute, trait, or characteristic or to predict an outcome. All good psychological tests have three characteristics in common:

1. They representatively sample the behaviors thought to measure an attribute or thought to predict an outcome. For example, suppose we are interested in developing a test to measure your physical ability. One option would be to evaluate your performance in every sport you have ever played. Another option would be to have you run the 50-meter dash. Both of these options have drawbacks. The first option would be very precise, but not very practical. Can you imagine how much time and energy it would take to review how you performed in every sport you have ever played? The second option is too narrow and unrepresentative. How fast you run the 50-meter dash does not tell us much about your physical ability in general. A better method would be to take a representative sample of performance in sports. For example, we might require you to participate in some individual sports (for example, running, tennis, gymnastics) and team sports (for example, soccer, basketball) that involve different types of physical abilities (for example, strength, endurance, precision). This option would include a more representative sample.

2. All good psychological tests include behavior samples that are obtained under standardized conditions. That is, a test must be administered the same way to all people. When you take a test, various factors can affect your score besides the characteristic, attribute, or trait that is being measured. Factors related to the environment (for example, room temperature, lighting), the examiner (for example, examiner attitude, how the instructions are read), the examinee (for example, disease, fatigue), and the test (for example, understandability of questions) all can affect your score. If everyone is tested under the same conditions (for example, the same environment), we can be more confident that these factors will affect all test takers similarly. If all of these factors affect test takers similarly, we can be more certain that a person's test score accurately reflects the attribute being measured. Although it is possible for test developers to standardize factors related to the environment, the examiner, and the test, it is difficult to standardize examinee factors. For example, test developers have little control over what test takers do the night before they take a test.

3. All good psychological tests have rules for scoring. These rules ensure that all examiners will score the same set of responses in the same way. For example, teachers might award 1 point for each multiple-choice question you answer correctly, and they might award or deduct points based on what you include in your response to an essay question. Teachers might then report your overall exam score either as the number correct or as a percentage of the number correct (the number of correct answers divided by the total number of questions on the test).

Although all psychological tests have these characteristics, not all exhibit these characteristics to the same degree. For example, some tests may include a more representative sample of behavior than do others. Some tests, such as group-administered tests, may be more conducive to administration under standardized conditions than are individually administered tests. Some tests may have well-defined rules for scoring, and others might have general guidelines. Some tests may have very explicit scoring rules, for example, "If Question 1 is marked true, then deduct 2 points." Other tests, such as those that include short answers, may have less explicit rules for scoring, for example, "Award 1 point for each concept noted and defined."

INTERIM SUMMARY 1.2
THE THREE DEFINING CHARACTERISTICS OF PSYCHOLOGICAL TESTS

All psychological tests have three common characteristics:

- First, a good test should representatively sample the behaviors thought to measure an attribute or predict an outcome. This ensures that the test measures what it says it measures.
- Second, the behavior samples should be obtained under standardized conditions. That

is, a test must be administered exactly the same way to all individuals so that we can be confident that a person's score accurately reflects the attribute being measured or the outcome being predicted.
- Third, there must be rules for scoring so that all examiners will score the test in the same way.

Assumptions of Psychological Tests

There are many assumptions that must be made when using psychological tests. The following are what we consider the most important assumptions:

1. *Psychological tests measure what they purport to measure or predict what they are intended to predict.* In addition, any conclusions or inferences that are drawn about the test takers based on their test scores must be appropriate. This is also called test validity. If a test is designed to measure mechanical ability, we must assume that it does indeed measure mechanical ability. If a test is designed to predict performance on the job, then we must assume that it does indeed predict performance. This assumption must come from a personal review of the test's validity data.

2. *An individual's behavior, and therefore test scores, will typically remain stable over time.* This is also called test–retest reliability. If a test is administered at a specific point in time and then we administer it again at a different point in time (for example, two weeks later), we must assume, depending on what we are measuring, that an individual will receive a similar score at both points in time. If we are measuring a relatively stable trait, we should be much more concerned about this assumption. However, there are some traits, such as mood, that are not expected to show high test–retest reliability.

3. *Individuals understand test items similarly* (Wiggins, 1973). For example, when asked to respond *true* or *false* to a test item such as "I am almost always healthy," we must assume that all test takers interpret "almost always" similarly.

4. *Individuals will report accurately about themselves* (for example, about their personalities, about their likes and dislikes; Wiggins, 1973). When we ask people to remember something or to tell us how they feel about something, we must assume that they will remember accurately and that they have the ability to assess and report accurately on their thoughts and feelings. For example, if we ask you to tell us whether you agree or disagree with the statement "I have always

liked cats," you must remember not only how you feel about cats now but also how you felt about cats previously.

5. *Individuals will report their thoughts and feelings honestly* (Wiggins, 1973). Even if people are able to report correctly about themselves, they may choose not to do so. Sometimes people respond how they think the tester wants them to respond, or they lie so that the outcome benefits them. For example, if we ask test takers whether they have ever taken a vacation, they may tell us that they have even if they really have not. Why? Because we expect most individuals to occasionally take vacations, and therefore the test takers think we would expect most individuals to answer *yes* to this question. Criminals may respond to test questions in a way that makes them appear neurotic or psychotic so that they can claim they were insane when they committed crimes. When people report about themselves, we must assume that they will report their thoughts and feelings honestly, or we must build validity checks into the test.

6. *The test score an individual receives is equal to his or her true ability plus some error, and this error may be attributable to the test itself, the examiner, the examinee, or the environment.* That is, a test taker's score may reflect not only the attribute being measured but also things such as awkward question wording, errors in administration of the test, examinee fatigue, and the temperature of the room in which the test was taken. When evaluating an individual's score, we must assume that it will include some error.

Although we must accept some of these assumptions at face value, we can increase our confidence in others by following certain steps during test development. For example, in Section III of this textbook, which covers test construction, we talk about how to design test questions that are understood universally. We also talk about the techniques that are available to promote honest answering. In Section II, which covers psychometric principles, we discuss how to measure a test's reliability and validity.

Test Classification Methods

As we have already discussed, there are tens of thousands of commercially available psychological tests, and professionals refer to these tests in various ways. Sometimes professionals refer to them as tests of maximal performance, behavior observation tests, or self-report tests. Sometimes professionals refer to tests as being standardized or nonstandardized, objective or projective. Other times professionals refer to tests based on what the tests measure. In this section, we discuss the most common ways that professionals classify and refer to psychological tests.

Maximal Performance, Behavior Observation, or Self-Report

Most psychological tests can be defined as being tests of maximal performance, behavioral observation tests, or self-report tests.

- **Tests of maximal performance** require test takers to perform a particular well-defined task such as making a right-hand turn, arranging blocks from smallest to largest, tracing a pattern, or completing mathematical problems. Test takers try to do their best because their scores are determined by their success in completing the task. Intelligence tests, tests of specific abilities (for example, mechanical ability), driving tests (road and written), and classroom tests all are good examples of tests of maximal performance.

- **Behavior observation tests** involve observing people's behavior and how people typically respond in a particular context. Unlike with tests of maximal performance, many times people do not know that their behavior is being observed and there is no single defined task for the individual to perform. Many restaurants use this technique to assess food servers' competence in dealing with customers. Sometimes managers hire trained observers to visit their restaurant disguised as a typical customer. In exchange for a free meal or some predetermined compensation, observers agree to record specific behaviors performed by a food server. For example, observers may document whether a food server greeted them in a friendly manner. Other examples of behavior observations include documenting job performance for performance appraisals or clinical interviews.

- **Self-report tests** require test takers to report or describe their feelings, beliefs, opinions, or mental states. Many personality inventories, such as the Hogan Personality Inventory (HPI), are self-report tests. The HPI, a test used primarily for personnel selection and individualized assessment, asks test takers to indicate whether each of more than 200 statements about themselves is true or false.

Most psychological tests fit one of the above categories, and some tests contain features of more than one category. For example, a structured job interview (which involves asking all job applicants a standard set of interview questions) could include both technical questions and questions about one's beliefs or opinions. Technical questions, which are well defined for the interviewee, qualify the interview as a test of maximal performance. Questions about beliefs and opinions qualify it as a self-report test. The interviewer may also observe the interviewees' behaviors, such as their greetings, which would qualify the interview as a behavioral observation.

Standardized or Nonstandardized

Standardized tests are those that have been administered to a large group of individuals who are similar to the group for whom the test has been designed. For example, if a test is designed to measure the writing ability of high school students, the test would be administered to a large group of high school students. This group is called the **standardization sample**—people who are tested to obtain data to establish a frame of reference for interpreting individual test scores. These data, called **norms,** indicate the average performance of a group and the distribution of scores above and below this average.

For example, if you took the SAT, the interpretation of your score included comparing it with the SAT standardization sample to determine whether your score was high or low in comparison with others and whether you scored above average, average, or below average. In addition, standardized tests always have specific directions for administration and scoring.

Nonstandardized tests do not have standardization samples and are more common than standardized tests. Nonstandardized tests are usually constructed by a teacher or trainer in a less formal manner for a single administration. For example, in many cases, the exams you take in your college courses are nonstandardized tests.

Objective or Projective

Sometimes people make a distinction between objective and projective tests. **Objective tests** are structured and require test takers to respond to structured true/false questions, multiple-choice questions, or rating scales. What the test taker must do is clear, for example, answer *true* or *false,* circle the correct multiple-choice answer, or circle the correct item on the rating scale. The GRE, Stanford-Binet Intelligence Scales, General Aptitude Test Battery, and most classroom tests are examples of objective tests.

> More detail about the GRE can be found in Test Spotlight 13.1 in Appendix A. More detail about the Stanford–Binet Intelligence Scales can be found in Test Spotlight 1.2 in Appendix A.

Another example of an objective test is the NEO Personality Inventory, an objective self-report instrument designed to identify what makes individuals unique in their thinking, feeling, and interaction with others. Although there are two forms of the inventory, both measure five broad personality dimensions: neuroticism, extroversion, openness, agreeableness, and conscientiousness. Test takers are asked to indicate whether they strongly disagree, disagree, are neutral, agree, or strongly agree with each of 240 statements. These statements are about their thoughts, feelings, and goals. For sample questions from the NEO Personality Inventory, see For Your Information Box 1.2.

On the other hand, **projective tests** are unstructured. They require test takers to respond to unstructured or ambiguous stimuli such as incomplete sentences, inkblots, and abstract pictures. The role of the test taker is less clear than with a standardized test. People who use projective tests believe that test takers project themselves into the task they are asked to perform and that their responses are based on what they believe the stimuli mean and on the feelings they experience while responding. These tests tend to elicit highly personal concerns. They are often used to detect

> More detail about the NEO Personality Inventory can be found in Test Spotlight 1.3 in Appendix A.

unconscious thoughts or personality characteristics, and they may be used to identify the need for psychological counseling. The TAT is an example of a projective test. (Chapter 14 contains more information on the TAT and other projective tests.)

Dimension Measured

Psychological tests are often discussed in terms of the dimensions they measure. For example, sometimes we distinguish among achievement tests, aptitude tests, intelligence tests, personality tests, and interest inventories. We refer to these as dimensions because they are broader than a single attribute or trait level. Often these types of tests measure various personal attributes or traits.

Achievement Tests

Achievement tests measure a person's previous learning in a specific academic area (for example, computer programming, German, trigonometry, psychology). A test that requires you to list the three characteristics of psychological tests would be considered an achievement test. Achievement tests are also referred to as tests of knowledge.

Achievement tests are used primarily in educational settings to determine how much students have learned or what they can do at a particular point in time. Many elementary schools and high schools rely

FOR YOUR INFORMATION BOX 1.2

Sample Items From the NEO Personality Inventory

The NEO Personality Inventory is an objective self-report instrument designed to identify what makes individuals unique in their thinking, feeling, and interaction with others. The inventory measures five broad personality dimensions: neuroticism, extroversion, openness, agreeableness, and conscientiousness. Test takers are asked to indicate whether they strongly disagree (SD), disagree (D), are neutral (N), agree (A), or strongly agree (SA) with each of 240 statements. These statements are about their thoughts, feelings, and goals. In the following, we list a sample item from three of the five scales:

Neuroticism

Frightening thoughts sometimes come into my head.	SD	D	N	A	SA

Extroversion

I don't get much pleasure from chatting with people.	SD	D	N	A	SA

Openness

I have a very active imagination.	SD	D	N	A	SA

SOURCE: Reproduced by special permission of the publisher, Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, from the NEO Personality Inventory-Revised, by Paul T. Costa, Jr., PhD, and Robert R. McCrae, PhD, Copyright 1978, 1985, 1989, 1991, 1992 by Psychological Assessment Resources (PAR). Further reproduction is prohibited without permission from PAR.

on achievement tests to compare what students know at the beginning of the year with what they know at the end of the year, to assign grades, to identify students with special educational needs, and to measure students' progress.

Aptitude Tests

Achievement tests measure a test taker's knowledge in a specific area at a specific point in time. **Aptitude tests** assess a test taker's potential for learning or ability to perform in a new job or situation. Aptitude tests measure the product of cumulative life experiences—or what one has acquired over time. They help determine what "maximum" can be expected from a person.

Schools, businesses, and government agencies often use aptitude tests to predict how well someone will perform or to estimate the extent to which an individual will profit from a specified course of training. Vocational guidance counseling may involve aptitude testing to help clarify the test taker's career goals. If a person's score is similar to scores of others already working in a given occupation, the test will predict success in that field.

Intelligence Tests

Intelligence tests, like aptitude tests, assess the test taker's ability to cope with the environment, but at a broader level. Intelligence tests are often used to screen individuals for specific programs (for example, gifted programs, honors programs) or programs for the mentally challenged. Intelligence tests are typically used in educational and clinical settings.

Interest Inventories

Interest inventories assess a person's interests in educational programs for job settings and provide information for making career decisions. Because these tests are often used to predict satisfaction in a particular academic area or employment setting, they are administered primarily to students by counselors in high schools and colleges. Interest inventories are not intended to predict success; rather, they are intended only to offer a framework for narrowing career possibilities.

Personality Tests

Personality tests measure human character or disposition. The first personality tests were designed to assess and predict clinical disorders. These tests remain useful today for determining who needs counseling and who will benefit from treatment programs. Newer personality tests measure "normal" personality traits. For example, the Myers–Briggs Type Indicator (MBTI) is often used by industrial/organizational psychologists to increase employees' understanding of individual differences and to promote better communication between members of work teams. Career counselors also use the MBTI to help students select majors and careers consistent with their personalities.

> More detail about the MBTI can be found in Test Spotlight 1.4 in Appendix A.

Personality tests can be either objective or projective. The MBTI is an example of an objective personality test. Projective personality tests, such as the TAT, serve the same purpose as some objective personality tests, but they require test takers to respond to unstructured or ambiguous stimuli.

Subject Tests

Many popular psychological testing reference books also classify tests by subject. For example, the *Seventeenth Mental Measurements Yearbook* (Geisinger, Spies, Carlson, & Plake, 2007) classifies thousands of tests into 19 major subject categories:

- Achievement
- Behavior assessment
- Developmental
- Education
- English
- Fine arts
- Foreign languages
- Intelligence
- Mathematics
- Miscellaneous (for example, courtship and marriage, driving and safety education, etiquette)
- Multiaptitude batteries
- Neuropsychological
- Personality
- Reading
- Science
- Sensorimotor

- Social studies
- Speech and hearing
- Vocations

Reference books such as the *Mental Measurements Yearbook* often indicate whether a test is (a) a test of maximal performance, a behavior observation test, or a self-report test; (b) standardized or nonstandardized; and (c) objective or projective. We discuss the *Mental Measurements Yearbook*, as well as other reference books, later in this chapter.

INTERIM SUMMARY 1.3
ASSUMPTIONS AND TEST CLASSIFICATION METHODS

When using psychological tests, the following assumptions must be made:

- Psychological tests measure what they say they measure, and any inferences that are drawn about test takers based on their test scores are appropriate.
- An individual's behavior, and therefore test scores, will remain unchanged over time.
- Individuals understand test items similarly.
- Individuals can report about themselves accurately.
- Individuals will report their thoughts and feelings honestly.
- The test score an individual receives is equal to his or her true ability plus some error.

Psychological tests can be classified in many different ways:

- As tests of maximal performance, behavior observation tests, or self-report tests
- As standardized or nonstandardized
- As objective or projective
- Based on the dimensions they measure
- Based on subject

Psychological Assessment, Psychological Tests, Measurements, and Surveys

Before discussing much more, we should spend some time discussing some terms that students often confuse—psychological assessment, psychological tests, measurement, and surveys. Students often think of psychological assessment and psychological testing as one and the same. Similarly, students often do not understand the difference between psychological tests and surveys. This section is designed to help you distinguish among these terms that are commonly used in psychological testing.

Psychological Assessments and Psychological Tests

Psychological assessments and psychological tests both are methods of collecting important information about people, and both are also used to help understand and predict behavior (Kline, 2000, Maloney & Ward, 1976). Assessment, however, is a broader concept than psychological testing. **Psychological assessment** involves multiple methods, such as personal history interviews, behavioral observations, and psychological tests, for gathering information about an individual. Psychological assessment involves *both* an

objective component and a subjective component (Matarazzo, 1990), and psychological tests are only one tool in the assessment process. For example, a clinical psychologist may conduct a psychological assessment of a patient and, as a part of this assessment, may administer a psychological test such as the MMPI.

Psychological Tests and Measurements

Although the meanings overlap, *psychological test* and *measurement* are not synonyms. **Measurement,** broadly defined, is the assignment of numbers according to specific rules. The concept of measurement is represented by the darker circle in Figure 1.2.

Psychological tests require test takers to answer questions or perform tasks to measure personal attributes. The concept of a psychological test is represented by the lighter circle in the figure. With psychological tests, test takers' answers to questions or their performance on some task is not initially expressed in physical units of any

> More detail about the MBTI can be found in Test Spotlight 1.4 in Appendix A.

kind; instead, scores are derived according to some predetermined method. In some cases, the end result of a psychological test is not a derived score at all, but rather a verbal description of an individual. For example, there are some personality tests that, although they have rules for scoring or summarizing information, do not produce overall scores. Instead, these tests yield profiles. The MBTI is an example of such a test.

Psychological tests can be considered psychological measurements when a sample of behavior can be expressed as a numerical score. This is represented by the overlapping section of the two circles in Figure 1.2.

