SYARAHAN PERDANA 2011

ASSESSMENT TEST ANXIETY AND ACHIEVEMENT
SYARAHAN PERDANA 2011

ASSESSMENT TEST ANXIETY AND ACHIEVEMENT

Prof. Dr. Sulaiman Bin Yamin
Faculty of Technical Education
Universiti Tun Hussein Onn Malaysia (UTHM)
Malaysia

Penerbit UTHM
2011
# TABLE OF CONTENT

**Preface**

**CHAPTER 1 ASSESSMENT, TEST ANXIETY AND ACHIEVEMENT**

1.0 **ASSESSMENT, TEST AND MEASUREMENT**

1.1 General Principles of Assessment 2
1.2 Assessment and the Instructional Process 5
1.3 Placement, formative, diagnostic and Summative Assessment 9
1.4 Norm-referenced and Criterion-referenced Measurement 11

**CHAPTER 2**

2.0 **INSTRUCTIONAL GOALS & OBJECTIVES: FOUNDATION FOR ASSESSMENT**

2.1 Instructional Objectives as Learning Outcomes 15

**CHAPTER 3**

3.0 **VALIDITY AND RELIABILITY**

3.1 Natures of Validity 28
3.2 Nature of Reliability 38
3.3 Usability 58

**CHAPTER 4 PLANNING CLASSROOM TESTS AND ASSESSMENTS**

4.0 **TESTING AND ASSESSMENT DURING INSTRUCTION**

4.1 End of Instruction Testing and Assessment 62
4.2 Selecting appropriate of Items and Assessment Tasks 64
4.3 General Suggestions for Writing Test Items and Assessment Tasks 73
4.4 Focusing on Improving Learning and Instruction 74
CHAPTER 5 ASSEMBLING, ADMINISTERING & APPRAISING CLASSROOM TESTS & ASSESSMENTS

5.0 ASSEMBLING THE CLASSROOM TEST

5.1 Recording Test Items 73
5.2 Reviewing Test Items and Assignment Tasks 73
5.3 Arranging Items in the Test 77
5.4 Preparing Directions for the Test or Assessment 77
5.5 Administering and Scoring Classroom Test & Assessment 81
5.6 Determining item and Task Effectiveness 84

CHAPTER 6 ELEMENTARY STATISTICS

6.0 ORGANIZING AND DISPLAYING SCORES 92

6.1 Simple Ranking 93
6.2 Graphic Presentations of Frequency Distribution 95
6.3 Measures of Central Tendency 97
6.4 The Range 99
6.5 Coefficient of Correlation 100
6.6 Linear Regression Coefficients 100

CHAPTER 7

7.0 TEST ANXIETY 103

7.1 Theories of Effect of Test Anxiety 106
7.2 Statement of Problem 112
7.3 The Research Hypotheses 112
7.4 Selection of Sample 113
7.5 Research Variables 115
7.6 Criterion Instruments 115
7.7 Achievement in Chemistry 115
7.8 Scale and Scoring 116
7.9 Content Validity 116
7.10 Reliability 116
7.11 Test Anxiety 117
7.12 Validation of the Translation 117
7.13 Attitude toward Science 117
7.14 Reliability 118
7.15 Research Designs and Procedures 118
7.16 Testing the Hypotheses 118
CHAPTER 8

8.0 CRITICAL ISSUES IN ASSESSMENT 131

8.1 Transformation of the School-based Assessment 132
8.2 Test Anxiety and Achievement 133

Bibliography 141
Curriculum Vitae 148
PREFACE

This booklet provides general idea on assessment and its attributes. It covers detailed picture of assessment; the importance of validity and reliability in assessment. How to measure validity and reliability are also being discussed.

The booklet also discusses the administration of the test, scoring the test, and analysis of the test.

Test anxiety which is one of the attributes of assessment or testing is mentioned in great length.

Finally, the booklet compare between public assessment and school based assessment. The positive and negative aspects of these two forms of assessment are clarified in detail. Also mentioned is the effect of test anxiety in assessment.

In conclusion, the writer wants the teachers, and parents and audiences to have more responsible attitudes and actions towards the assessment as it will affect their students or children.
ACKNOWLEDGEMENT

I would like to express my sincere thanks and gratitude to my colleagues in the office, especially to my Personal Assistant, Pn. Mollyza bt. Abdul Majid and Pn. Norazreen bt. Mohamad@Kasbani; Assistant Office of Information Technology.

I would also like to convey my gratitude and appreciation to the top management of the University for providing this opportunity in sharing the knowledge and practice on assessment, test anxiety and achievement.