You will find that many people use the terms *psychological test* and *psychological measurement* interchangeably. Although most psychological tests are measurements, not all psychological tests, strictly defined, meet the definition of a measurement. Throughout the remainder of this text, we follow the common practice of referring to all psychological tests as measurements because most of them are, but keep in mind the distinctions we have drawn in this section.
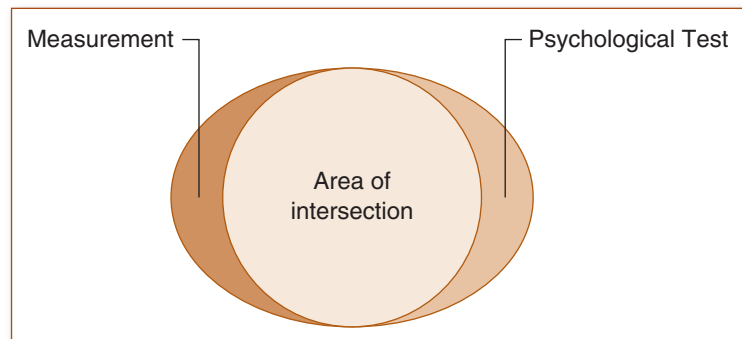
Psychological Tests and Surveys

Surveys, like psychological tests (and psychological assessments), are used to collect important information from individuals. Surveys differ from psychological tests in two important ways. First, psychological tests focus on individual outcomes, and surveys focus on group outcomes. Psychological tests provide important information about individual differences and help individuals and institutions make important decisions about individuals. For example, a psychological test may suggest that a child is unusually intelligent and therefore should be placed in a gifted or honors program. Surveys, on the other hand, provide important information about groups and help us make important decisions about groups. For example, an organizational survey may suggest that employees are displeased with a company benefits program and that a new benefits program is needed.

Figure 1.2 A Comparison of Measurement and Psychological Testing (the area of intersection represents samples of behavior expressed as numerical scores)

Measurement ⎯   ⎯ Psychological Test

Area of intersection

Second, the results of a psychological test are often reported in terms of an overall derived score or scaled scores. Results of surveys, on the other hand, are often reported at the question level by providing the percentage of respondents who selected each answer alternative. Of course, in some cases, surveys focus on individual outcomes and are constructed using scales. In such cases, the survey approximates a psychological test. (Chapter 10 is devoted to an in-depth discussion of surveys.)

Locating Information About Tests

With so many psychological tests available, we are sure you can imagine that finding the most appropriate one for your specific purpose can be a difficult task. To choose an appropriate test for a particular circumstance, you must know the types of tests that are available and their merits and limitations. Prior to the 1950s, test users had few resources for obtaining such information. Today, however, numerous resources are available. Although all have the same general purpose—to help test users make informed decisions—the information such resources contain varies. Some resources provide only general descriptive information about psychological tests, such as the test's name, author, and publisher, and others contain detailed information, including test reviews and detailed bibliographies. Some resources focus on commercially available, standardized published tests, and others focus on unpublished tests. Some references include information about tests for particular groups (for example, children), and others include a broad range of tests for various populations.

Some of the most commonly used resource books, including a brief synopsis of the contents as well as a library catalog number, are described in For Your Information Box 1.3. The first four resource books, the *Mental Measurements Yearbook* (*MMY*), *Tests in Print* (*TIP*), *Tests*, and *Test Critiques*, are often viewed as the most useful and popular (American Psychological Association, 2010b). Note that although different libraries may give a particular reference a different catalog number, the one we have supplied will direct you to the general area where you will find the book. If you cannot find a particular book, ask the librarian for assistance; your library might not carry the reference book, and the librarian can help you find the book at another location.

FYI

FOR YOUR INFORMATION BOX 1.3

Commonly Used Resource Books

Book Title	Contents	Reference Number
Tests in Print (multiple volumes)	*Tests in Print (TIP)* is published in multiple volumes. Each volume contains descriptive listings of commercially published psychology and achievement tests that are available for purchase. *TIP* also serves as a comprehensive index to the contents of previously published *Mental Measurements Yearbooks* (see below for a description of the *Mental Measurements Yearbook*). Each descriptive listing, or test entry, contains extensive information, including but not limited to the title of the test, the purpose of the test, the intended population, publication dates, the acronym used to identify the test, scores the test provides, whether the test is an individual test or group test, whether the test has a manual, the author(s), the publisher, the cost of the test, and available foreign adaptations. Each entry also contains brief comments about the test as well as cross-references to reviews in the *Mental Measurements Yearbooks*.	LB3051.T47

Book Title	Contents	Reference Number
Mental Measurements Yearbook (multiple volumes)	The *Mental Measurements Yearbook (MMY)* is published in multiple volumes. Each volume contains descriptive information and test reviews of new English-language, commercially published tests and tests that have been revised since the publication of the previous *MMY* edition. The *MMY* is cumulative, meaning that later volumes build on earlier ones rather than replace them. Each descriptive listing, or test entry, contains extensive information about a particular test. If the test is a revision of a previous test, the entry also includes the volume of the *MMY* in which the test was originally described. Each entry also typically includes information about the test's reliability and validity, one or two professional reviews, and a list of references to pertinent literature. For a guide to descriptive entries in the *MMY*, see Figure 1.3. The *MMY* is very likely accessible electronically through your college's library system.	LB3051.M4
Tests	*Tests* contains descriptions of a broad range of tests for use by psychologists, educators, and human resource professionals. Each entry includes the test title, the author, the publisher, the intended population, the test purpose, major features, the administration time, the cost, and the availability.	BF176.T43
Test Critiques (multiple volumes)	*Test Critiques* is published in multiple volumes. Each volume contains reviews of frequently used psychological, business, and educational tests. Each review includes descriptive information about the test (for example, author, attribute measured, norms) and information on practical applications and uses. *Test Critiques* also contains in-depth information on reliability, validity, and test construction.	BF176.T419
Personality Test and Reviews (multiple volumes)	*Personality Test and Reviews* is published in volumes. Each volume contains a bibliography of personality tests that are contained in the *MMY*. Each entry contains descriptive information about the test as well as test reviews.	BF698.5B87
Tests in Education	*Tests in Education* contains descriptive and detailed information about educational tests for use by teachers, administrators, and educational advisers.	LB3056.G7.L49
Measures for Psychological Assessment	*Measures for Psychological Assessment* contains annotated references to journal articles and other publications in which measures of primarily mental health are described.	BF698.5C45
Testing Children	*Testing Children* contains descriptions of tests available for children. These descriptions include the knowledge, skills, and abilities measured by each test; the content and structure of the test; the time required to administer the test; the scores that are produced; the cost; and the publisher.	BF722.T47
Test and Measurements in Child Development: A Handbook	*Tests and Measurements in Child Development* contains a listing of unpublished measures for use with children as well as detailed information about each measure.	BF722.J64
Measures for Psychological Assessment: A Guide to 3,000 Original Sources and Their Applications	*Measures for Psychological Assessment is* a guide that contains annotated references to thousands of less recognized assessment devices developed and described in journal articles.	155.28016.C559

Whether you are trying to locate tests that measure intelligence, self-esteem, or some other attribute, trait, or characteristic, we suggest that you begin your search with one of the first four resource books in For Your Information Box 1.3. *TIP* and the *MMY* are two of the most helpful references, and students often find it most helpful to begin with *TIP*. Figure 1.3 includes a descriptive guide of the type of information you will find in the *MMY*. Figure 1.4 includes a summary of how to use *TIP* to find tests. You can find more information on how to use both of these resources, as well as how to use the information contained in these resources to evaluate a test, on the Buros homepage discussed in On the Web Box 1.2.

Figure 1.3 A Guide to Descriptive Entries in the *Mental Measurements Yearbook*

SOURCE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln. www.unl.edu/buros

Because there is a wealth of psychological tests available, there is a wealth of resources available for you to use in gathering information about psychological tests. You are not limited to print resources; advances in technology now allow you to access the Internet and gather information about psychological tests on demand. On the Web Box 1.2 discusses some websites you can access to locate information on psychological tests. For Your Information Box 1.4 discusses where you can locate unpublished psychological tests.

| Figure 1.4 | How to Use *Tests in Print* |

BUROS INSTITUTE OF MENTAL MEASUREMENTS

How to Use
Tests in Print

Tests in Print (TIP) *consists of descriptive listings, without reviews, of commercially published tests in print.* TIP *is also a comprehensive index to the contents of previously published* Mental Measurements Yearbooks.

1. If you know the TEST TITLE:

Use the "**Index of Titles**." The index lists all tests in that volume plus all tests out of print since last being listed. "2458," for example, refers to test 2458 in that volume; "9:1128" refers to now out-of-print test 1128 in the *Ninth Mental Measurements Yearbook*. *Citation numbers refer to entry numbers, not to page numbers.*

Example from "Index of Titles":

> Short Tests of Clerical Ability, 2458
> Shortened Edinburgh Reading Test, 2459
> Shortened Aptitude Test, T4:2195
> Signals Learning Test, 2461
> Silver Burdett Music Competency Tests, 9:1128
> Silver Drawing Test of Cognitive Skills and Adjustment, 2462
> Simile Interpretations, T4:2198
> Similes, T4:2199

2. If you know the TYPE OF TEST:

Use the "**Classified Subject Index**" to locate various categories of tests, such as achievement, intelligence, personality, etc. This index organizes all tests into 18 major categories; tests appear alphabetically within each category. *Citation numbers refer to entry numbers, not to page numbers.*

Example from "Classified Subject Index, Education":

> Gifted Program Evaluation Survey, Gifted and talented programs, see 1040
> Graduate Records Examinations Education Test, Graduate School candidates, see 1063
> High School Characteristics Index, Grades 9-13, 4-13, see 1157
> How a Child Learns, Classroom teachers, see 1175
> Hudson Educational Skills Inventory, Grades K-12, see 1184

3. If you know the NAME OF THE TEST AUTHOR OR REVIEWER:

Use the "**Index of Names**." This index includes test authors (for example, "test, 1460"), review authors ("rev, 2589"), and authors of referenced articles ("ref, 2222"). (Parenthesized numbers indicate the reference number.) *Citation numbers refer to entry numbers, not to page numbers.*

Example from "Index of Names":

> Caeglio, G.: test, 1460
> Caffey, C. A.: ref, 2222(1)
> Caggiula, A. A.: ref, 2563(449)
> Cahalane, J.: ref, 268(395), 1043(39)
> Cahen, L. S.: rev, 2589
> Cahill, C.: ref, 1705(65), 2937(935)
> Cahir, N.: ref, 1135(14), 2674(188)
> Cahn, T. S.: ref, 268(90)
> Cain, J.: ref, 93(84), 1690(84)
> Cain, L. F.: test, 2844

Buros Institute of Mental Measurements
University of Nebraska-Lincoln
21 Teachers College Hall
Lincoln, NE 68588-0348

This is not copyrighted material. Reproduction and dissemination are encouraged.
See the Buros Institute website at **www.unl.edu/buros**

ON THE WEB BOX 1.2

Locating Information About Tests on the Web

Computer technology lets us connect to the World Wide Web and locate websites containing valuable information about psychological tests. These websites include information such as the following:

- Frequently asked questions about psychological testing
- How to find a particular type of psychological test
- How to locate reviews of psychological tests
- How to select an appropriate test
- What qualifications are necessary to purchase psychological tests
- How to contact test publishers
- How to obtain copies of specific psychological tests

Although there are many available websites, here are four that we have found to be extremely valuable:

(Continued)

Website	Description
American Psychological Association www.apa.org/science /programs/testing/find -tests.aspx#findinfo	Although the American Psychological Association (APA) does not sell or endorse specific testing instruments, it does provide guidance on testing resources and how to find psychological tests. This website contains answers to the most frequently asked questions about psychological testing. One section focuses on questions about published psychological tests (those that can be purchased from a test publisher); here you will find advice on how to find information about a particular test and about the proper use of tests, how to contact test publishers and purchase tests, and available software and scoring services. Another section focuses on unpublished psychological tests and measures (those that are not commercially available); here you will find advice on how to find unpublished tests in your area of interest and important information regarding your responsibilities as a user of unpublished tests.
Buros Institute of Mental Measurement www.unl.edu/buros/ bimm/index.html	The Buros Institute of Mental Measurement promotes the appropriate use of tests and provides professional assistance, expertise, and information to those who use commercially published tests. This website contains a number of instructional resources, tools, and links. For example, it contains detailed instructions on what information can be found in two popular Buros publications that we have already discussed: the *Mental Measurements Yearbook* and *Tests in Print.* This site also contains some great "how to" resources such as how to use *Tests in Print* and the *Mental Measurements Yearbook* and how to use the information in these resources to evaluate a test. In addition, it contains a link to *Test Reviews Online,* a service that provides access to more than 2,000 test reviews, beginning with those that are published in the *Ninth Mental Measurements Yearbook.* Likewise, there are links to the *Code of Fair Testing Practices* (discussed further in Chapter 3) and the APA's frequently asked questions website mentioned previously.
Educational Testing Service Test Link www.ets.org/testcoll/	Educational Testing Service Test Link is the world's largest database of tests and measurement instruments that have been available since the early 1900s. This online database contains descriptions of more than 20,000 tests (published and unpublished) and research instruments, collected from test publishers and test authors from around the world. Each description includes the title of the test/instrument, the author, the publication date, availability (how to obtain the test or measurement), the intended population, and specific uses of the test/instrument. In addition to providing information about specific tests, this database contains valuable information on how to order tests.
O-Net Resource Center www.onetcenter.org/ guides.html	The Occupational Information Network (O-Net) is sponsored by the U.S. Department of Labor and is a primary source for occupational information. Consisting of a comprehensive database of worker attributes and job characteristics, O-Net also provides valuable resources on testing and assessment—resources intended to support public and private sector efforts to identify and develop the skills of the American workforce. This website provides access to three extremely valuable testing and assessment guides: • *Testing and Assessment: A Guide to Good Practices for Workforce Investment Professionals* includes information on how assessment instruments can be used to promote talent development in career counseling, training, and other talent development activities. It discusses how to evaluate and select assessment instruments, administer and score assessments to meet business and individual client needs, and accurately and effectively interpret assessment results. It also lists the professional and legal standards related to assessment use in talent development. • *Tests and Other Assessments: Helping You Make Better Career Decisions* includes an explanation of how assessment instruments are used in employment selection and career counseling and provides tips and strategies for taking tests and other assessments. • *Testing and Assessment: An Employer's Guide to Good Practices* helps managers and workforce development professionals understand and use employment testing and assessment practices to meet their organizations' human resources goals.

FYI

FOR YOUR INFORMATION BOX 1.4

Locating Unpublished Psychological Tests

Although there are thousands of commercially available tests, there are just as many, if not more, unpublished tests designed and used by researchers. A number of print and nonprint resources are available for locating information on unpublished tests.

Two of the most popular print resources are the *Directory of Unpublished Experimental Measures* and *Measures for Psychological Assessment: A Guide to 3,000 Original Sources and Their Applications*. Three of the most popular nonprint resources for locating information about unpublished or noncommercial tests are Tests in Microfiche, the PsycINFO database, and the Health and Psychosocial Instruments database.

Directory of Unpublished Experimental Measures (Goldman, Mitchell, & Egelson, 1997)

This directory provides easy access to more than 5,000 experimental mental measures, tests, and surveys that have been used by other researchers but are not commercially available. Topics range from educational adjustment and motivation to personality and perception. The measures, tests, and surveys are arranged in a 24-category system and grouped according to function and content, noting purpose, format, psychometric information (where available), and related research. First published in 1974 and currently in its seventh edition, this resource is updated periodically by the publisher.

Measures for Psychological Assessment: A Guide to 3,000 Original Sources and Their Applications (Chun, Cobb, & French, 1975)

This guide includes annotated references to psychological measures that have appeared in journal articles and other publications. Although a bit outdated, this can be a useful resource. It has two sections: primary references and applications. The primary references section includes the name of each measure, the reference in which the measure originally appeared, and one or more other researchers who have used the measure in experimental research. The applications section includes other research studies that have used the original measures and references other experimental tests.

Tests in Microfiche

This resource can be accessed through the Educational Testing Service Test Link. It contains a variety of educational and psychological instruments that are cited in the literature but are either out of date or unpublished. It contains more than 1,000 tests, and new tests are added each year. For more information, go to www.ets.org/testcoll or check with your college's library.

PsycINFO Database

This bibliographic database indexes published studies in psychology. By using the Form/Content field "Tests & Measures" to search the PsycINFO database, you can find tests that have been used in research and written about in the literature. For more information, go to www.apa.org/pubs/databases/psycinfo/index.aspx.

Health and Psychosocial Instruments Database (HAPI)

This computerized database includes citations to unpublished health and psychosocial evaluation and measurement tools (for example, questionnaires, interviews, tests, checklists, rating scales) that have appeared in journals and technical reports since 1985. HAPI is updated quarterly and contains more than 15,000 measurement instruments. HAPI is provided online by Ovid Technologies, which typically must be accessed through BRS Information Technologies at your college's library. Some libraries maintain the database on CD-ROM. For more information, see www.ovid.com/site/catalog/DataBase/866.jsp

Chapter Summary

By now, we hope you understand that psychological testing extends well beyond the use of intelligence and personality tests. Anything that requires a test taker to perform a behavior that is used to measure some personal attribute, trait, or characteristic or to predict an outcome can be considered a psychological test. The quizzes and exams you take in class are psychological tests. The written and road portions of driving exams are psychological tests. Even the structured job interviews you have participated in, or will participate in as you conduct your job search, qualify as psychological tests.

Psychological tests have various similarities and many differences. All psychological tests require an individual to perform one or more behaviors, and these behaviors are used to measure some personal attribute, trait, or characteristic thought to be important in describing or understanding behavior or to predict an outcome. However, psychological tests can and do differ in terms of the behaviors they require individuals to perform, the attributes they measure, their content, how they are administered and formatted, how they are scored and interpreted, and their psychometric quality.

Although the use of psychological tests can be traced to ancient China, most scholars agree that the advent of formal psychological testing did not begin until Binet published the first test of intelligence in 1905. Today, psychological testing is a big business, with tens of thousands of commercially available, standardized psychological tests as well as thousands of unpublished tests.

All good tests have three defining characteristics in common. First, they include a representative sample of behaviors. Second, they collect the sample under standardized conditions. Third, they have rules for scoring. When using psychological tests, we must make some assumptions. We must assume that a test measures what it says it measures, that any inferences that are drawn about test takers from their scores on the test are appropriate, that an individual's behavior (and therefore test scores) will remain stable over time, that individuals understand test items similarly, that individuals can and will report accurately about their thoughts and feelings, and that the test score an individual receives is equal to his or her true behavior/ability in the real world plus some error.

Testing professionals refer to psychological tests in various ways. Sometimes they refer to them as tests of maximal performance, behavior observations, or self-reports. Sometimes they refer to them as standardized or nonstandardized. Other times they refer to them as objective or projective. Professionals also refer to tests based on the dimensions they measure.

It is important to remember the distinctions among four commonly misunderstood terms: *psychological assessment, psychological test, measurement,* and *survey.* First, although both psychological assessments and psychological tests are used to gather information, a psychological test is only one of many tools in the psychological assessment process. Second, a psychological test can be considered to be a measurement when the sampled behavior can be expressed in a derived score. Third, psychological tests are different from surveys in that psychological tests focus on individual differences and often report one overall derived score (or scaled scores), and surveys focus on group similarities and typically report results at the question or item level.

Last, but not least, a number of resources are available, in print and online, to locate information about published and unpublished psychological tests and measures. The *Mental Measurements Yearbook* and *Tests in Print* are two of the most popular references for learning more about available tests.

Engaging in the Learning Process

KEY CONCEPTS

After completing your study of this chapter, you should be able to define each of the following terms. These terms are bolded in the text of this chapter and defined in the Glossary.

- achievement tests
- aptitude tests
- behavior
- behavior observation tests
- emotional intelligence
- intelligence tests
- interest inventories
- measurement
- nonstandardized tests
- norms
- objective tests

- personality tests
- projective tests
- psychological assessments
- psychological tests
- psychometrics
- self-report tests
- standardization sample
- standardized tests
- surveys
- tests of maximal performance
- vocational tests

LEARNING ACTIVITIES

The following are some learning activities you can engage in to support the learning objectives for this chapter.

Learning Objectives	Study Tips and Learning Activities
After completing your study of this chapter, you should be able to do the following:	*The following study tips will help you meet these learning objectives:*
Define what a psychological test is, and understand that psychological tests extend beyond personality and intelligence tests.	• Write your definition of a psychological test. List examples of psychological tests, from what comes to your mind first to what comes to your mind last. Compare your list of examples with Figure 1.1. • Ask various professionals, in and outside of the psychology field, to define what a psychological test is. Compare and contrast their definitions. Compare these definitions with the definitions provided in this textbook. Discuss why definitions might vary.
Trace the history of psychological testing from Alfred Binet and intelligence testing to the tests of today.	• Reflect on the history of testing. Create a timeline showing significant events in testing, beginning with testing in ancient China and ending with testing today.
Describe the ways in which psychological tests can be similar to and different from one another.	• Think about two exams you recently took. Make two lists: one of how they were similar and another of how they were different. Compare your lists with Interim Summary 1.1.
Describe the three characteristics that are common to all psychological tests, and understand that psychological tests can demonstrate these characteristics to various degrees.	• Recall the three characteristics common to all psychological tests. Make three columns, and label them Representative Sample of Behaviors, Standardized Conditions, and Rules for Scoring. Select one or two psychological tests that you have taken. Write how the test(s) demonstrate(s) each characteristic.

(Continued)

Learning Objectives	Study Tips and Learning Activities
	• Construct an eight-question quiz, with one question for each learning objective. Give the quiz to your classmates (your professor will determine the logistics of this). As a class, discuss whether the quiz meets all of the characteristics of a psychological test. What were the strengths of your quiz? How could your quiz have been improved?
Describe the assumptions we must make when using psychological tests.	• Describe the six assumptions we must make when using psychological tests. Without looking in your book, see how many assumptions you can write. Compare your written assumptions with the assumptions in the book. Explain why we must make these assumptions.
Describe the ways that psychological tests can be classified.	• Review the test classification methods in your book. Think about the road portion of the driving test, the SAT, a job interview, the NEO Personality Inventory, and a multiple-choice test you took recently. Classify each test using the different test classification methods.
Describe the differences among four commonly used terms that students often get confused: psychological assessment, psychological tests, measurement, and surveys.	• Draw a picture or diagram illustrating how these four commonly confused terms overlap.
Identify and locate printed and online resources that are available for locating information about psychological tests.	• Go to your college library and find *Tests in Print* and the *Mental Measurements Yearbook*. Write the names of three tests and what they measure. • Go to each of the websites referenced in your book. Compare and contrast the information found on these websites. • Select a psychological test that is mentioned in Chapter 1 or 2 or that is suggested by your instructor. Using reference books available at your college library and online, collect as much of the information as possible about your test. Keep track of where you found the information.

PRACTICE QUESTIONS

The following are some practice questions to assess your understanding of the material presented in this chapter.

Multiple Choice

Choose the one best answer to each question.

1. What do all psychological tests require that you do?
   a. Answer questions
   b. Fill out a form
   c. Perform a behavior
   d. Sign a consent form

2. According to the textbook, which one of the following is least typical of psychological tests?
   a. Personality tests
   b. Intelligence tests
   c. Structured interviews
   d. Classroom tests

3. Who published the first
   test of intelligence in 1905?
   a. Lewis Binet
   b. Alfred Simon
   c. Robert Woodworth
   d. Alfred Binet

4. Who published the Stanford–Binet?
   a. Henry Murray
   b. Robert Woodworth
   c. Lewis Terman
   d. Alfred Binet

5. What test did Robert Woodworth
   develop during World War I to help
   the U.S. military detect soldiers
   who would not be able to handle
   the stress associated with combat?
   a. Thematic Apperception Test
   b. Stanford–Binet
   c. Personal Data Sheet
   d. Rorschach Inkblot Test

6. What was the first widely used personality
   inventory?
   a. Woodworth Psychoneurotic Inventory
   b. Personal Data Sheet
   c. Rorschach Inkblot Test
   d. Thematic Apperception Test

7. A test that requires you to demonstrate your
   driving ability can best be classified as what
   type of test?
   a. Test of maximal performance
   b. Self-report test
   c. Behavior observation test
   d. Projective test

8. A test that requires you to
   respond to test questions about
   your feelings and beliefs can best be described
   as what type of test?
   a. Test of maximal performance
   b. Self-report test
   c. Behavior observation test
   d. Projective test

9. The role of the test taker is least clear in which
   one of the following?
   a. Objective tests
   b. Projective tests
   c. Standardized tests
   d. Self-report tests

10. What type of test is
    administered to a large
    group of individuals who
    are similar to the group for
    which the test has been designed?
    a. Nonstandardized test
    b. Standardized test
    c. Objective test
    d. Subjective test

11. What type of test would a classroom
    teacher most likely administer?
    a. Achievement test
    b. Aptitude test
    c. Intelligence test
    d. Interest inventory

12. What type of test assesses test
    takers' potential for learning or
    ability to perform in an area in
    which they have not been
    specifically trained?
    a. Achievement test
    b. Intelligence test
    c. Aptitude test
    d. Vocational test

13. What type of test requires
    test takers to respond to
    structured true/false questions,
    multiple-choice questions,
    and/or rating scales?
    a. Projective test
    b. Nonstandardized test
    c. Subjective test
    d. Objective test

14. What type of test would a
    career development counselor
    most likely administer?
    a. Achievement test
    b. Aptitude test
    c. Intelligence test
    d. Interest inventory

15. Which one of the following
    would be the best source for locating
    a professional test review for a
    commercially available published test?
    a. *Tests in Print*
    b. *Tests in Microfiche*
    c. *Mental Measurements Yearbook*
    d. *Measures for Psychological Assessment*

Short Answer/Essay

Read each of the following, and consider your response carefully based on the information presented in this chapter. Write your answer to each question in two or three paragraphs.

1. What is a psychological test?
2. Why should you care about psychological tests?
3. What three characteristics do all psychological tests have in common? Explain and provide an example of each.
4. Summarize the ways in which psychological tests can be similar to and different from one another.
5. When using a psychological test, what assumptions must be made? Why are these assumptions important?
6. What are the similarities and differences among intelligence tests, aptitude tests, and achievement tests? Provide an example of each.
7. How are psychological assessments, psychological tests, and measurement similar? How are they different?
8. How are psychological tests and surveys similar? How are psychological tests and surveys different?

ANSWER KEYS

Multiple Choice

1. c	2. c	3. d	4. c	5. c
6. a	7. a	8. b	9. b	10. b
11. a	12. c	13. d	14. d	15. c

Short Answer/Essay

Refer to your textbook for answers. If you are unsure of an answer and cannot generate the answer after reviewing your book, ask your professor for clarification.

WikipediA

List of standardized tests in the United States

A **standardized test** is a test administered and scored in a standard manner. The following are such tests as administered across the United States.

Contents

Ability/Achievement tests

Ability/ Achievement tests are used to evaluate a student's or worker's understanding, comprehension, knowledge and/or capability in a particular area. They are used in academics, professions and many other areas.

A general distinction is usually made between tests of ability/ aptitude (intelligence tests) versus tests of achievement (academic proficiency).

IQ tests

- Stanford-Binet Intelligence Scales (SB5)
- Wechsler Adult Intelligence Scale (WAIS)
- Wechsler Intelligence Scale for Children (WISC)
- Wechsler Preschool and Primary Scale of Intelligence (WPPSI)
- Otis-Lennon School Ability Test
- Differential Ability Scales (DAS)

- Woodcock-Johnson Tests of Cognitive Abilities (WJ)

Achievement tests

- Wechsler Individual Achievement Test (WIAT)
- Kaufman Test of Educational Achievement (KTEA)
- Woodcock-Johnson Tests of Achievement (WJ)
- Peabody Individual Achievement Test (PIAT-R)
- Wide Range Achievement Test, 5th Ed. (WRAT-5)

Public schools

- National Assessment of Educational Progress (NAEP)
- State achievement tests are standardized tests. These may be required in American public schools for the schools to receive federal funding, according to the US Public Law 107-110 originally passed as Elementary and Secondary Education Act of 1965, and currently authorized as Every Student Succeeds Act in 2015. No Child Left Behind was the controversial version of the law signed by President G. W. Bush in 2001; it was reauthorized in 2015 by President B. Obama.
- Exit examinations for high school graduation

Other tests

The test of General Educational Development (GED) and Test Assessing Secondary Completion TASC evaluate whether a person who has not received a high school diploma has academic skills at the level of a high school graduate.

Private tests are tests created by private institutions for various purposes, such as progress monitoring in K-12 classrooms.

- ACT
  - PLAN
  - EXPLORE[1]
- California Achievement Test
- ITBS - Iowa Test of Basic Skills[2]
- SAT - formerly Scholastic Aptitude Test
  1. SAT Subject Tests
- CLT - Classic Learning Test (https://www.cltexam.com/)
- Former English Language Proficiency Test - ELPT
- PSAT/NMSQT - Preliminary SAT/National Merit Scholarship Qualifying Test
- STAR Early Literacy, STAR Math, and STAR Reading
- Stanford Achievement Test
- TerraNova
- WorkKeys

Admissions tests

Admissions tests are used in the admission process at elite or private elementary and secondary schools, as well as most colleges and universities. They are generally used to predict the likelihood of a student's success in an academic setting.[3]

Secondary school

- ISEE - Independent School Entrance Examination
- SSAT - Secondary School Admission Test
- HSPT - High School Placement Test
- COOP- Cooperative admissions examination program
- SHSAT - Specialized High School Admissions Test

Undergraduate

- SAT - formerly Scholastic Aptitude Test
  - SAT Subject Tests
  - Former English Language Proficiency Test - ELPT
- ACT - formerly American College Testing Program or American College Test
- ACCUPLACER - community colleges and 4 year colleges placement test
- CLT - Classic Learning Test (https://www.cltexam.com/)

Graduate/professional schools

- Allied Health Professions Admission Test (AHPAT)
- Dental Admission Test (DAT)- (United States)
- Graduate Management Admission Test (GMAT) - (US)
- Graduate Record Examination (GRE) - (US and Canada)
- Law School Admission Test (LSAT) - (US and Canada)
- Miller Analogies Test (MAT)
- Medical College Admission Test (MCAT) - (US and Canada)
- Optometry Admission Test (OAT) - Optometry Admission Test
- Pharmacy College Admission Test (PCAT)
- Veterinary College Admission Test (VCAT) – no longer administered; American veterinary schools now use either the GRE or MCAT
- California Basic Educational Skills Test
- Wiesen Test of Mechanical Aptitude (WTMA)

Language proficiency

- TOEIC - Test of English for International Communication
- TOEFL - Test of English as a Foreign Language
- IELTS - International English Language Testing System

Psychological tests

- 16 Personality Factors
- Achievement Motivation Inventory

- Beck Depression Inventory
- Millon Clinical Multiaxial Inventory
- Minnesota Multiphasic Personality Inventory (MMPI)
- Personality Assessment Inventory
- Myers-Briggs Type Indicator (MBTI)
- Revised NEO Personality Inventory
- Thematic Apperception Test

Professional certification tests

- Certified Public Accountant (CPA) for Accountants
- Chartered Financial Analyst (CFA)
-
- COMLEX-USA for osteopathic physicians
- Examination for Professional Practice in Psychology (EPPP), the most common certification for practitioners of Clinical Psychology in the U.S.
- Fundamentals of Engineering (FE), the first of two exams that must be passed to become a Professional Engineer
- General Securities Representative Examination, more commonly known as the Series 7 Exam, required to receive a license as a stockbroker in the U.S.
- Investment Company Products/Variable Life Contracts Representative Examination, more commonly known as the Series 6 Exam, for U.S. licensing to sell a limited set of securities such as mutual funds and variable life insurance
- Multistate Bar Examination (MBE), part of the bar examination in almost all United States jurisdictions
- Multistate Pharmacy Jurisprudence Examination (MPJE), a prerequisite for licensure as a pharmacist in the vast majority of U.S. jurisdictions
- Multistate Professional Responsibility Examination (MPRE), a requirement for bar admission in addition to the bar examination in almost all U.S. jurisdictions
- NAPLEX, required by all U.S. jurisdictions for licensure as a pharmacist
- NCLEX-PN for Licensed Practical Nurses
- NCLEX-RN for Registered Nurses
- Physician Assistant National Certifying Exam for physician assistants (PA)
- PRAXIS for Teacher certification
- Principles and Practice of Engineering Exam the second of the two exams someone must pass to become a Professional Engineer
- Uniform Certified Public Accountant Examination
- Uniform Combined State Law Examination, more commonly called the Series 66 Exam, required by some U.S. states for state certification as both a securities agent and investment adviser representative
- Uniform Securities Agent State Law Examination, more commonly known as the Series 63 Exam, required by almost all U.S. states for state certification as a securities agent
- United States Medical Licensing Examination for physicians (holders of either Doctor of Medicine or Doctor of Osteopathic Medicine degrees)
- USPTO registration examination, a requirement of the United States Patent and Trademark Office for registration as a patent attorney or agent

Armed Forces

ASVAB (United States) required for entry into any branch of The United States Military. Other tests, such as AFOQT and ASTB are used for officers.

See also

- List of admissions tests
- Standards-based assessment

References

1. "EXPLORE web page" (https://web.archive.org/web/20070713203006/http://www.act.org/explore/). Archived from the original (http://www.act.org/explore/) on 2007-07-13. Retrieved 2007-07-15.
2. "Iowa Testing Programs - College of Education - the University of Iowa" (https://web.archive.org/web/20120729221024/http://itp.education.uiowa.edu/). Archived from the original (http://www.itp.education.uiowa.edu/) on 2012-07-29. Retrieved 2012-08-05.
3. "Glossary" (http://www.getcollegefunds.org/glossary.html), Oregon Student Admissions Commission. Retrieved 4/1/08.

Retrieved from "https://en.wikipedia.org/w/index.php?title=List_of_standardized_tests_in_the_United_States&oldid=1105857040"

This page was last edited on 22 August 2022, at 03:47 (UTC).

2.1 Meaning, Purpose and Construction of Achievement Test

2.1.1 Meaning of Achievement Test

Achievement means one's learning attainments, accomplishments, proficiencies, etc. It is directly related to the pupil's growth and development in educational situations. An achievement test is a test of developed skill or knowledge. The most common type of achievement test is a standardized test developed to measure skills and knowledge learned in a given grade level, usually through planned instruction, such as training or classroom instruction. It is an important tool in school evaluation and has great significance in measuring instructional progress and progress of the students in the subject area. Tests should give an accurate picture of students' knowledge and skills in the subject area or domain being tested. Accurate achievement data are very important for planning curriculum and instruction and for program evaluation. Test scores that overestimate or underestimate students' actual knowledge and skills cannot serve these important purposes.

Definition

- "Any test that measures the attainments and accomplishments of an individual after a period of training or learning".   - NM Downie

- 'The type of ability test that describes what a person has learned to do'.   - Throndike and Hagen

- "A systematic procedure for determining the amount a student has learned through instructions."   - Groulund

Objectives

- Identify and explain reasons for performing tests.

- Understand testing terminology to communicate clearly with students and colleagues.

- Evaluate a test's validity and reliability.
- Select appropriate tests.
- Administer test protocols properly and safely.

Functions of Test

- It provides basis for promotion to the next grade.
- To find out where each student stands in various academic areas.
- It helps in determination about the placement of the students in a particular section.
- To motivate the students before a new assignment has taken up.
- To know effectively the student is performing in theory as well as in clinical areas.
- To expose pupil's difficulties which the teacher can help them to solve.

Characteristics of a Good Test

Test preparation activities which promote quality, long-term learning are appropriate, even essential. Good test-taking skills and appropriate content learning can reduce the likelihood that extraneous factors will influence students' test scores. The various characteristics of a good test are:

- It can be tried out and selected on the basis of its difficulty level and discriminating power.
- Directly related to the educational objectives.
- It should possess description of measure behavior in realistic and practical terms.
- Contains a sufficient number of test items for each measured behavior; concerned with important and useful matter; comprehensive, brief, precise and clear.
- It should be divided into different knowledge and skills according to behavior to be measured.
- Standardized the items and made instructions clear so that different users can utilize it.
- Rules and norms have to be developed so that various age groups can use at various levels.
- It provides equivalent and comparable forms of the test.
- A test manual has to be prepared, which can act as a guide for administering and scoring.

2.1.2 Purpose of Achievement Test

The purpose of achievement testing is to measure some aspect of the intellectual competence of human beings: what a person has learned to know or to do. Teachers use achievement tests to measure the attainments of their students.

Achievement testing serves many purposes:

- Assess level of competence
- Diagnose strength and weaknesses
- Assign Grades
- Achieve Certification or Promotion
- Advanced Placement/College Credit Exams
- Curriculum Evaluation
- Accountability
- Informational Purposes

2.1.3 Construction of an Achievement Test

Achievement Test

- Any test designed to assess the achievement in any subject with regard to a set of predetermined objectives

Major steps involved in the construction of achievement test

- Planning of test
- Preparation of a design for the test
- Preparation of the blue print
- Writing of items
- Preparation of the scoring key and marking scheme
- Preparation of question-wise analysis

Planning of Test

- Objective of the Test
- Determine the maximum time and maximum marks

Preparation of a design for the test

Important factors to be considered in design for the test are:

- Weightage to objectives

- Weightage to content
- Weightage to form of questions
- Weightage to difficulty level.

Weightage to Objectives

- This indicates what objectives are to be tested and what weightage has to be given to each objectives.

Sl.No.	Objectives	Marks	Percentage
1	Knowledge	3	12
2	Understanding	2	8
3	Application	6	24
4	Analysis	8	32
5	Synthesis	4	16
6	Evaluation	2	8
	Total	**25**	**100**

Weightage to Content

- This indicates the various aspects of the content to be tested and the weightage to be given to these different aspects.

Sl.No	Content	Marks	Percentage
1	Sub topic - 1	15	60
2	Sub topic - 2	10	40
	Total	**25**	**100**

Weightage to Form of Questions

- This indicates the form of the questions to be included in the test and the weightage to be given for each form of questions.

Sl.No.	Form of questions	No. of Questions	Marks	Percentage
1	Objective type	14	7	28
2	Short answer type	7	14	56
3	Essay type	1	4	16
	Total	**22**	**25**	**100**

Weightage to Difficulty Level

- This indicates the total mark and weightage to be given to different level of questions.

Sl.No.	Form of Questions	Marks	Percentage
1	Easy	5	20
2	Average	15	60
3	Difficult	5	20
	Total	25	100

Preparation of the Blueprint

- Blueprint is a three-dimensional chart giving the placement of the objectives, content and form of questions (Blueprint table on page 43).