Finally, I am indebted to my wife and family for their never ending support and encouragement towards my work and carrier.
1.0 ASSESSMENT, TEST AND MEASUREMENT

The terms assessment, test and measurement are easily confused because all may be involved in a single process. **Assessment** is a general term that includes the full range of procedure used to gain information about student learning (observations, ratings of performances or projects, paper-and-pencil tests) and the formation of value judgments concerning learning progress. A **test** is a particular type of assessment that typically consists of a set of questions administered during a fixed period of time under reasonably comparable conditions for all students. **Measurement** is the assigning of numbers to the results of a test or other type of assessment according to a specific rule (e.g., counting correct answers or awarding points for particular aspects of an essay).

Assessment is a much more comprehensive and inclusive term that measurement or testing. The term measurement is limited to quantitative descriptions of students; that is, the results of measurement are always expressed in numbers (e.g., Ahmad correctly solved 35 of the 40 arithmetic problems). It does not include qualitative descriptions (e.g. Ahmad’s works was neat) nor does it imply judgments concerning the worth or value of the obtained results. Assessment, on the other hand, may include both quantitative descriptions (measurement) and qualitative descriptions (non-measurement) of students. In addition, assessment always includes value judgments concerning the desirability of the results. **Figure 1** shows the comprehensive nature of assessment and the role of measurement and non-measurement techniques in the assessment process.
1.1 GENERAL PRINCIPLES OF ASSESSMENT

Assessment is an integrated process for determining the nature and extent of student learning and development. This process will be most effective when the following principles are taken into consideration.

1. **Clearly specifying what is to be assessed has priority in the assessment process.** The effectiveness of assessment depends as much on a careful description of what to assess as it does on the technical qualities of the assessment procedures used. Thus, specification of the characteristics to be measured should precede the selection or development of assessment procedures. When assessing student learning, this means clearly specifying the intended learning goals before selecting the assessment procedures to use.

2. **As assessment procedure should be selected because of its relevance to the characteristics or performance to be measured.** Assessment procedures are frequently selected on the basics of their objectivity, accuracy or convenience. Although these criteria are important, they are secondary to the main criterion: is this procedure the most effective method of measuring the learning or development to be assessed? Every procedure is appropriate for some uses and inappropriate for others. In assessing student achievement, for example, we need a close match between the intended learning goals and the types of assessment tasks used. If the development of the ability to organize ideas and write a well-integrated composition is a learning goal, then a multiple-choice test on the mechanics of writing would be a poor substitute for assessment based on analyses of student writing under a variety of conditions (e.g., in-class essay tests, writing projects, and term papers).
FIGURE 1
The Assessment Process

Assessment

Measurement (e.g. testing) [plus]

Value Judgments (e.g. good learning progress)

and/or

Non-measurement (e.g. informal observation) [plus]
3. **Comprehensive assessment requires a variety of procedures.** No single type of instrument or procedure can assess the vast array of learning and development outcomes emphasized in a school program. Multiple-choice and short-answer tests of achievement are useful for measuring knowledge, understanding, and applications outcomes, but essay tests and other written projects are needed to assess the ability to organize and express ideas. Projects that require students to formulate problems, accumulate information through library research, or collect data (e.g., through experimental observations or interviews) are needed to measure certain skills in formulating and solving problems. Observational techniques are needed to assess performance skills and various aspects of student behavior. And self-report techniques are useful for assessing interests and attitudes. A complete picture of student achievement and development requires the use of many different assessment procedures.

4. **Proper use of assessment procedures requires and awareness of the limitations.** Assessment procedures range from very highly developed measuring instruments (e.g., standardized aptitude and achievement tests) to rather crude assessment devices (e.g., observational and self-report techniques). Even the best educational and psychological measuring instruments yield results that are subject to various types of measurement error.

   Sampling error is one common problem in educational and psychological measurement. An achievement test may not adequately sample a particular domain of instructional content. An observational instrument designed to assess a student’s social adjustment may not sample enough behavior for a dependable index of this trait. Fortunately, sampling error can be controlled through careful application of established measurement procedures.

   A second source of error is caused by chance factors influencing assessment results, such as guessing on objective tests, subjective scoring on essay tests, errors in judgment on observation devices, and inconsistent responding on self-report instruments (e.g., attitude scales). Because of these problems, students just a few points apart on an educational assessment should not be considered to be different. In fact, no score on an educational or psychological assessment should be treated as a totally accurate measurement of the characteristic in question. Through the
careful use of assessment procedures we are able to keep these errors of measurement to a minimum.

The incorrect interpretation of measurement results constitutes to a minimum source of error. Users of educational assessments sometimes interpret results as more precise than they are, or as an indication of characteristics beyond those the assessment is designed to measure. For instance, scholastic aptitude scores are sometimes interpreted as measures of innate abilities rather than modifiable abilities, or as a measure of general personal worth rather than as a limited measure of verbal and numerical reasoning. Misinterpretation of test results is all too common, but it can be controlled by careful attention to what the test actually measures and how accurately it does so.