Writing of Items

- The paper setter writes items according to the blue print.
- The difficulty level has to be considered while writing the items.
- It should also check whether all the questions included can be answered within the time allotted.
- It is advisable to arrange the questions in the order of their difficulty level.

Preparation of the scoring key and marking scheme

- In the case of objective type items where the answers are in the form of some letters or other symbol a scoring key is prepared.

Scoring Key

Q.No	Answer	Marks
1	A	½
2	C	½
3	A	½
4	D	½
5	B	½

- In the case of short answer and essay type questions, the marking scheme is prepared.
- In preparing marking scheme the examiner has to list out the value points to be credited and fix up the mark to be given to each value point.

Preparation of the Blueprint

Objectives / Form of Questions / Content	Knowledge			Understanding			Application			Analysis			Synthesis			Evaluation			Grand Total
	O	SA	E	O	SA	E	O	SE	E	O	SA	E	O	SA	E	O	SA	E	
Sub Topic-1	2(4)			1(2)			2(4)	2(1)				4(1)		2(1)			2(1)		15
Sub Topic–2	1(2)			1(2)				2(1)			4(2)			2(1)					10
Total Marks	3	0	0	2	0	0	2	4	0	0	4	4	0	4	0	0	2	0	25
Grand Total	3			2			6			8			4			2			

Note: O – Objective Type, SA – Short Answer Type, E – Essay Type

The number outside the bracket indicates the marks and those inside indicates the number of questions.

Marking Scheme

Q.No	Value points	Marks	Total Marks
1	Value Point – 1	½	2
	Value point – 2	½	
	Value point – 3	½	
	Value point – 4	½	
	Value Point – 1		
2	Value point – 2	½	2
	Value point – 3	½	
	Value point – 4	½	

Preparation of Question-wise Analysis

Question-wise Analysis

Q. No.	Content	Objectives	Form of Questions	Difficulty Level	Marks	Estimated Time (In Mts.)
1	Sub topic–1	Knowledge	Objective Type	Easy	½	1
2	Sub Topic–2	Understanding	Objective Type	Average	½	1
3	Sub Topic–2	Application	Objective Type	Easy	½	1
4	Sub Topic–1	Knowledge	Objective Type	Easy	½	1
5	Sub Topic–2	Understanding	Objective type	Average	½	1
5	Sub Topic–1	Analysis	Short answer	Average	2	3
6	Sub Topic–1	Synthesis	Short Answer	Difficult	2	3
7	Sub topic–2	Application	Short answer	Easy	2	3
8	Subtopic–1	Analysis	Essay	Average	4	10

A Guide for Writing and Improving Achievement Tests

Teresa L. Flateby, Ph.D.

University of South Florida
Tampa, FL 33620
(813) 974-2742

INTRODUCTION

Achievement tests can be written to ascertain students' level of learning within a course, in a major, or across their entire undergraduate education. For test results to be useful, they must follow basic measurement standards. In this document, procedures are presented which should help produce a valid test reflecting appropriate coverage of content and a reliable test with repeatable results. Both qualities are important for the two most widely used types of items, essay and multiple-choice, which are compared. Examples of printouts available from the Office of Evaluation and Testing are described and used to explain the item and test evaluation processes in the multiple-choice examination discussion.

Although a similar process is followed when designing a test for a single course or program, multiple faculty and a measurement consultant should be involved when designing a test to measure achievement within a major.

Fundamentals of Achievement Testing

The purposes of classroom achievement tests and their results are many and varied.

Some of the possibilities are to:

- measure an individual's achievement of course objectives
- assess the group's performance
- evaluate the test and the items
- evaluate and improve instruction and the curriculum

Always remember, however, that the fundamental purpose of achievement testing is to promote learning.

Achievement test results should accurately measure individual differences or achievement at a certain pre-specified mastery level and should always foster learning.

To accomplish these purposes, a test must be valid and reliable. Validity is addressed when a test plan is formulated to accurately represent the course content and depth of learning achieved in a course. Test results must be reliable or repeatable to be confident that a student's score is a true reflection of an examinee's achievement. When a test is constructed which closely adheres to the test plan and other guidelines presented in this manual, the likelihood of gaining repeatable test results that accurately reflect achievement of the course content is improved.

Several testing factors have been shown to contribute to learning. First, the type of test students expect guides study behaviors. If a multiple-choice test is planned, students typically will study only for recognition. If it is known that a test will emphasize factual information, a student will memorize facts, which usually are forgotten quickly. Second, test questions written above the rote level have a greater potential for promoting transfer and retention. Therefore, most tests should be written to include items to stimulate higher cognitive levels.

Essay or Multiple-Choice Achievement Test?

There have been many criticisms directed toward multiple-choice tests. It is often heard that they only measure rote learning. Moreover, they are often referred to as "multiple guess" exams. Thus, essay examinations are deemed necessary to measure cognitive levels above the simplest level of learning. However, measurement experts believe that these criticisms are unwarranted if a multiple-choice exam is well constructed.

Because the assessment of student learning requires an adequate and accurate sampling of course content, the multiple-choice test is recommended for achievement-

type testing. If the achievement test requires measuring the highest cognitive by

combining both essay and multiple-choice items into a single test is commonly

suggested. This is advocated especially if the test writer is new to achievement test

construction.

Table 1 illustrates the appropriate uses of multiple-choice and essay

examinations and the strengths and weaknesses of each.

TABLE 1

A Comparison of Essay and Multiple Choice Tests

	Essay	Multiple Choice
Recommended Uses	1. When measuring the highest cognitive levels (synthesis and evaluation levels of Bloom's Taxonomy – Appendix A). 2. When a response needs to be created. 3. When evaluating writing ability.	1. When measuring achievement at the knowledge, comprehension, application and analysis cognitive levels.
Advantages	1. Relatively short amount of time required to construct the items. 2. Allows for creativity, originality and composition.	1. Objective scoring. (Once "correct" answers are decided). 2. Evaluation of validity is possible by comparing the test to the table of item specifications. 3. Evaluation of reliability is possible. 4. Thorough sampling of course content is possible. 5. Item analysis resulting from test scores can reveal particular problems in the exam and/or in the instruction or learning.
Disadvantages	1. Objective scoring is questionable, and more difficult. 2. No generally acceptable criteria for demonstrating the validity and reliability of test. 3. Course sampling is limited. 4. Time consuming to evaluate responses.	1. Time consuming to construct. 2. Difficult to construct items at the highest cognitive levels. 3. Faculty must have some training or knowledge in test construction and item analysis techniques to write valid and reliable test.

ESSAY EXAMINATIONS

Developing Essay Items

These essay development suggestions should maximize the effectiveness of essay items for measuring achievement of course content. For guidance on assessing the quality of writing, refer to *Cognitive Level of Quality Writing Assessment: Building Better Thought Through Better Writing, 2001*.

1. Use a table of item specifications, also called a test blueprint (discussed in "How to Construct a Valid Multiple-Choice Test" section), to ensure that items are relevant and appropriate for the course content.

2. Prepare students for taking an essay exam. Provide practice in writing essay responses. Score these and give feedback to the students about their responses. Give students the grading criteria before the test.

3. Focus the questions. Be precise so that students clearly understand what is expected of them.

4. Have all students respond to the same essay questions; do not let them choose among the questions. Course content cannot be adequately sampled if students select content on which they wish to be tested. Also, students' performance cannot be compared if they are tested on different content.

5. Write more essay questions that allow for restricted responses rather than one or two essay questions that require long responses. This improves content sampling, which is especially important if the essay is being used to measure achievement rather than writing ability. Validity and reliability are improved if content is accurately and adequately sampled.

6. Students should have sufficient time to plan, prepare, and review their responses. Consider this when planning the number of essay items to include on the test.

7. Have a colleague review the questions for ambiguities.

8. Write items that measure the application, analysis, synthesis and evaluation levels of Bloom's Taxonomy of Educational Objectives. (See Appendix A.)

Scoring Essay Items

Although essay exams have been criticized for being unreliable, procedures exist which, when followed, can improve the consistency of scoring and therefore their reliability.

1. Review lecture notes and course materials before scoring students' essay responses.

2. Read each individual's response to a single item one time before scoring and before reading responses to the next item.

3. Have students sign their names on the backs of the papers so the examinees are anonymous.

4. Know what should be contained in each response before reading any papers. Specify the content to be covered. Also, determine the weight to be given to each element expected. Allow for unanticipated, but valid, responses. This is the reason for reading students' responses to an item one time before actually scoring them.

5. If achievement of the content is the sole emphasis of the course, ensure that achievement and not writing ability is being evaluated. Many measurement experts believe that sentence structure, grammar and other aspects of writing should not be considered in the scoring of a paper unless they are part of the course content. The general measurement perspective is that students should be tested only on the material taught in the course. Other educators disagree. If writing will factor into a student's grade, the importance of writing skills should be clearly emphasized before students prepare for the test.

In short, when developing essay questions for achievement exams, course content should be adequately sampled and expected responses should be specified as precisely as possible. Therefore, for testing achievement, more questions, restricted responses and specified response criteria are recommended.

MULTIPLE-CHOICE EXAMINATIONS

How to Construct Valid and Reliable Classroom Achievement Tests

Certain guidelines should be followed when developing an exam to adequately and accurately measure achievement of course content and to achieve reliable scores. Constructing a valid test which reflects course objectives is an integral part of planning the course. Thus, "teaching to the test" is desirable, even necessary. Also, confidence in test scores is imperative for assessing differences in learning, assigning grades, or for determining mastery. Instead, a test should produce repeatable scores. A 60 earned by a student will remain about 60 if a reliable test is repeated or an equivalent form of the test is given. In other words, a 60 is a close approximation of the student's "true" or theoretical score.

Achievement tests are either norm-referenced or criterion-referenced. Norm-referenced tests emphasize individual differences, how students compare with each other; criterion-referenced tests highlight how examinees' performance compares to a specific standard or level of mastery, logically or empirically determined. Identification of this standard is sometimes difficult to accomplish, especially at the more complex learning levels. Although difference in orientation exists between norm and criterion-referenced tests, Hopkins, Stanley and Hopkins (1990) maintain that all good achievement tests should be based on either explicit or implicit objectives or topics reflected in a table of item specifications. This implies that there is a great deal of overlap between the two types of tests and the development of each type begins similarly. The differences pertain mostly to the presentation and interpretation of the

results. These similarities and differences will be discussed further in the Item Analysis section.

A well-constructed test blueprint, also referred to as a table of item specifications, provides the necessary structure to foster validity. (Tables 2-5 are examples of test blueprints.) Thus, to measure achievement of a unit's, course's, or program's objectives, the test blueprint must be an accurate representation of the content and cognitive levels taught.

TABLE 2

CONTENT	TASK			TOTALS
	Knows Specific Facts	Understands Concepts	Applies Principles	
Newton's Laws of Motion	4	4	12	**20**
Types of Forces	4	2	7	**13**
Buoyancy	2	4	4	**10**
Acceleration of Gravity	2	3	5	**10**
Friction	2	2	3	**7**
TOTALS	**14**	**15**	**31**	**60**

Zimmerman, B.B., Sudweeks, R.R., Shelley, M.F., Wood, Bud, 1990.

TABLE 3

OUTCOMES CONTENT	KNOWS			Comprehends Principles	Applies Principles	Total Number of Items
	Terms	Facts	Procedures			
Role of Tests in Instruction	4	4		2		**10**
Principles of Testing	4	3	2	6	5	**20**
Norm-Referenced versus Criterion-Referenced	4	3	3			**10**
Planning the Test	3	5	5	2	5	**20**
Total Number of Items	**15**	**15**	**10**	**10**	**10**	**60**

Gronlund, 1982.

TABLE 4

MAJOR CONTENT STRATA	TAXONOMY LEVEL			TOTAL
	Knowledge	Comprehension	Application, Synthesis, etc.	
The functions of measurement in education	3	2	0	**5 (10%)**
Basic statistical concepts, central tendency and variability	1	2	2	**5 (10%)**
Norms: types, meaning, interpretation	3	3	4	**10 (20%)**
Validity: content, construction, criterion-related validity and correlation	4	6	5	**15 (30%)**
Reliability: concepts, theory, and methods of estimation	4	7	4	**15 (30%)**
TOTALS	**15 (30%)**	**20 (40%)**	**15 (30%)**	**50 (100%)**

Hopkins, Stanley, Hopkins 1990.

TABLE 5

CONTENT	LEARNING OUTCOMES			Total Number of Items
	Knowledge	Comprehension	Application and above	
Purposes of Testing	3	2		**5 (10%)**
Necessary Criteria for Tests				
Reliability	2	1		**3 (6%)**
Validity	2	2		**4 (8%)**
Test Development				
Table of Item Specifications	2	3	3	**8 (16%)**
Proper Item Construction	5	3	2	**10 (20%)**
Criteria for Evaluating Test				
Relevance, Variability, Difficulty, Discrimination, Reliability	3	2		**5 (10%)**
Item Analysis				
Principles/Printout	3	2	2	**7 (14%)**
Discrimination, Difficulty, Distractor Analysis	4	3	1	**8 (16%)**
TOTALS	**24 (48%)**	**18 (36%)**	**8 (16%)**	**50 (100%)**

Test Construction Workshop, Flateby.

To construct a test blueprint, first list the important course content, which are reflected in the syllabus and lesson plans. These will be listed on the far left column. Next, determine the cognitive levels of understanding students should achieve for each of the content areas. Bloom's Taxonomy of Educational Objectives and Cognitive Domain (1956), or a similar hierarchy, is typically used to specify the depth of learning expected. How thoroughly should students understand the material? Should they be able to recognize an appropriate step in a process (knowledge), explain a concept in their own words (comprehension), apply a principle or process to a new set of circumstances (application), compare and contrast components of schema (analysis), or create a plan to solve a problem (synthesis)? Knowledge, the foundation of Bloom's hierarchy, represents remembering or recognizing facts, followed by comprehension, the basic level of understanding. At the application level, a learner uses the content, skill or concept learned in a situation not encountered in class, the readings, or assignments. Analysis, the fourth level, requires a person to divide the material, a concept or process, into its component parts, to interrelate, compare, and contrast the parts. The fifth level, synthesis, involves the combination of components into a whole product, plan, or procedure. In the highest cognitive level, evaluation, a person judges a product or process based upon a specific set of criteria. (See Appendix A for a complete description of Bloom's Taxonomy.) The cognitive levels provide headings for the next columns. Typically, the highest cognitive levels are grouped into a single heading. (Refer to Tables 2–5.)

After the content and cognitive levels have been specified and placed on the table, determine the percentage of items to be assigned to each of the content areas

and cognitive levels. These percentages are based upon the importance of the content, the emphasis given the content in the course or program, the potential it has to increase the retention and transfer of learning, and the cognitive levels fostered in classroom assignments. Calculate the number of items to accompany the percentages based upon the total number of test items. When deciding upon the total number of test items, keep in mind that all students should have adequate time to finish the exam, but reliability is usually strengthened with well-written items. Also, allow at least one minute for each item written above the knowledge level.

Consideration of several other factors should help produce valid test results. There should be no surprises on the test. If only facts were presented in class and in the assignments, do not include analysis-type questions. Similarly, if concepts were analyzed, write an appropriate number of items requiring analysis, which could be short essay-type items. If the test blueprint is reflective of the content and cognitive levels, and is followed carefully when writing the items, a reliable assessment of students' achievement should result.

Table 5 represents a plan to measure learning from a one-day test construction and item analysis workshop. The important content topics are listed in the first column and the cognitive levels are presented in the next three columns. The Item Analysis content area (discrimination etc.), items will be written to elicit multiple cognitive levels, four items at the knowledge or factual level, three items at the basic understanding level and one item at the application level. These numbers suggest that students were given little time to apply the information presented.

To summarize the steps for constructing a test blueprint to achieve a valid achievement test:

1. List important course content or topics.

2. Identify appropriate cognitive levels using Bloom's Taxonomy of Educational Objectives for each of the course objectives.

3. Determine the number of items for
    a. the entire test
    b. each cell, i.e. course content by cognitive level.

Addressing Reliability

Reliability coefficients, appropriate for norm-referenced exams, are typically calculated by correlating or comparing two sets of scores. The most appropriate method to estimate reliability of a classroom achievement test is the Kuder-Richardson 20 (KR-20), an internal consistency measure which relates scores within one administration. Basically, the KR-20 is calculated by comparing the totals of the correct and incorrect responses for each item (the sum of the individual item variances) to the total test variance.

A reliability coefficient can range from 0.0, representing no consistency, to 1.00, representing perfect consistency. Typically, a reliability coefficient of at least .70 or higher is considered necessary to place confidence in the scores of a norm-referenced achievement test, which is critical for assigning grades. Below .70 it is less probable that scores are attributed to achievement rather than to chance or testing error. By adhering to the following guidelines, the likelihood of constructing a reliable norm-referenced test with results reflecting a normal distribution is increased.

1. Write longer tests, with well-constructed items. (Refer to Developing Items.)

2. Include items which are positive discriminators. This means that, in general, students who perform well on the exam answer the item correctly.

3. Write items which are moderately difficult. Because variation in scores contributes to reliability, very easy and very difficult items do not add to reliability as much as items of moderate difficulty and also have less potential to discriminate. However, a few easy items at the beginning of the test might build examinees' confidence. (Refer to Evaluating Tests and Items.)

Reliability coefficients calculated by correlating two sets of scores will be lower for a minimum competency criterion-referenced test than for a norm-referenced test because more students should answer items correctly, resulting in less variation. Also, these results should form a negatively skewed distribution, with most scores clustering at the high end of the distribution. Therefore, a different set of criteria is necessary to evaluate the consistency of scores for criterion-referenced exams. Hopkins, Stanley, and Hopkins (1990) recommend using the standard error of measurement as an indicator of score consistency, which is discussed in the Item Analysis section.

To summarize, the appropriate statistical method to estimate reliability and the acceptable level will vary depending upon the intent of the achievement test. If the test was developed strictly to measure individual differences, the reliability coefficient should be .70 or above. Although variation, and ultimately reliability is enhanced by including moderately difficult terms, it is acceptable to begin with a few easier items to promote confidence and to end with more difficult items to challenge the better prepared students. If a mastery-level or a minimum competency criterion-referenced test was constructed, less variability is expected because more students should answer the items correctly. Thus, a KR-20 calculated for a criterion-referenced test is expected to be weaker if mastery is achieved.

Developing Items

Before reading this section, take a moment to answer the questions in Activity 1.

Compare your answers with the correct answers provided in Appendix B.

Activity 1: Testing Your Multiple-Choice Test-Wiseness
(Adapted from Eison, 1985)

DIRECTIONS: The seven multiple-choice questions below cover historical topics that you are not likely to know. See if you are able to determine the correct answer by carefully reading each item. Please circle the correct answer for each item.

1. The Locarno pact:
    a. is an international agreement for the maintenance of peace through the guarantee of national boundaries of France, Germany, Italy, and Belgium.
    b. allowed France to occupy the Ruhr Valley.
    c. provided for the dismemberment of Austria-Hungary.
    d. provided for the protection of Red Cross bases during war times.

2. The disputed Hayes-Tilden election of 1876 was settled by an:
    a. resolution of the House of Representatives.
    b. decision of the United States Supreme Court.
    c. Electoral Commission
    d. joint resolution of Congress.

3. The august character of the work of Pericles in Athens frequently causes his work to be likened to that in Rome of:
    a. Augustus.
    b. Sulla.
    c. Pompey.
    d. Claudius.

4. The Declaration of the Rights of Man was:
    a. adopted by the French National Assembly.
    b. adopted by every Western European legislature.
    c. immediately ratified by every nation in the world.
    d. hailed by every person in England.

5. The Locarno pact:
    a. was an agreement between Greece and Turkey.
    b. gave the Tyrol to Italy.
    c. was a conspiracy to blow up the League of Nations' building at Locarno.
    d. guaranteed the boundary arrangements in Western Europe.

6. Horace in the 16th Epode emphasizes the:
    a. despair of the average man confronted by sweeping social change.
    b. elation of the average man confronted by sweeping social change.
    c. optimism of the common man about sweeping social change.
    d. all of the above.

7. About what fraction of the 1920 population of the United States was foreign-born?
    a. less than five percent.
    b. between fourteen and twenty-eight percent.
    c. twenty-five percent.
    d. between thirty and fifty percent.

Some of these items may have been answered correctly even with little or no exposure to the content. All of these "clues" should be avoided when developing items. Many "test-wise" students are able to guess the correct answers to all of the items in the exercise and would make the same guesses if the same items were administered a second time. Thus, the test would be reliable, but the results would not be valid as a measure of achievement of course content. By following the test blueprint and the guidelines offered on the next pages, the chances of achieving valid and reliable test results are increased.

General Item Writing Guidelines

An item or question contains three parts:

*   the stem, in which the question is asked or the problem is stated
*   the correct option
*   the incorrect options, also called foils or distractors.

The item should have only one correct answer and should be based upon significant information or concepts, not trivia. The item also should be clearly defined and be worded precisely without ambiguities. Remember, reading achievement is not being tested unless that is what is being taught, so be as brief and concise as possible.

To achieve a test with valid results, adhere to the test blueprint and ensure the items are written to reflect the appropriate cognitive levels. A test which promotes retention and transfer has items written at various levels in Bloom's Taxonomy (or some other defensible learning taxonomy). Be certain that items are written to accurately reflect the cognitive levels encouraged in the course for each important content area. Appendix C presents verbs which are appropriate for the various cognitive levels in Bloom's Taxonomy and may be useful when developing items. Follow the guidelines below when developing multiple-choice test items.

Write the stem:

1. as a complete sentence or question, or an incomplete statement which is completed by selecting one of the responses. It is easier to write complete statements or questions without ambiguity. Measurement experts recommend complete sentences for those new to test construction.

2. in a positive form. Negative items are easier to write and easier for students to answer. For example, "Which of the following does not promote reliability in a norm-referenced test?"

3. with a single correct answer. The stem may ask for the best answer, which elicits finer discriminations.

4. as precisely as possible. However, given a choice between a longer stem or longer options, lengthen the stem.

5. in more detail, instead of lengthening the options. When words are repeated in the options, lengthen the stem to include the repeated words. Attempt to write the stem as briefly as possible.

All options, both incorrect (called distractors) and correct, should:

- be brief.
- be grammatically consistent with the stem.
- be approximately the same length.
- be equally complex.
- cover the same type of content.
- be independent of each other.
- follow the rules of grammar.

The distractors should:

- be written for students who have a partial understanding or misunderstanding of the content.
- be plausible.
- be similar to the correct answer.

It is unnecessary to have the same number of distractors in all of the items. Stop adding distractors when they are no longer plausible. The inclusion of implausible distractors only increases the amount of reading time and does not add anything to the item or test.

The correct responses should not:

- provide clues, be longer, more technical or repeat any important words from the stem.

It is typically advised to refrain form using "all of the above" because a student is able to select the correct response from partial information. If he or she knows two answers are correct, "all of the above" will be selected even if he or she is unsure of the other options. Also, if a student knows one of the options is incorrect, then "all of the above" will not be selected. "None of the above" merely shows that a student is able to identify what is incorrect but does not provide evidence that the accurate information is known.

Evaluating Tests and Items

The following criteria adapted from Ebel and Frisbie (1991) provide a useful framework for evaluating norm-referenced achievement tests and items. If the evaluation processes or criteria are different for criterion-referenced tests, a separate description is presented.

Relevance: Does an item belong in the test? Do the items measure what the test author intended to measure? These are validity questions and address the extent to which the test blueprint was used.

Balance: Do the items adequately represent all content areas and cognitive processes specified in the test blueprint?

Efficiency: This refers to the number of items per unit of testing time. The more information about a student's achievement level obtained in a specific amount of time, the better.

Specificity: Items should be written to measure learning objectives only, not reading or writing ability, general intelligence, or test taking ability.

Difficulty

Norm-referenced: A difficulty level represents the proportion of examinees responding correctly to an item. Measurement specialists suggest an ideal mean difficulty for a norm-referenced achievement test to be halfway between a perfect score and a chance score. For example, if there are four response options, a chance score is 25% and 62.50 is the ideal average difficulty. Also, measurement experts believe that four-option multiple-choice items with difficulty levels below .5 (less than 50% passing) are too difficult. Either there is a problem with the item itself or the content is not understood. Another possibility is that students are accustomed to studying for multiple-choice tests written at the rote level, and may not be prepared for a test requiring higher cognitive levels. Thus, provide students with examples of the types of items the test will include.

Criterion-referenced: For minimum-competency criterion-referenced tests, because a large proportion of examinees should answer correctly, the same average criterion does not apply. The difficulty for a criterion-referenced test should be consistent with the logically or empirically based predetermined criterion. For example, if 80% is the criterion identified, the difficulty levels should be similar to that percentage.

Discrimination

Norm-referenced: This index shows how well items discriminate between the high and low achieving students. Discrimination indices range from −1.00 to +1.00, with a positive index indicating that students who performed well on the test tended to answer the item correctly. A negative discriminator, suggesting that the poorly performing students tended to answer the item correctly, is undesirable.

Criterion-referenced: Since the discrimination index included in the item analysis printout provided by Evaluation and Testing is based upon correlation, which is dependent upon variability, the discrimination index would be expected to be low for mastery-type items.

Variability

Norm-referenced: If grades are to be based upon a normal curve, a wide range or spread of scores is necessary. Very easy or difficult items do not contribute to variability.

Criterion-referenced: If mastery is being evaluated, wide variation in test scores is not expected or desired. The resulting distribution should be negatively skewed rather than normally distributed.

<u>Reliability</u>

Norm-referenced: This is an overall test statistic indicating score consistency, and is the single most important statistic for a norm-referenced achievement test. Reliability indices range from 0.0 to 1.00 with a .70 considered to be the minimum value acceptable. High score variability, high discrimination, and moderate difficulty levels are associated with high reliability. Although necessary if grades are to be assigned and to have confidence in the scores, reliability does not ensure validity or relevance.

Criterion-referenced: Most of the contributors to high reliability are not sought for criterion-referenced tests. It was previously suggested that the standard error of measurement could be used to evaluate the consistency of scores for either type of test and is the preferred statistic to measure stability of results from criterion-referenced tests. Bear in mind that validity, which is represented by Relevance and Balance in this discussion, is the most important criterion for either norm or criterion-referenced tests.

USE OF THE ITEM ANALYSIS TO IMPROVE
ITEMS, TESTS AND TEACHING

This section explains the information provided in the item analysis printout produced by Evaluation and Testing. Item evaluations and interpretations are presented from both the norm-referenced and criterion-referenced perspectives. Refer to Appendix D when reading.

Reading and Using the Item Analysis Printout

The item analysis printout begins with a frequency distribution, and includes the following information: the scores earned ("Score"); the frequency of each score ("Freq"); the cumulative frequency ("Cum F") starting with the lowest score and summing the frequency from the lowest to the specific score; the proportion of students earning a particular score ("PRP"); the cumulative proportion ("Cum P"); and "Z" score. Each score has been transformed into a Z, or standard score, which ranges from –3.00 to +3.00. The Z score can be compared to a standard normal curve to determine the percentile. The mean, median, or point below which 50% of the scores fall, standard deviation ("Std. Dev."), which is roughly the average variability around the mean, and the number of cases ("N") are listed at the bottom of the printout.

The actual item analysis follows the frequency distribution. The "Item No.", located in the first and last columns (Appendix D, Pg. 2) identifies the item. The second column, "Key" presents the correct response as indicated by the answer key. The headings "1-5" refer to the specific response options, both the correct response and the distractors. Numbers will be in the "other" column only if students gridded more than one option for that specific item. The "N" and "P" under each number represents the number and proportion of students selecting that particular response option. The "Pro.

Passing" column contains the difficulty value for the item, with numbers close to 1.00 indicating that the majority of students answered correctly. The "D" column lists the discrimination index, which can range from −1.00 to +1.00. Positive indices suggest that the better performing students tended to answer the item correctly.

There are also important data at the bottom of the last page: 1) "N", the number of examinees tested, 2) "Mean", the arithmetic average for the test, 3) "St. Dev.", the standard deviation, 4) "KR-20", the reliability estimate, which can range from 0.0 to +1.00, 5) "S.E.", the standard error of measurement, which estimate the error attributed to the test by comparing the reliability estimates to the standard deviation.

A student's theoretical "true score" is the summation of the observed or actual score and error score. By using these two values, one can estimate the "true score" with a certain degree of accuracy. The standard error of measurement can be related to the Z score and a student's score to determine the range of possible "true scores" with the formula: "True Score = $X \pm Z$ x S.E." For example, if the S.E. is 3 and the student's score is 50, the "true score" would fall between $50 \pm (1)$ 3, or 47-53, 68% of the time since $\pm$ (1) Z encompasses 68% of the standard normal distribution. The "true score" should be within the range of $50 \pm (2)$ 3, or 44-56, 95% of the time based on the probabilities associated with the standard normal curve, because 95% of the distribution falls within two standard deviations. The "true score" range within three standard deviations of the mean is $50 \pm (3)$ 3, or 41-59, which would capture the "true score" 99 times out of 100. This statistic and method can be used to judge the precision of scores for either criterion-referenced or norm-referenced tests.

Alternatively, the size of the standard error of measurement can be evaluated by itself to judge the precision of scores for either criterion-referenced or norm-referenced tests.

Please note that the impact of an atypical score on any statistic is greater when the statistic is calculated from a small set of numbers (n < 25) than when the statistic is calculated from a large set of numbers.

Using the Item Analysis Printout to Evaluate the Test

For an achievement test to be useful for assigning grades, the table of item specifications representing the course content and appropriate cognitive levels must be followed. The test must also produce consistent results, and therefore have adequate reliability estimates. Items and instruction can be evaluated by using the difficulty index, discrimination index, and the distractor analysis.

Begin the evaluation of the test and items by reviewing the distribution of scores. If a test has been written to identify individual differences, a normal or bell-shaped curve or a flatter version of this curve is expected. If a criterion-referenced test or one which includes mastery items has been written, a skewed distribution with the majority of scores in the upper end with fewer scores on the negative end of the distribution is anticipated. Because variability is typically lower for the criterion-referenced exam, lower reliability and discrimination indices may not suggest problems with the items on a criterion-referenced test as they would with a norm-referenced test which should have wider score variation.

The reliability coefficient (KR-20) is evaluated next, with a .70 or higher value expected if a norm-referenced test has been developed. Since mastery-type tests should have less variation, the reliability coefficient will be lower.

An inspection of the data presented in Appendix D reveals a KR-20 of .754, an acceptable statistic, but expected given the number of items on the test (n = 100). (Refer to Addressing Reliability.)

After reviewing the distribution and reliability estimate (KR-20) of the items, the difficulty index, the discrimination index and the distractors should be scrutinized.

Difficulty Index (Proportion Passing) Analysis

The difficulty index, the proportion of students answering an item correctly, can range from 0.0 to +1.00. A D = 0.0 indicates that not one student answered the item correctly and the item was very difficult, ambiguous, or miskeyed. When 100% of the examinees answer an item correctly (D = 1.00), either clues indicating the correct response were given or the content was very basic and was mastered. Measurement experts maintain that an item with a difficulty index under .50 is too difficult.   A low index may occur because the content was complex, the item was faulty, instruction was inadequate or students were not prepared. In addition to the difficulty level for each item, the average difficulty for a norm-referenced test should be located halfway between a chance score and a perfect score (100%). Thus, for a four-option test, the average difficulty should be 62.5 and is calculated in the following manner:

$$\frac{(100 - 25)}{2} = 37.5 \quad (100\% = \text{perfect score}, 25\% = \text{chance score} - 25 \text{ of } 100)$$

$$37.5 + 25 = 62.5 \quad (\text{added to the chance score})$$

The average difficulty calculated for the printout in Appendix D is approximately .71, which is higher than the recommended difficulty for a norm-referenced test. An observed difficulty that is much higher or lower than the recommended difficulty could have a varying negative impact on the reliability index because moderate difficulty adds to variability, which is a contributor to the reliability of a norm-referenced test. A very high average difficulty (very easy items) for a norm-referenced test could indicate a poorly written test, an easy test, or one which provides clues to the correct responses. However, if a criterion-referenced test has been administered, high average difficulty is expected. A higher than average difficulty may represent a test which is not strictly norm-referenced, which may or may not have been the instructor's intent. The test may be truly reflective of course content and the table of item specifications but does not have the identification of individual differences or gaining the full range of achievement as its primary purpose. Another possibility is that the instructor did not adhere to acceptable test construction procedures.

Item Discrimination Analysis

The item discrimination index compares students' performance on an item to their performance on the entire examination. The point-biserial correlation index, the method used by Evaluation and Testing to measure item discrimination, compares the performance of all students on each item to their performance on the total test. Another method of deriving an item discrimination index compares examinees' performance in the upper and lower groups on the examination (e.g. upper and lower 27%) for each item. Item discrimination indices vary from −1.00 to +1.00, with a negative index suggesting that poorly performing students on the exam answered the particular item

correctly, or conversely, high performing students answered the item incorrectly. Items should have a positive discrimination index, indicating that those who scored high on the test also tended to answer the item correctly, while those who scored low on the test tended to answer incorrectly.

A discrimination index should be evaluated with reference to the difficulty level of the item (see Item Analysis Summary), because a correlation method is used to assess the item's success in discriminating between low and high achieving students. If the items are very easy or difficult, indicating homogenous performance, there is less variation in the scores, thus resulting in a reduced potential for discrimination. For example, item #31 has a slightly negative discrimination index, but is an extremely difficult item. Therefore, at least one poorly performing student answered this very difficult item correctly. This item could be evaluated as being too difficult and having ambiguous options and should be revised. It is also possible that the correct option was miskeyed or the content was not taught as thoroughly as the instructor thought. Regardless of the reason for this result, negative discrimination indices are undesirable.

Item #8 is a negative discriminator, with nearly 20% of the students selecting the incorrect response. This could be a result of an ambiguously worded stem or options, or some students who performed poorly on the test answering this item correctly. Several items on this exam are relatively to very difficult (prop. passing < .50). In most cases, although difficult, the discrimination index (D) is relatively strong (.2 and above).

The discrimination index can be used to evaluate an item from either a criterion-referenced test or a norm-referenced achievement test, however, less variation is expected on the criterion-referenced test. Students who were successful in mastering

the material overall should answer the items correctly, resulting in a positive but lower discrimination index.

Distractor Analysis

It was mentioned in the Developing Items section that distractors must be plausible. This suggests that at least for norm-referenced tests, some students should be attracted to every distractor. Ideally, at least one person should select each one of the incorrect options. If this does not happen, the options should be reviewed for plausibility. In a criterion-referenced test, it is also necessary to be assured that the criterion has been attained. Therefore, distractors from these tests also should be plausible.

Item Analysis Summary

The three components of an item analysis are interdependent. One approach to item evaluation begins with reviewing the difficulty level. An instructor should have expectations about how students will perform on the item regardless if a norm-referenced or criterion-referenced test has been written. If the item is expected to be easy for most students, and it is not, the discrimination index should be reviewed. If the index is positive, indicating that the better performing students answered the item correctly, then the content covered on the item may not have been not fully understood by the students. This suggests that further instruction is needed, not that the item was poor. If the discrimination index is negative, the item needs revision. When the content is complex, a low difficulty index (few answering the item correctly) may be expected unless the class is composed of a homogeneous group of high achievers who have thoroughly grasped the content. If most students correctly answered an item expected

to be very difficult, the correct response option could contain clues or the distractors could be implausible. At this point, the distractors should be reviewed.

TABLE 6

Item #	Key	1		2		3		4		5		Other		Prop. Passing	Mean of Passers	D	Item #
		N	P	N	P	N	P	N	P	N	P	N	P				
1	1	21	1.00	0	.00	0	.00	0	.00	0	.00	0	.00	1.00	74.5	.000	1
3	4	2	.10	2	.10	0	.00	17	.81	0	.00	0	.00	.81	77.8	.609	3
4	2	0	.00	18	.86	3	.14	0	.00	0	.00	0	.00	.86	76.8	.510	4
8	3	0	.00	1	.05	17	.81	2	.10	1	.05	0	.00	.81	73.5	-.183	8
9	2	2	.10	11	.52	5	.24	3	.14	0	.00	0	.00	.52	77.7	.301	9
11	2	3	.14	18	.86	0	.00	0	.00	0	.00	0	.00	.86	75.9	.317	11
17	3	4	.19	5	.24	10	.48	0	.00	2	.10	0	.00	.48	75.8	.111	17
35	2	4	.19	15	.71	1	.05	0	.00	1	.05	0	.00	.71	78.0	.492	35
40	3	0	.00	1	.05	20	.95	0	.00	0	.00	0	.00	.95	74.7	.088	40
97	1	6	.29	1	.05	0	.00	10	.48	4	.19	0	.00	.29	84.0	.532	97

Another item evaluation approach begins with an analysis of the discrimination index, followed by a review of the difficulty with an and finally an inspection of the item options, if warranted. Items from Appendix D (see Table 6 above) were selected to explain this process from both the norm-referenced and criterion-referenced perspectives. Item #1 has no discrimination ability (D = 0), with 100% of the examinees answering correctly (Pro. Passing = 1.00). This indicates an easy item, either due to content mastery or clues in the correct response. The instructor was probably trying to build examinees' confidence. Since Item #4 was answered correctly by the majority of examinees and was a strong positive discriminator, only the most poorly performing students answered incorrectly. The discrimination index and proportion passing suggest the item is performing properly and is appropriate for either testing purpose, assuming the test blueprint was followed carefully. It is acceptable to begin even a

norm-referenced test with easier items. Item #40 has a low positive discrimination index and a high proportion passing. Although the discrimination potential is limited, the better students appeared to answer the item correctly. This item is satisfactory if the instructor was evaluating a rudimentary fact or concept which should be thoroughly understood. Item #35 is a strong positive discriminator with moderate difficulty. This is an excellent item for measuring individual differences. However, if these results are obtained on a mastery-type item, a higher passing rate may be expected. If a higher rate was anticipated, additional instruction may be needed. Item #97 has a positive discrimination index, but a low proportion passing. Because the item was a positive discriminator, it is probable that this item is satisfactory, although difficult. These results suggest the need for additional instruction since many students do not understand this material. Items #9 and #17 could be evaluated similarly. While being positive discriminators, approximately 50% of the students answered incorrectly, indicating they did not fully grasp the content. Alternatively, they did not expect this type of multiple choice test, and therefore did not study adequately. An inspection of the item and test blueprint should provide information regarding the plausibility of these explanations. Ideally, the proportion passing should be higher. From either a norm-referenced or criterion-referenced perspective, additional instruction may be indicated. Item #11 is a positive discriminator, with a larger proportion of students answering the item correctly than incorrectly. This is probably a good item for either testing purpose. The results from Item #3 reveal a majority of students passing and a positive discrimination index. This is probably a good item for either a criterion-referenced or norm-referenced exam. If the table of item specifications calls for an item with a large percentage of students

answering correctly, then this item is suitable for a norm-referenced test. Item #8 is a

negative discriminator, indicating the better performing students did not answer the item

correctly. This is undesirable for either a criterion-referenced or norm-referenced test.