These limitations of assessment procedures do not negate the value of tests and other types of assessments. A healthy awareness of the limitations of assessment instruments makes it possible to use them more effectively. Keep in mind that the cruder the instrument, the greater its limitations and, consequently, the more caution required in its use.

5. **Assessment is a mean to an end, not an end in itself.** The use of assessment procedures implies that some useful purpose is being served and that the user is clearly aware of this purpose. To blindly gather data about students and then file the information away is a waste of both time and effort. Assessment is best viewed as a process of obtaining information on which to base educational decisions.

### 1.2 ASSESSMENT AND THE INSTRUCTIONAL PROCESS

The main purpose of classroom instruction is to help students achieve a set of intended learning goals. These goals should typically include desired changes in the intellectual, emotional, and physical spheres. When classroom instruction is viewed in this light, assessment becomes and integral part of the teaching-learning process. The intended learning outcomes are established by the instructional goals, the desired changes in students are brought about by the planned learning activities, and the students’ learning progress is periodically assessed by tests and other assessment devices. While the interdependent nature of teaching and learning is beyond dispute, the interdependent nature of teaching, learning, and assessment is less often
recognized. The statements in “Measurement and Assessment – Essential Components of Effective Teaching” attest to the fact that leaders in the advancement of teaching as a profession recognize the critical role of assessment as an integral part of effective teaching. The interdependence of these three facets of education can be clearly seen in the following steps included in the instructional process.

1.2.1 Identifying Instructional Goals
The first step in both teaching and assessment is determining the learning outcomes to be expected from classroom instruction. How should students think and act when they complete the learning experience? What knowledge and understanding should the students possess? What skills should they be able to display? What interest and attitudes should they have developed? What changes in habits of thinking, feeling, and doing should have taken place? In short, what specific changes are we striving for, and what are students like when we have succeeded in bringing about these changes?

1.2.2 Pre-assessing the Learners’ needs
When the instructional goals have been clearly specified, it is usually desirable to make some assessment of the learner’s needs in relation to the learning outcomes to be achieved. Do the students possess the abilities and skills needed to proceed with the instruction? Have the students already developed the skills and understanding intended? Assessing students’ knowledge and skill at the beginning of instruction enables us to answer such questions. This information is useful in planning work for students who lack the prerequisite skills and in modifying our instructional plans to fit the needs of the learners.

1.2.3 Providing Relevant Instruction
Relevant instruction takes place when course content and teaching methods are integrated into planned instructional activities designed to help students achieve the intended learning outcomes. During this instructional phase, measurement and assessment provide a means of monitoring learning progress and diagnosing learning difficulties. Thus, periodic assessment during instruction provides a type of feedback-corrective procedure that aids in continuously adapting instruction to group and individual needs.
1.2.4 Assessing the Intended Learning Outcomes

The final step in the instructional process is to determine the extent to which the learning objectives were achieved by the students. This is accomplished by using tests and other types of assessments that are specifically designed to measure the intended learning outcomes. Ideally, the instructional goals will clearly specify the desired changes in students and the assessment instruments will provide a relevant measure or description of the extent to which those changes have taken place. Matching a range of assessment procedures to the intended learning outcomes, with particular emphasis on those that are judged to be most important, is basic to effective classroom assessment and will receive considerable attention in later chapters.

1.2.5 Using the Results

Student assessment is often regarded as being essentially for the benefit of teachers and administrators. This attitude overlooks the direct contribution assessment can make to students. Properly used assessment procedures can contribute directly to improved student learning by 1) clarifying the nature of the intended learning outcomes, 2) providing short-term goals to work toward, 3) providing feedback concerning learning progress, and 4) providing information for overcoming learning difficulties and for selecting future learning experiences. Although these purposes are probably best served by the periodic assessment during instruction, the final assessment of intended outcomes also should contribute to these ends.

Information from carefully developed tests and other types of assessments also can be used to improve instruction. Such information can aid in judging 1) the appropriateness and attainability of the instructional goals, 2) the usefulness of the instructional materials, and 3) the effectiveness of instructional methods. Thus, assessment procedures can contribute to improvements in the teaching-learning process itself, as well as contributing directly to improved student learning.
Assessment results are, of course, also used for assigning marks and reporting student progress to parents. The systematic use of a wide range of assessment procedures provides an objective and comprehensive basis for reporting on each student's learning progress. In addition to marking and reporting, assessment results also are used in the school for various administrative and guidance functions. They are useful in curriculum development, in aiding students with educational and vocational decision, and in assessing the effectiveness of the school program. The simplified instructional model shown in Figure 1.2 summarizes the basic steps in the instructional process and illustrates the interrelated nature of teaching, learning and assessment.