If an item is moderately difficult, a strong and positive discrimination index is expected

from a norm-referenced test. If the discrimination index is low and the item is

moderately difficult, typically the item should be revised. Determine if the correct

response is ambiguous, if the option or answer was confusing, or if the material was not

understood.

Very easy or very difficult items are not expected to have high discrimination

indices, since the discrimination index is based upon correlation. Because a strong

correlation requires variation in the distribution, with little variation the potential for

discrimination is reduced. Thus, when the discrimination index is low and the item is

very easy or difficult, there is less concern than when an item with a low discrimination

index (index close to 0.0) or moderate difficulty are observed.

To summarize, each item should be evaluated within the framework of the type of

test, norm-referenced or criterion-referenced. A general level of difficulty should be

expected for each item and should be compared to the observed difficulty. Positive

discrimination indices (.2 and above) are desired. If a problem is uncovered, evaluate

the options and decide if the item and/or the instruction need revision.

Synthesis

There has been considerable discussion about the purposes and interpretations

of criterion-referenced and norm-referenced achievement examinations. There are

commonalties in the two evaluation approaches. In fact, an inspection of the

development stages may not reveal which evaluation approach has been used, because both should begin with a table of item specifications, which reflects the course content and learning outcomes. The two types of tests have different purposes; the criterion-referenced approach evaluates how well each student performs in relation to a predetermined criterion and the norm-referenced approach compares students in relation. The appropriate approach to testing should be based upon instructional goals. If an exact criterion can be established, for example, if 80% of the items on the test represent mastery of specific content, and the achievement of this criterion is critical, then criterion-referenced testing is the appropriate vehicle to use. If students will be ranked based upon their achievement of content and learning objectives, then a norm-referenced test which emphasizes items that maximize these differences is appropriate.

To develop a norm-referenced test which will reliably identify individual differences, construct and adhere to a table of item specifications and write items that are moderately difficult and positive discriminators. However, if content is covered that is fundamental, the difficulty index should be high (majority of students responded correctly), and the discrimination index should be positive, but may be low. If testing a homogeneous group of high achievers, such as an advanced graduate class, the range of performance on any item or the entire test should be restricted, with a large proportion of students passing each item. For this class, relatively low discrimination indices may result. While it may be necessary to measure individual differences in student achievement for this group of students, the items and scores may reflect a distribution more typical of a criterion-referenced exam than of a norm-referenced exam.

To summarize, the main differences between criterion-referenced and norm-referenced tests occur at the criterion-setting point and at the item analysis, test statistic review point. The table of item specifications guides the development of both the norm-referenced and criterion-referenced tests.

APPENDIX A
Condensed Version of the Taxonomy of Educational Objectives

Cognitive Domain

KNOWLEDGE

1.00 KNOWLEDGE

Knowledge, as defined here, involves the recall of specifics and universals, the recall of methods and processes, or the recall of a pattern, structure, or setting. For measurement purposes, the recall situation involves little more than bringing to mind the appropriate material. Although some alteration of the material may be required, this is a relatively minor part of the task. The knowledge objectives emphasize most the psychological processes of remembering. The process of relating is also involved in that a knowledge test situation requires the organization and reorganization of a problem such that it will furnish the appropriate signals and cues for the information and knowledge the individual possesses. To use an analogy, if one thinks of the mind as a file, the problem in a knowledge test situation is that of finding in the problem or task the appropriate signals, cues, and clues which will most effectively bring out whatever knowledge is filed or stored.

1.10 KNOWLEDGE OF SPECIFICS

The recall of specific and isolated bits of information. The emphasis is on symbols with concrete referents. This material, which is at a very low level of abstraction, may be thought of as the elements from which more complex and abstract forms of knowledge are built.

1.11 KNOWLEDGE OF TERMINOLOGY

Knowledge of the referents for specific symbols (verbal and non-verbal). This may include knowledge of the most generally accepted symbol referent, knowledge of the variety of symbols which may be used for a single referent, or knowledge of the referent most appropriate to a given use of a symbol.

- To define technical terms by giving their attributes, properties, or relations.
- Familiarity with a large number of words in their common range of meanings.

1.12 KNOWLEDGE OF SPECIFIC FACTS

Knowledge of dates – events, persons, places, etc. This may include very precise and specific information such as the specific date or exact magnitude of a phenomenon. It may also include approximate or relative information such as an approximate time period or the general order of magnitude of a phenomenon.

- The recall of major facts about particular cultures.
- The possession of a minimum knowledge about the organisms studied in the laboratory.
- Illustrative education objectives selected from the literature.

1.20 KNOWLEDGE OF WAYS AND MEANS OF DEALING WITH SPECIFICS

Knowledge of the ways of organizing, studying, judging, and criticizing. This includes the methods of inquiry, the chronological sequences, and the standards of judgement within a field as well as the patterns of organization through which the areas of the fields themselves are determined and internally organized. This knowledge is at an intermediate level of abstraction between specific knowledge on the one hand and knowledge of universals on the other. It does not so much demand the activity of the student in using the materials as it does a more passive awareness of their nature.

1.21 KNOWLEDGE OF CONVENTIONS

Knowledge of characteristic ways of treating and presenting ideas and

phenomena. For purposes of communication and consistency, workers in a field

employ usages, styles, practices and forms which best suit their purposes and/or which

appear to suit best the phenomena with which they deal. It should be recognized that

although these forms and conventions are likely to be set up on arbitrary, accidental, or

authoritative bases, they are retained because of the general agreement or concurrence

of individuals concerned with the subject, phenomena, or problem.

- Familiarity with the forms and conventions of the major types of works, e.g., verse, plays, scientific papers, etc.
- To make pupils conscious of correct form and usage in speech and writing.

1.22 KNOWLEDGE OF TRENDS AND SEQUENCES

Knowledge of the processes, directions, and movements of phenomena with

respect to time.

- Understanding of the continuity and development of American culture as exemplified in American life.
- Knowledge of the basic trends underlying the development of public assistance programs.

1.23 KNOWLEDGE OF CLASSIFICATIONS AND CATEGORIES

Knowledge of the classes, sets, divisions, and arrangements which are regarded

as fundamental for a given subject field, purpose, argument, or problem.

- To recognize the area encompassed by various kinds of problems or materials.
- Becoming familiar with a range of types of literature.

1.24 KNOWLEDGE OF CRITERIA

Knowledge of the criteria by which facts, principles, opinions, and conduct are tested or judged.

- Familiarity with criteria for judgement appropriate to the type of work and the purpose for which it is read.
- Knowledge of criteria for the evaluation of recreational activities.

1.25 KNOWLEDGE OF THE METHODOLOGY

Knowledge of the methods of inquiry, techniques, and procedures employed in a particular subject field as well as those employed in investigating particular problems and phenomena. The emphasis here is on the individual's knowledge of the method rather than his ability to use the method.

- Knowledge of scientific methods for evaluating health concepts.
- The student shall know the methods of attack relevant to the kinds of problems of concern to the social sciences.

1.30 KNOWLEDGE OF THE UNIVERSALS AND ABSTRACTIONS IN A FIELD

Knowledge of the major schemes and patterns by which phenomena and ideas are organized. These are the large structures, theories, and generalizations which dominate a subject field or which are quite generally used in studying phenomena or solving problems. These are at the highest levels of abstraction and complexity.

1.31 KNOWLEDGE OF PRINCIPLES AND GENERALIZATIONS

Knowledge of particular abstractions which summarize observations of phenomena. These are the abstractions which are of value in explaining, describing, predicting, or in determining the most appropriate and relevant action to be taken.

- Knowledge of the important principles by which our experience with biological phenomena is summarized.
- The recall of major generalizations about particular cultures.

1.32 KNOWLEDGE OF THEORIES AND STRUCTURES

Knowledge of the body of principles and generalizations together with their interrelations which present a clear, rounded, and systematic view of a complex phenomenon, problem, or field. These are the most abstract formulations, and they can be used to show the interrelation and organization of a great range of specifics.

- The recall of major theories about particular cultures.
- Knowledge of a relatively complete formulation of the theory of evolution.

INTELLECTUAL ABILITIES AND SKILLS

Abilities and skills refer to organized modes of operation and generalized techniques for dealing with materials and problems. The materials and problems may be of such a nature that little or no specialized and technical information is required. Such information as is required can be assumed to be part of the individual's general fund of knowledge. Other problems may require specialized and technical information at a rather high level such that specific knowledge and skill in dealing with the problem and the materials are required. The abilities and skills objectives emphasize the mental processes of organizing and reorganizing material to achieve a particular purpose. The materials may be given or remembered.

2.00 COMPREHENSION

This represents the lowest level of understanding. It refers to a type of understanding or comprehension such that the individual knows what is being communicated and can make use of the material or idea being communicated without necessarily relating it to other material or seeing its fullest implications.

2.10 TRANSLATION

Comprehension as evidenced by the care and accuracy with which the

communication is paraphrased or rendered from one language or form of

communication to another. Translation is judged on the basis of faithfulness and

accuracy, that is, on the extent to which the material in the original communication is

preserved although the form of the communication has been altered.

- The ability to understand non-literal statements (metaphor, symbolism, irony, exaggeration).
- Skill in translating mathematical verbal material into symbolic statements and vice versa.

2.20 INTERPRETATION

The explanation or summarization of a communication. Whereas translation

involves an objective part-for-part rendering of a communication, interpretation involves

a reordering, rearrangement, or a new view of the material.

- The ability to grasp the thought of the work as a whole at any desired level of generality.
- The ability to interpret various types of social data.

2.30 EXTRAPOLATION

The extension of trends or tendencies beyond the given data to determine

implications, consequences, corollaries, effects, etc.; which are in accordance with the

conditions described in the original communication.

- The ability to deal with the conclusions of a work in terms of the immediate inference made from the explicit statements.
- Skill in predicting continuation of trends.

3.00 APPLICATION

The use of abstractions in particular and concrete situations. The abstractions

may be in the form of general ideas, rule or procedures, or generalized methods. The

abstractions may also be technical principles, ideas, and theories which must be

remembered and applied.

- Application to the phenomena discussed in one paper of the scientific terms or concepts used in other papers.
- The ability to predict the probable effect of a change in a factor on a biological situation previously at equilibrium.

4.00 ANALYSIS

The breakdown of a communication into its constituent elements or parts such

that the relative hierarchy of ideas is made clear and/or the relations between the ideas

expressed are made explicit. Such analyses are intended to clarify the communication,

to indicate how the communication is organized, and the way in which it manages to

convey its effects, as well as its basis and arrangement.

4.10 ANALYSIS OF ELEMENTS

Identification of the elements included in a communication.

- The ability to recognize unstated assumptions.
- Skills in distinguishing facts from hypotheses.

4.20 ANALYSIS OF RELATIONSHIPS

The connections and interactions between elements and parts of a

communication.

- Ability to check the consistency of hypotheses with given information and assumptions.
- Skill in comprehending the interrelationships among the ideas in a passage.

4.30 ANALYSIS OF ORGANIZATIONAL PRINCIPLES

The organization, systematic arrangement, and structure which hold the

communication together. This includes the "explicit" as well as "implicit" structure. It

includes the bases, necessary arrangement, and the mechanics which make the communication a unit.

- The ability to recognize form and pattern in literary or artistic works as a means of understanding their meaning.
- Ability to recognize the general techniques used in persuasive materials, such as advertising, propaganda, etc.

5.00 SYNTHESIS

The putting together of elements and parts so as to form a whole. This involves the process of working with pieces, parts, elements, etc., and arranging and combining them in such a way as to constitute a patter or structure not clearly there before.

5.10 PRODUCTION OF A UNIQUE COMMUNICATION

The development of a communication in which the writer or speaker attempts to convey ideas, feelings, and/or experiences to others.

- Skill in writing, using an excellent organization of ideas and statements.
- Ability to tell a personal experience effectively.

5.20 PRODUCTION OF A PLAN, OR PROPOSED SET OF OPERATIONS

The development of a plan of work or the proposal of a plan of operations. The plan should satisfy requirements of the task which may be given to the student or which he may develop for himself.

- Ability to propose ways of testing hypotheses.
- Ability to plan a unit of instruction for a particular teaching situation.

5.30 DERIVATION OF A SET OF ABSTRACT RELATIONS

The development of a set of abstract relations either to classify or explain particular data or phenomena, or the deduction of propositions and relations from a set of basic propositions or symbolic representations.

- Ability to formulate appropriate hypotheses based upon an analysis of factors involved, and to modify such hypotheses in the light of new factors and considerations.
- Ability to make mathematical discoveries and generalizations.

6.00 EVALUATION

Judgements about the value of material and methods for given purposes.

Quantitative and qualitative judgements about the extent to which material and methods satisfy criteria. Use of a standard of appraisal. The criteria may be arranging and combining them in such a way as to constitute a patter or structure not clearly there before.

6.10 JUDGEMENTS IN TERMS OF INTERNAL EVIDENCE

Evaluation of the accuracy of a communication from such evidence as logical accuracy, consistency, and other internal criteria.

- Judging by internal standards, the ability to assess general probability of accuracy in reporting facts from the care given to exactness of statement, documentation, proof, etc.
- The ability to indicate logical fallacies in arguments.

6.20 JUDGEMENTS IN TERMS OF EXTERNAL CRITERIA

Evaluation of material with reference to selected or remembered criteria.

- The comparison of major theories, generalizations, and facts about particular cultures.
- Judging by external standards, the ability to compare a work with the highest known standards in its field – especially with other works of recognized excellence.

APPENDIX B
Correct Responses and Clues for Activity 1

Question	Correct Response	Clue
1	A	This option is much longer than the other options. Also, a pact is a formal agreement usually between nations.
2	C	This option is the only option grammatically correct with the stem.
3	A	This option repeats information in the stem.
4	A	This is the only option without the word "every". Typically "every" and other absolute words are incorrect.
5	D	Question 1 provides information about the stem.
6	A	Options B and C are similar and A is dissimilar, making Option D incorrect. Also, B and C are incorrect because they are similar.
7	C	This is the only option that is specific.

APPENDIX C
Bloom's Taxonomy of Educational Objectives

I.   KNOWLEDGE
    A. Remembers by recall or recognition; should not be too different from way in which knowledge was originally learned.
    B. Behavior tasks

1. Defines…	5. Recites…
2. Recalls…	6. Names…
3. Lists…	7. Describes…
4. States…	8. Selects…

II. COMPREHENSION
    A. Grasps the meaning of the material; deals with the content. Requires interpretation or translation from abstract to simple phrases or making generalizations.
    B. Behavioral tasks

1. States in own words…	6. Classifies…
2. Gives an example of…	7. Explains…
3. Illustrates…	8. Predicts…
4. Summarizes… between…	9. Distinguishes
5. Interprets…	

III. APPLICATION
    A. Uses information in real-life problems.
    B. Behavioral tasks
        1. Chooses appropriate procedure…
        2. Applies a principle…
        3. Uses an approach…
        4. Solves a problem…
        5. Computes…
        6. Relates…
        7. Demonstrates…

IV. ANALYSIS
    A. Breaks material into constituent parts and identifies relationships of the parts to each other and to the whole.
    B. Distinguishes fact from hypothesis and from value statements.
    C. Identifies conclusions and generalizations.
    D. Separates relevant from trivia.
    E. Differentiates one symbol from another symbol.
    F. Behavioral tasks

1. Distinguishes…	5. Differentiates
2. Discriminates between…	6. Infers…

3. Discovers…   7. Subdivides…
4. Detects…

V.   SYNTHESIS
      A. Combines parts to make a whole.
      B. Emphasizes originality.
      C. Organizes ideas into new patterns.
      D. Tries various approaches.
      E. Ability to use results of research in solving a problem.
      F. Behavioral tasks
            1. Develops…   4. Creates…
            2. Writes…   5. Combines…
            3. Designs…   6. Composes…

VI. EVALUATION
      A. Makes a judgement concerning the value of ideas, principles, methods,
          solutions, etc.
      B. Uses set criteria.
      C. Not opinions.
      D. Recognizes fallacies.
      E. Behavior tasks
            1. Compares…   5. Concludes…
            2. Judges…   6. Discriminates…
            3. Determines the best possible…   7. Supports…
            4. Applies criteria…

BIBLIOGRAPHY

Bloom, Benjamin S. 1956 Ed., <u>Taxonomy of Educational Objectives</u>, David McKay
      Company, Inc., New York.

Carey, Lou M. 1988, <u>Measuring and Evaluating School Learning</u>, Allyn and Bacon, Inc.,
      Newton, MA.

Crocker, Linda and Algina, James. 1986, <u>Introduction to Classical and Modern Test
      Theory</u>, CBS College Publishing, New York.

Ebel, Robert L. and Frisbie, David A. 1991, <u>Essentials of Educational Measurement</u>, 5th
      ed., Prentice-Hall, Englewood Cliffs, NJ.

Ebel, Robert L. 1980, <u>Practical Problems in Educational Measurement</u>, D.C. Health and
      Co., Lexington, MA.

Eison, James. 1985, <u>Constructing Classroom Tests: Why and How</u>, Southeast Missouri
      State University, Cape Girardeau, MO.

Gronlund, Norman E. 1982, <u>Constructing Achievement Tests</u>, 3rd ed., Prentice-Hall,
      Inc., Englewood Cliffs, NJ.

Gronlund, Norman E. 1993, <u>How to Make Achievement Tests and Assessments</u>, 5th
      ed., Allyn and Bacon, Needham Heights, MA.

Hall, Bruce. 1993, <u>Handouts distributed in workshops</u>, University of South Florida,
      Tampa, FL. Permission granted to use materials.

Hills, John R. 1976, <u>Measurement and Evaluation in the Classroom</u>, Charles E. Merrill
      Publishing Co., Columbus, OH.

Hopkins, K.D., Stanley, J.C. and Hopkins, B.R. 1990, <u>Educational and Psychological
      Measurement and Evaluation</u>, 7th ed., Prentice-Hall, Englewood Cliffs, NJ.

Sax, Gilbert. 1989, <u>Principles of Educational and Psychological Measurement and
      Evaluation</u>, 3rd ed., Wadsworth, Inc., Belmont, CA.

Zimmerman, Beverly B., Sudweeks, R.R., Shelley, M.F., Wood, Bud. 1990, <u>How to
      Prepare Better Tests: Guidelines for University Faculty</u>, Brigham Young
      University Testing Services and the Department of Instructional Science, Provo,
      UT.

UNIT 8   ACHIEVEMENT TESTS

Structure

8.1   INTRODUCTION

As a teacher one is involved directly in the evaluation of the learner. The theory that you have learnt will have to be applied by you as a teacher in the classroom situation. The present Block 'Learner's Evaluation' in general and this unit in particular is concerned with this very important activity of teachers.

Teachers teach and help the learners to learn. The learning that takes place is assessed or evaluated not only for the learner's benefit but also for the teacher to evaluate his/her own work. At the end of a lesson or a group of lessons, the teacher needs to get feedback on what the learner has achieved, as a result of the teacher's efforts and also, indirectly to assess his/her own achievement as a teacher. This feedback comes with the help of a tool, generally an achievement test. An achievement test is designed to evaluate a unit during the teaching-learning process. The unit of teaching-learning may be, as has already been mentioned, one lesson or a group of lessons transacted in a particular time period. You have already read about achievement tests in the previous unit. In this unit we will discuss the same in detail.

8.2 OBJECTIVES

After going through this unit you will be able to :

- discuss the purpose of achievement tests,
- describe the steps involved in constructing an achievement test,
- explain/illustrate how the design and blueprint of an achievement test are prepared,
- write a variety of questions - objective, short answer and essay,
- prepare a sample achievement test with a marking scheme,
- describe how an achievement test should be administered,
- mark/score an achievement test and interpret test scores, and
- discuss the different types of grading and their purpose.

8.3 PURPOSE OF ACHIEVEMENT TESTS

Achievement tests are universally used in the classroom mainly for the following purposes :

1. To measure whether students possess the pre-requisite skills needed to succeed in any unit or whether the students have achieved the objective of the planned instruction.
2. To monitor students' learning and to provide ongoing feedback to both students and teachers during the teaching-learning process.
3. To identify the students' learning difficulties - whether persistent or recurring.
4. To assign grades.

8.4 CONSTRUCTION OF ACHIEVEMENT TESTS

There are several steps involved in the construction of Achievement Tests. We will now discuss these in detail one by one.

8.4.1 Instructional Objectives

The first and the most important step in planning a test is to identify the instructional objectives. Each subject has a different set of instructional objectives. In the subjects of Science, Social Sciences, and Mathematics the major objectives are categorised as knowledge, understanding, application and skill, while in Languages the major objectives are categorised as knowledge, comprehension and expression. Knowledge objective is considered to be the lowest level of learning whereas understanding, application of knowledge in sciences or behavioural sciences are considered higher levels of learning. You have already read about this in detail earlier in unit 3.

8.4.2 Design

The second step in planning a test is to make the "Design". The Design specifies weightages to different (a) instructional objectives, (b) types (or forms) of questions, (c) units and sub-units of the course content, (d) levels of difficulty. It also indicates as to whether there are any options in the question paper, and if so, what their nature is.

The Design, in fact, is termed as an instrument which reflects major policy decisions of the examining agency, whether it is a Board or an individual. A sample format for presenting design of a test is given on the next page.

8.4.3 Blueprint

The third step is to prepare the "Blueprint". The policy decisions, as reflected in the design of the question paper, are translated into action through the Blueprint. It is at this stage that the paper setter decides as to how many question are to be set for different objectives. Further he/she decides under which unit/topic a particular question is to be set. Further more, he/she picks up various forms of questions. Thereafter, the paper setter decides how all the questions are to be distributed over different objectives and content areas so as to obtain the weightages

decided in the Design. The three dimensions of the blueprint consist of content areas in horizontal rows and objectives and forms of questions in vertical columns. Once the blueprint is prepared, the paper setter can write/select the items and prepare the question paper. A sample format of Blueprint is given on the next page :

DESIGN

SUBJECT :

CLASS :

THE WEIGHTAGE OF THE DISTRIBUTION OF MARKS OVER THE DIFFERENT DIMENSIONS OF THE QUESTION PAPER IS/SHALL BE AS FOLLOWS :

1. WEIGHTAGE TO INSTRUCTIONAL OBJECTIVES/LEARNING OUTCOMES

S. No.	OBJECTIVES	MARKS	% AGE OF MARKS
(1)	KNOWLEDGE		
(2)	UNDERSTANDING		
(3)	APPLICATION		
(4)	SKILL		
	TOTAL		

2. WEIGHTAGE TO CONTENT/SUBJECT UNITS :

S.NO.	UNITS & THEIR SUB-UNITS	MARKS	UNITS & THEIR SUB-UNITS	MARKS
(1)				
(2)				
(3)				
(4)				
(5)				
(6)				

3. WEIGHTAGE TO TYPES/FORMS OF QUESTIONS

S.NO.	FORMS OF QUESTIONS	MARKS FOR EACH	NUMBER OF QUESTIONS	TOTAL MARKS
(1)	L.A.			
(2)	S.A.			
(3)	V.S.A.			

*NOTE: THE EXPECTED LENGTH OF THE ANSWERS OF DIFFERENT TYPES OF QUESTIONS WOULD BE AS FOLLOWS.

THIS IS ONLY AN APPROXIMATION. THE ACTUAL LENGTH, HOWEVER, MAY VARY. AS THE TOTAL TIME IS CALCULATED ON THE BASIS OF THE NUMBER OF QUESTIONS REQUIRED TO BE ANSWERED AND THE LENGTH OF THEIR ANTICIPATED ANSWERS. IT WOULD, THEREFORE, BE ADVISABLE, TO BUDGET TIME PROPERLY BY CUTTING OUT THE SUPERFLUOUS LENGTH AND BE WITHIN THE EXPECTED LIMITED.

S.NO.	TYPE/FORMS OF QUESTIONS	MARKS	EXPECTED LENGTH (NO. OF WORDS/SENTENCES)	EXPECTED TIME FOR EACH QUESTION (MINUTES)
(1)	L.A.			
(2)	S.A.			
(3)	V.S.A.			

L.A.	=	(LONG ANSWER)	DIFFICULTY LEVEL :	EASY
S.A.	=	(SHORT ANSWER)	(GIVE	AVERAGE
V.S.A.	=	(VERY SHORT ANSWER)	PERCENTAGE)	DIFFICULT

*NOTE : INTERNAL OPTIONS IN L.A. QUESTIONS ONLY.

BLUE PRINT

EXAM :

SUBJECT :   PAPER :

UNIT :   CLASS :

MAXIMUM MARKS :   TIME :

OBJECTIVES	KNOWLEDGE			UNDERSTANDING			APPLICATION			SKILL			TOTAL		
FORMS OF QUESTION/ CONTENT UNIT	E	SA	VSA	E	SA	VSA	E	SA	VSA	E	SA	VSA	E	SA	VSA
1.															
2.															
3.															
4.															
5.															
6.															
SUB TOTAL															
TOTAL															

NOTE : PLEASE PUT THE NUMBER OF QUESTION WITHIN BRACKETS AND THE MARKS OUTSIDE THE BRACKETS.

SUMMARY

EASSY OR LONG ANSWER   (LA)   MARKS :

SHORT ANSWER   (SA)   MARKS :

VERY SHORT ANSWER   (VSA)   MARKS :

SCHEME OF OPTIONS :

SCHEME OF SECTIONS :

DELETE WHICHEVER IS NOT APPLICABLE

8.4.4 Writing of Questions

The next step after the finalization of the blueprint is writing appropriate questions in accordance with the broad parameters set out in the blueprint. One should take one small block of the blueprint at a time and write out the required questions. Thus, for each block of blueprint which is filled in, questions have got to be written one by one. Once it is done, we have all the questions meeting the necessary requirements laid down in the blueprint. While selecting each small block for writing a question, you can proceed in several ways.

a) either writing all questions (one by one) belonging to one objective at a time i.e. knowledge or understanding or application followed by other objectives, or

b) by taking up questions according to their form or type i.e. Essay Type followed by Short Answer and Very Short Answer Type or in any other order, or

c) by writing questions for one unit of the syllabus or portion to be covered by the test at a time.

Each approach has its advantages and disadvantages, too. Irrespective of the method followed, the questions then have to be arranged in a logical sequence.

8.4.5 Marking Scheme

The fifth step is to prepare the "Marking Scheme". The marking scheme helps prevent inconsistency in judgement. In the marking scheme, possible responses to items in the test are structured. The various value points for response are graded and the marks allowed to each value point indicated. The marking scheme ensures objectivity in judgement and eliminates differences in score which may be due to idiosyncrasies of the evaluator. The marking scheme, of course, includes the scoring key, which is prepared in respect of objective type questions. Let us discuss this in detail.

Apart from the quality of the question paper, reliability of assessment, to a great extent, depends on the degree of consistency of scores assigned to the students by different examiners or by the same examiner on two different occassions. Thus, variation can occur because of any one of two different reasons :

i) Due to inconsistency of the same examiner when he/she examines different answer scripts adopting different standards.

ii) Due to different examiners using different standards of judgement.

If an answer script is awarded the same grade or marks on repeated exposure to the same examiner, the examiner is said to be consistent in awarding the marks. As such, the assessment done by him/her could be said to be more reliable and consistent than the other examiner in whose case variation in award of marks is higher.

The factor contributing to variations in the standards of assessment, both at the intra-and the inter-examiner levels, can be controlled by supplying a detailed scheme of marking along with the expected answers so that every examiner may interpret the questions in the same way and attain the same standard of marking without being too lenient or strict or varying in his/her assessment. Subjectivity, is thus minimised and it is believed to give a more reliable picture of the students' performance.

Highlights of a good marking scheme

1) It is a three column statement showing serial number of the questions, their expected outline answers and the marks allotted to each value point under them.

2) In respect of long answer or essay type questions, the expected outline answers should :

   i) be complete and cover all possible or major areas as demanded by the questions

   ii) clearly indicate each expected point or the parts under the outlined major areas

   iii) provide direction as to whether all points will count towards a complete or correct answer or a set of points will be adequate enough for full credit (All this should be clearly reflected), and

   iv) indicate marks for each expected point. Marks so distributed over expected points or their sets should be equal to the total marks assigned for a question.

3) In respect of short answer questions a complete answer may be provided with its break-ups where ever necessary along with the break-up of marks.

4) Out of the total marks assigned for a question, each point so enumerated/explained may be assigned marks according to their significance in the answer.

5) In some situations, apart from the content, other qualities of answer may also matter significantly, particularly in long answer or essay type questions. These could be logical approach, coherence, lucidity of expression, the style of presentation etc. Some marks may also be set apart for such overall quality of answer which cannot be usually covered in enumeration of the content points.

6) The scheme of marking needs to be comprehensive enough not to leave any point unexpected and thus should provide clear guidelines in respect of the break-up of marks over different points or parts of the answer.

7) If a question entails some other points beyond one's expectation, a provision may also be made to take them into account and suitably reward them.

8.4.6 Question-wise Analysis

The sixth and the last step is that of question-wise analysis. Such an exercise helps the paper setter to ensure that there is no imbalance in the question paper. During question-wise analysis, the paper setter analyses each question on various parameters stated in the blueprint.

Check Your Progress 1

1) List the steps involved in constructing an achievement test.

....

....

....

iii) Describe the purpose that the marking scheme serves.

8.5   TYPES OF QUESTIONS

There are mainly three kinds of questions – essay, short answer and objective type. We will learn about the essay type questions in detail in Unit 9. Here, we will discuss the other two in detail.

8.5.1   Short Answer Questions

Short answer questions generally require exact answers and, although taking many forms, they share the following distinctive features.

i)   They usually take less than five minutes to read and answer, many take less than a minute.

ii)   They include some guidance on the extent of the answer required e.g. the size of answer, space or specific instruction such as "In not more than 20 words..."

iii)   The answer is supplied by the pupil, not pre-selected as in objective questions.

They can be grouped into two broad categories :

a)   extended answer

b)   insert and completion

8.5.1.1   Extended Answer Type

The extended answer version includes questions which require pupils to write a brief description, draw a map, make a list, perform a calculation, translate a sentence, write down a definition or formula and so on. They are probably the commonest form of questions used in schools and are frequently used by examining Boards. They are deceptively easy to set and usually difficult to mark with any degree of speed and consistency.

Examples

1.   Give the titles of two novels of Prem Chand and for each the names of three of the principal characters.

    a)   Title   :   _____

         Characters : i) _____ ii) _____

                      iii) _____

    b)   Title   :   _____

         Characters : i) _____ ii) _____

                      iii) _____

2.   Describe briefly two factors which contributed to the outbreak of the First World War. Do not write more than thirty words about each factor.

3.   Name the two types of chemical compound which combine to form salts, and then name one inorganic and one organic salt.

4.   Make diagrams to show the structure of

    i)   an atom

    ii)   a human cell

    iii)   Methyl alcohol

5.   Write two separate sentences, one containing the word 'eminent' and the other containing 'imminent'.

6.   Construct

    a)   An equilateral triangle with the sides measuring 4 cm.

b) A square with 5 cm side.

c) Find out the difference between the areas of the triangle and the square you have constructed.

8.5.1.2 Completion Type

The commonest form of **completion** questions is one where the pupil is required to add one or two words to complete an incomplete statement correctly. Where the missing words are in the body of the statement to be completed it is usually called an insert type. A **completion type** is where the words are required at the end of the statement. The use of insert or completion questions is not, however, limited to written statements and can be used to prepare extremely good questions based on incomplete maps, drawings, diagrams, formulae, calculations, and the like.

Examples

1. In the human eye light enters the (I)....., which is surrounded by the part called the (2) ....., As the amount of light increases this part (3) ...., but (4).....again when the amount of light decreases. On reaching the (5).....at the back of the eye the light stimulates two types of nerve cell called (6) r......and (7) c......

    (1) .... (2) .... (3) .... (4) ....

    (5) .... (6) .... (7) ....

2. Complete the missing words in this paragraph.

    That night there was so little hotel a ...tion that they had to take an expensive ......of rooms. After paying the bill they were almost p......less.

    (A useful technique for testing vocabulary and spelling).

3. Complete the following formulae :

    Ammonia   :   N.........

    Sulphuric acid   :   H.......

    Sodium carbonate :   ...... CO

    (Incompleteness of formulae can be adjusted in accordance with what is to be tested.)

4. Complete the expansion by filling in the blanks :

    $$(a+b)^2 = a^2+b^2+2ab$$

8.5.2 Objective Type Questions

What is an objective question? Simply, an objective question is one which is free from any subjective bias – either from the tester or the marker. Confusingly, in educational jargon, the adjective 'objective' usually means 'not subjective' while the noun 'objective' usually means an aim, a goal, target or intention. This sub-section is not about course objectives-aims, intended learning outcomes, etc. – but about testing which is free from subjective elements. There can only be one right or objective answer to an objective question. Objective questions can take various forms, but invariably they require brief answers with little or no writing. A simple tick or a quick oral answer may be enough.

8.5.2.1 Simple Recall

The most common used objective type question by teachers as part of their day-to-day teaching is simple recall. The teacher asks a short question, expecting a quick one-word answer or a simple statement completed. Let us have a look at some examples.

a) Direct question – After which battle was the Mughal empire established in India?

    Expected answer – First Battle of Panipat.

b) Incomplete statement – A writer called Jane .... wrote
                                          Pride and Prejudice.

    Expected answer – Austen

This, of course, is the kind we have discussed under short answer questions also. There is a definite overlapping.

Rather more complex than tests for simple recall are the types of questions which can be

grouped under the heading : Choice-Response. These may be sub-divided into the three varieties : multiple choice, true/false, matching block. We will now discuss these one by one.

8.5.2.2 Multiple Choice

A multiple choice-item consists of three parts – a stem, a key and a number of distractors. The key and distractors together are often referred to as options. The stem can be either a direct question or an incomplete statement; the key is the correct answer and the distractors are plausible but incorrect answers. Some examples of multiple choice-items are shown below, with the first one labelled to identify the components.

Examples

a) **Direct Question**

STEM		What is the essential characteristic of an objective item?
DISTRACTOR	A	It is written to test a specified learning outcome.
KEY	B	No subjective judgement is required in its marking.
DISTRACTOR	C	It is based on a verifiable fact, problem or principle.
DISTRACTOR	D	Its subject matter and wording are unambiguous.

b) **Incomplete Statement**

   i) Objective Questions are so called because :

   A) They are written to test specified learning outcomes.

   B) No subjective judgement is required to mark them.

   C) They are based on verifiable facts or principles.

   D) Their content is chosen objectively.

   ii) His wife is working with .....Australian company.

   * an

   * a

   * the

   * none of the above

   iii) I .... my home work before he arrived.

   * have done

   * am doing

   * have been doing

   * had done

c) **Incomplete Options**

   i) Which expression, when completed correctly, is the largest state in India in terms of population.

   a) M P

   b) U P

   c) W B

   d) J K

   ii) The surname of the first Prime Minister of India begins with the alphabet.

   A) R

   B) S

   C) G

   D) N

(The incomplete options can often be used when giving the options in full would make the answer too obvious or for testing vocabulary where the recognition element may make the item invalid.)

8.5.2.3 True-False

As its name implies, the basic true-false item requires the pupil to select either 'true' or 'false' as the answer. It is usually written in the form of a statement which the pupil must decide as being either 'true' or 'false' or alternatively choose between other work pairs relating to the statement such as greater than-less than, plus-minus, often-rarely, same-different, 'faster-slower' and so on. It is the possibilities offered by these other pairs which make the true/false form a particularly useful one.

Examples

State whether the following statements are true or false :

1) The Industrial Revolution first began in Asia.

2) Subhash Chandra Bose was the founder of Azad Hind Fauj.

3) 2/3 is a rational number.

4) H_2SO_4 is the formula of sulphuric acid.

5) Volume is indirectly proportionate to Temperature.

Within the category of true-false can be included another variety of alternative answer items. This provides three or more possible answers not differing in kind as in multiple-choice but simply occupying different position along a true/false, positive/negative scale. Common examples of such answer are - always false, sometimes false, always true; greater than, equal to, less than. There are many' possible developments on this theme. One is to couple the category of answer with a reason or supporting statement thereby bringing in some aspect of understanding. A possible answer structure for such an item which could make use of a variety of materials is :

A   (True) – The statement is supported by reference to evidence.

B   (?) – There is no evidence one way or the other to support or refute the statement.

C   (False) – The statement is not supported by reference to the evidence.

Often more than three scaled answer of this type can be useful but to go beyond five scaled answer places too much of a burden upon the pupil's memory in relation to the working of the answer categories.

Examples

1.   For each of the pairs of values listed below circle G if the first is greater, L if it is less, and E if it is equal to the second.

i)	density of ice and that of water vapours	G-L-E
ii)	speed of sound in air and speed of sound in water	G-L-E
iii)	velocity of radio waves and that of light	G-L-E
iv)	Rupee and Dollar	G-L-E
v)	Population of India and Population of China	G-L-E
vi)	100×1000 and 1000×100	G-L-E

2.   For each of the following statements :

a)   circle TT if both parts are true;

b)   circle FF if both parts are false;

c)   circle TF if the first part is true but the second part false;

d)   circle FT if the first part is false but the second true.

i)   'Bharat Ratna' was awarded to Pt. Jawahar Lal Nehru after his death.
TT/FF/TF/FT

ii)   India succeeded in regaining its territories after the Indo-China War of 1962.
TT/FF/TF/FT

iii)   In the nineteenth century, the Mughals lost their empire in India and the British established their supremacy over India.   TT/FF/TF/FT

3. For each of statements below:

a) Circle 'T' if there is conclusive evidence to support it;

b) Circle 'F' if there is conclusive evidence to refute it;

c) Circle '?' if there is no dependable evidence to support or refute it.

    i) Eating carrots improves the ability to see in the dark.   T/F/?

    ii) Vitamin C deficiency causes scurvy   T/F/?

Although set in a science context this type of question is equally applicable to the 'explanatory' aspects of other subjects such as History and Geography.

The True-False kind of question is very easy to set as well as to evaluate and so can be very useful for teachers. But it has certain pitfalls too. It is the most convenient format, permitting teachers to pick up statements straight away from the books. This may encourage rote memory, but if they are set with some imagination, they can test higher level objectives, too.

8.5.2.4 Matching Block

The matching block format consists of two lists and the pupil is required to correlate correctly one or more entries from one list with one or more entries from the other so that correct matching by elimination is not possible.

Examples

a) Indicate the city in which each of the landmarks in List I is located by completing the match-panel given after the list by entering the letter from list II.

List I (Landmarks)	List II (Locations)
1. India Gate	a. Jaipur
2. Gateway of India	b. Delhi
3. Hawa Mahal	c. Madras
4. Howrah Bridge	d. Lucknow
5. Fort St. George	e. Bhopal
	f. Calcutta
	g. Bombay

List I   1 2 3 4 5

List II

b) In the match panel given after the list enter the letter from List II to indicate the century of each event in List I.

List I	List II
Historical Event	Century A.D
1. The French Revolution	a. 15th
2. The First Indian War of Independence	b. 16th
3. Establishment of Mughal empire in India	c. 17th
4. The Russian Revolution	d. 18th
5. Renaissance in Europe	e. 19th
	f. 20th

List I   1 2 3 4 5

List II

(This application of matching block rapidly tests knowledge of sequence and similar 'frameworks')

8.5.2.5 Advantages and Disadvantages of Objective Tests

One problem immediately presents itself with any form of choice response, viz. guessing. Instructions not to guess are impossible to enforce and it is equally impossible for an evaluator to tell whether answers have or have not been guessed. This is particularly true of True/False tests where a candidate has a 50/50 chance of making a correct guess.

However, there are several positive advantages in using objective tests. Probably the most obvious advantage is the speed at which they can be marked. Marking doesn't need any special skill and so can be done by anyone, including pupils who have just taken the test. Moreover, as questions are short and easily answered, knowledge of a large syllabus content can be sampled. By using such tests to sample content knowledge, there is time available to use other assessment techniques to test skills. Objective tests also give an opportunity to pupils who are poor writers to demonstrate their knowledge without subjective elements creeping in.

However, objective tests are not appropriate for all occasions. Whereas they are excellent for sampling knowledge, it is much more difficult to construct them to test higher order skills. They can never test written expression, or ability to argue in one's own words. If well written, however, they can test higher order skills. The overuse of objective tests at the expense of other forms of assessment may result in an assessement which can be biased and invalid. But probably the main disadvantage of objective tests is the difficulty in writing good ones. Just as well-written tests are highly valid, so badly written objective tests are highly invalid.

Check Your Progress 2

Do you prefer objective type tests to the others? Why? or why not?

....

....

8.6 ADMINISTRATION OF AN ACHIEVEMENT TEST

Having prepared a good test, you should plan to administer it in such a way that each of your students will do his/her best.

Motivating students is very important, and this is an area in which each teacher will have his/her own special technique. If you can get your students to see this test as an interesting and challenging task, which will benefit them, they will surely do well. Let them understand the advantages of the class test. Make them understand that such tests help them to get a feedback on their weaknesses and misunderstandings, which can be corrected before they face external examinations. Experience has shown that students who are given frequent class tests and are subjected to continuous assessment do better in external examinations.

Some of the values of designing a good test and preparing students well for the test may be lost if you do not plan in advance for its administration. Detailed planning is necessary as any confusion in the administration of a test is found to disturb the examinee and lower the validity of the results. Some tips to be kept in mind while planning for the administration of a test are given below:

a) Time Schedule

Be sure you plan your time schedule carefully, ensuring teacher and pupil readiness. Much preparation may be done a day before. It will be wise to schedule enough time for briefing the invigilators.

If there is a deadline for finishing and leaving the room (e.g., the end of a class period), be specially sure to plan for adequate time at the end for the things which must be done. Even with a small class these take five to ten minutes, and with a large group they may take at least fifteen minutes. A hasty wind-up may result in non-fulfilment of the objectives of the test, or other disasters.

b) The Room

It is important for any examination to provide a quite, comfortable atmosphere, in which the students are encouraged to do their best. As much as possible, try to test in a quiet place with a minimum of distracting noises. Avoid rooms near cafeterias, important hallways, playing fields or other noisy places. Request nearby loudspeaker owners to shut them off for the duration of the examination hours. Hang signs on the door, saying "EXAMINATION IN PROGRESS : DO NOT DISTURB". Objective examinations generally require more intense concentration than essay type exams. The latter demand an excess of physical endurance (trying to write fast enough to keep up with one's thoughts). Objective tests require constant, careful and critical thinking and reasoning, with a minimum of physical work.

c) **Desks, Etc.**

Remember that the students will be writing on a single – thickness answer sheet, not a thick answer book. Be sure the writing surfaces are at least 30×38 cm., and as smooth as possible. If there are cracks or scratches a student's pencil may push through the answer sheet, spoiling it and making it hard to mark. Also be sure the room is clear of any charts, posters, etc. that might help some candidates.

d) **Equipment**

It is wise to make up a check-list, ahead of time, of what you will have to take with you to the examination hall. Be sure to include chalk to write necessary notices on the black board. If there is no black board make up placards or poster ahead of time. Also invest in a dozen or so soft pencils, preferably with erasers. This is in case some students (a) bring hard pencils, (which make the answer sheet much harder to score), or (b) break a pencil and don't have a spare pencil. For exact timing of the test (much more important for objective tests) it is better to have two watches or clocks, in case one stops.

e) **Invigilators**

For anything more than an informal, half-period quiz, you will probably need the help of one or more invigilators. Chose persons who are willing to give their full attention to the task. Neither you nor your invigilators should talk, read, correct papers or do any other work during the examination time. They should observe closely, circulating constantly, checking that the students are answering in the right place, using soft pencils, not copying etc. However, they should not hover too long over any student, as this makes the examinee nervous.

8.7 SCORING AND RECORDING OF TEST RESULTS

Despite the objectivity of scoring short answer tests, certain procedures are indispensable if scoring is to be done with maximum accuracy and efficiency. The necessity for extreme care in scoring has been indicated by several studies showing that scoring errors occur with appalling frequency. "Constant" errors can be due to failure to understand scoring directions, with resultant scores which are consistently too low or too high. "Variable" errors can be due to carelessness in marking, adding, computing, or transcribing scores. These errors warrant (1) the careful training and instruction of scorers and (2) the rescoring of at least a sample of any group of test booklets or answer sheets.

8.7.1 Order of Scoring

With essay tests it may be desirable to have one person score all answers to the first question, then to the second, and so on. If, for objective tests separate answer sheets are provided, the scorer may score a given page in all booklets first, then the next page, and so on, rather than scoring all of one booklet before going on to the next. If so many booklets must be scored that several scorers are needed, each person may specialize on a given page or group of pages of the booklet but should score only one page in all booklets at a time.

8.7.2 Rescoring

With a large number of booklets to be scored and sufficient help available, it is always worthwhile to rescore them so as to eliminate errors that otherwise are almost inevitable in a clerical task like this. If complete rescoring is not feasible every fifth or tenth booklet should be rescored to get a rough idea of the frequency and magnitude of scoring errors. Rescoring a sample sometimes uncovers such an inaccuracy as to make it desirable to rescore the remainder.

8.7.3 Keeping Records

As soon as possible after the tests have been administered, the answer sheet should be checked and scored, and the scores should be recorded on the permanent records of the school. Each teacher should be given copies of the score reports for the pupils in his/her classes. Usually schools have some type of permanent record for each pupil which provides space for recording standardized test results.

The form in which test results are recorded is often meaningless to anyone except the persons recording them. Sometimes permanent records for a pupil contain such information as the following:

| IQ | 104 | Mathematics | 97 |
| Reading | 68 | Science | 93 |

What do these scores mean ? What test of intelligence was used ? What was its standard deviation? Are the Reading, Mathematics and Science scores the raw scores, percentile ranks, or same other type of standard or derived score? Unless the cumulative record contains complete information about the test and the type of score, the effort involved in carrying on a testing program, scoring the tests, and reporting the scores is practically wasted. If the records are to have value, the following must be indicated : test title, form of the test, date when the test was given, the raw score or standard score, and percentile rank under properly identified captions. When percentile ranks are reported, the group on which the norms were based should be identified – for example, national, state, district, local, or other group – and the nature of the group should be specified.

8.8 NORMS AND INTERPRETATION OF TEST SCORES

Norms are based on frequency distributions and in general, half of any group of pupils falls below its average and half above it. A norm is not the "ideal performance" of a group of pupils; it is only the typical performance of typical students at a given time. Hence educators should avoid the fallacy of insisting on bringing everyone up to the norm. School practices resulting from attempts to get all pupils to be like the "average" have held back many superior students and created emotional problems among the less able. Such practices have caused much criticism of education through the years. Benjamin Franklin is reported to have said that "the schools are polishing the bricks and dulling the diamonds."

8.8.1 Local Factors

Many local factors should be taken into account when interpreting the standing of pupils according to norms derived on a nationwide basis. Among these factors are (1) the legal age for entering school, (2) the average age of actually entering school, (3) promotion and detention policies, (4) rate and selectivity of elimination from school, (5) efficiency of the teaching personnel, (6) the grade placement and time allowances, (7) general nature of the curriculum, (8) the standing of local pupils in mental ability and other aspects related to the one being evaluated, (9) the relative emphasis in the local school on academic, social, and vocational development, and (10) the home background of pupils. The meaning of the derived scores for a given group of pupils can then be interpreted in the light of these factors.

When interpreting the performance of individual pupils or of a class as a whole, the teacher should take into special consideration differences in the cultural background of families and communities. There are wide variations in the kind of experiences pupils have. We can expect that differences in language background, richness of home resources, and intensity of the desire for an education will be reflected in pupil performance.

Performance of pupils also varies with varying emphasis on different aspects of the school curriculum. In some subject matter areas, such as Arithmetic, the teacher usually cannot expect his/her pupils to go much beyond instructional materials. In other areas such as reading, there are many opportunities for students to develop skill and knowledge on their own outside the school program. Thus the performance of individuals and groups should be judged, in part at least, on the basis of the curriculum to which they have been exposed. When the performance of a class or an individual deviates considerably from the norms on standardized tests, a need for reappraisal of the school curriculum and of teaching emphases may be indicated. In many practical situations it is necessary to use norms whose applicability to local conditions is questionable or unknown. To the extent that it is so, the data obtained from them cannot be interpreted meaningfully for individual pupils.

A final precaution is to avoid using tests to punish pupils or to foster a spirit of rivalry among teachers or schools. Teachers and administrators must keep the welfare of pupils uppermost in their minds and be sensitive to the requirements of adequate human relations. Failure to do this in administering and interpreting a testing program has produced negative feelings about tests in both pupils and teachers. So much for qualitative interpretations. You will read about quantitative interpretations of test scores with the help of satistical analysis in Block 4.

8.9 GRADES

What is grading? In the system of grading, students are classified into a few ability groups or categories according to their level of achievement in an examination. The achievement is defined in the form of numerical or letter grades, each of which denotes a certain level of performance, generally not in absolute terms but in relation to the performance of the whole group.

As stated above, grading is essentially meant for categorising students into a few ability groups on the basis of their performance in the examination. There are two approaches to formation of groups that define the grades (a) on the basis of absolute marks and (b) on the basis of relative marks or rank order of marks. Let us see what these mean and also consider their merits and demerits.

8.9.1 Absolute Grading

This approach involves direct conversion of marks into grades. Whatever be the distribution of marks in a subject, the marks between two fixed points on 0-100 scale would correspond to a given grade. An example of this is the categorisation of students into 5 groups - Distinction, Ist, 2nd, 3rd Division and fail categories on the basis of marks as follows:

75 or above	:	Distinction
60 – 74	:	Ist Division
45 – 59	:	2nd Division
33 – 44	:	3rd Division
Below 33	:	Fail

It is possible to form any number of groups to correspond to grades (A,B,C etc.) in this way on the basis of marks. However, in view of the disparity in the distribution of marks of different subjects, Grade A of one subject can not be treated at par with Grade A of another subject though Grade A is based on the same cut-off point in both the subjects. For example, if it is decided to award Grade A to those scoring 90% or more whatever be the subject, there may be no student getting Grade A in English or History while quite a few will be getting grade A in Mathematics. In a sense it only serves as a substitute for individual marking system except that it gives a number of ability groups.

8.9.2 Comparative Grading

This involves conversion of marks into grades on the basis of rank order or percentiles. In this case the distribution of marks is taken into consideration while determining the range of marks corresponding to different grades. For example, the top 5% students may be given grade A, the next 10% grade B and so on. Here the actual cut-off score for grade A in one subject may be quite different from that of another subject. In this case the grade that a student gets depends on his/her relative performance, that is, on what his/her marks are in relation to the marks of others. This type of grading actually corresponds to norm-referenced testing about which you have read in unit 2.

8.9.3 Advantages

In general when we talk of grading it is only the type of grading based on relative marks that we have in mind. These grades are expressed in the form of letters A, B, C etc. The following are the main advantages of such grading:

i)   With the same uniform pattern being adopted for all subjects, grading would provide better comparability of the results of different years in the same subject.

ii) Grading is essentially based on rank ordering of students. Studies have shown that agreement among examiners on ranks to be awarded to examinees is much more in this than on absolute marks. Hence grades based on rank order in general, are more reliable.

iii) There is greater comparability among subjects when grades are used. When there is a choice of subjects, students need not avoid the subjects which are considered low scoring. Even with a so called low scoring subject, the proportion of students getting a grade would be nearly the same as in a so called high scoring subject.

iv) Grades in different subjects in an examination provide a meaningful profile of the achievement of a student. Unlike marks, one can easily find out in which subjects the performance is outstanding, good, fair or poor. With marks, one can arrive at such inference only on knowing what the range, average and dispersion are of the marks in the different subjects.

Check Your Progress 4

What are the two kinds of grading ? Which kind will you prefer as a teacher? Why?

....

....

....

8.10 LET US SUM UP

In this unit we were concerned with achievement tests which are invariably used by all teachers. We started with the purpose of the tests and went on to their construction. We discussed in detail the steps involved in it – Instructional Objectives, Design, Blueprint, Writing of Question Items, Marking Scheme and Question-wise Analysis.

Further, we learnt about different types of questions. We went into detailed descriptions of short answer and objective type questions.

We then discussed administration of standardised tests, their scoring, recording of results, norms and interpretation of scores. We concluded with a description of grades, their utility and their kind.

This unit is of great practical utility for all practising teachers.

8.11 UNIT-END EXERCISE

1. Prepare an achievement test in your subject taking up any lesson. Follow all the steps as given in the unit, starting with instructional objectives and ending with question-wise analysis.

8.12 POINTS FOR DISCUSSION

1. Are objectives tests superficial and hollow?

2. How will you use objective type tests for testing higher level instructional objectives?

8.13 ANSWERS TO CHECK YOUR PROGRESS

1. i) The steps involved in constructing an achievement test are:

a)   Identifying instructional objectives

b)   Designing the test

c)   Preparing the Blueprint

d) Write appropriate questions in accordance with the Blueprint

e) Prepare the marking scheme

f) Analyse each question on various parameters stated in the Blueprint (Question-wise Analysis)

ii) The purpose of the marking scheme is to :

a) Avoid inconsistency in judgement

b) Ensure objectivity in assessment

c) Eliminate differences in score

d) Make results more reliable

e) Reduce variations in marks - both intra-examiner as well as inter-examiner

2. Objective type tests are better than others because :

a) They can be marked easily and very fast

b) Marking doesn't need any special skill

c) Marking can be done by anyone

d) The whole syllabus can be covered, at least sampled

e) Well written objective tests can be highly valid and reliable for testing knowledge and recall of content

f) Marking doesn't vary and so is objective and free from bias

Objective type tests are not preferred because :

a) It is difficult to write good objective items

b) Guess work can be resorted to while answering them

c) Badly written objective tests are highly invalid and unreliable

d) They can't test higher order skills, at least it is very difficult to construct objective tests for higher order skills

e) They are not appropriate for all occasions

f) They can never test written expression or ability to think or argue in one's own words.

3. While scoring a test, remember :

a) The order of scoring - one question, if essay type, in all the answer books at a time or one page if objective should be done.

b) It is worthwhile to re-check and rescore to avoid mistakes, or at least a sample should be re-checked.

The important factors in interpreting the scores of a test are :

a) Differences in cultural background

b) Variations in the kind of experiences pupils have

c) Differences in language background

d) Richness of home resources

e) Intensity of the desire for an education

f) Varying emphasis on different aspects of the school curriculum.

4. The two kinds of grading are :

a) Absolute

b) Relative or Comparative

Comparative grading should be preferred because :

a) It is based on rank ordering or percentiles

b) It provides better comparability of scores irrespective of whether a subject is scoring or non-scoring

c)   The gaps between subjects (Scoring or Non-Scoring) are done away with since grading is based on rank order or percentiles and not on absolute marks

d)   They provide a more meaningful profile of the achievement of a student or a group

8.14   SUGGESTED READINGS

Ebel, Robert L., (1966) : *Measuring Educational Achievement*, Prentice-Hall of India, New Delhi.

Ebel, Robert L. and Frisbie, David A., (1991) : *Essentials of Educational Achievement*, Prentice-Hall of India, New Delhi.

Harper, A. Edwin, J. and Harper, Erika S., (1992) : *Preparing Objective Examinations*. A Handbook for Teachers, Students and Examiners, Prentice-Hall of India, New Delhi.

Popham, W. James, (1990) : *Modern Educational Measurement: Pracatitioners Perspective*. Prentice-Hall, USA.

Remmers, H.H et. al., (1967) : *A Practical Introduction to Measurement and Evaluation*. Universal Bookstall, Delhi.