**FIGURE 1.2**
Simplified Instructional Model

1. **Identify Instructional Goals**
2. **Pre-assess Learners’ needs**
3. **Provide relevant Instruction**
   1. Monitor learning progress
   2. Diagnose learning difficulties
4. **Assess Intended Outcomes**
   - Improvement of Learning and Instruction
   - Marking and Reporting to Parents
   - Use of Results for Other School Purposes
1.3 Placement, Formative, Diagnostic and Summative Assessment

Tests and other assessment procedures can also be classified in terms of their functional role in classroom instruction. One such classification system follows the sequence in which assessment procedures are likely to be used in the classroom. These categories classify the assessment of student performance in the following manner:

1. **Placement assessment**: to determine student performance at the beginning of instruction.
2. **Formative assessment**: to monitor learning progress during instruction.
3. **Diagnostic assessment**: to diagnose learning difficulties during instruction.
4. **Summative assessment**: to assess achievement at the end of instruction.

Each of these types of classroom assessment typically requires instruments specifically designed for the intended use.

1.3.1 Placement Assessment. Placement assessment is concerned with the student’s entry performance and typically focuses on questions such as the following: 1) Does the student possess the knowledge and skills needed to begin the planned instruction? For example, does the beginning reader have the necessary reading readiness skills, or does the beginning algebra student have a sufficient command of essential arithmetic concepts? 2) to what extent has the student already developed the understanding and skills that are the goals of the planned instruction? Sufficient levels of comprehension and proficiencies might indicate the desirability of skipping certain units or of being placed in a more advanced course. 3) To what extent do the student’s interest, work habits, and personality characteristics indicate that one mode of instruction might be better than another (e.g. group instruction versus independent study)? Answers to questions like these require the use of a variety of techniques: records of past achievement, pretests on course objectives, self-report inventories, observational techniques, and so on. The goal of placement assessment is to determine for each student the position in the instructional sequence and the mode of instruction that is most beneficial.
1.3.2 Formative Assessment. Formative assessment is used to monitor learning progress during instruction. Its purpose is to provide continuous feedback to both student and teacher concerning learning successes and failures. Feedback to students provides reinforcement of successful learning and identifies the specific learning errors that are in need of correction. Feedback to the teacher provides information for modifying instruction and for prescribing group and individual work. Formative assessment depends heavily on specially prepared tests for each segment of instruction (e.g., unit, chapter). Tests and other types of assessment used for formative assessment are most frequently teacher made, but customized tests made available by publishers of textbooks and other instructional materials also can serve this function. Observational techniques are, of course, also useful in monitoring student progress and identifying learning errors. Because formative assessment is directed toward improving learning and instruction, the results are typically not used for assigning course grades.

1.3.3 Diagnostic Assessment. Diagnostic assessment is a highly specialized procedure. It is concerned with the persistent or recurring learning difficulties that are left unresolved by the standard corrective prescriptions of formative assessment. If a student continues to experience failure in reading, mathematics, or other subjects, despite the use of prescribed alternative methods of instruction, then a more detailed diagnosis is indicated. To use a medical analogy, formative assessment provides first-aid treatment for simple learning problems and diagnostic assessment searches for the underlying causes of those problems that do not respond to first-aid treatment. Thus, diagnostic assessment is much more comprehensive and detailed. It involves the use of specially prepared diagnostic tests as well as various observational techniques. Serious learning disabilities also are likely to require the services of educational, psychological, and medical specialists. The aim of diagnostic assessment is to determine the causes of persistent learning problems and to formulate a plan for remedial action.
1.3.4 **Summative Assessment.** Summative assessment typically comes at the end of a course (or unit) of instruction. It is designed to determine the extent to which the instructional goals have been achieved and is used primarily for assigning course grades or for certifying student of the intended learning outcomes. The techniques used in summative assessment are determined by the instructional goals but they typically include teacher-made achievements tests, ratings on various types of performance (e.g. laboratory, oral report), and assessments of products (e.g., themes, drawings, research reports). Although the main purpose of summative assessment is grading, or the certification of student achievement, it also provides information for judging the appropriateness of the course objectives and the effectiveness of the instruction.

1.4 **Norm-Referenced and Criterion-Referenced Measurement**

How the results of tests and other assessment procedures are interpreted also provides a method of classifying these instruments. There are two basic ways of interpreting student performance. **Norm-referenced** interpretation describes the performance in terms of the relative position held in some known group (e.g., Noraini typed better than 90 percent of the class members). **Criterion-referenced** interpretation describes the specific performance that was demonstrated (e.g., Molly typed 40 words per minute without error). When interpretations are confined to the attainment of a specific objective (e.g., capitalized all proper nouns) they are sometimes called objective referenced. This is a type of criterion-referenced interpretation, but it does not cover as broad a domain if tasks as that typically used in criterion referencing. These concepts are defined more specifically in the accompanying box.

Norm-referenced interpretations might be based on a local, state, or national group, depending on the use to be made of the results. Using national norms, for example, we might describe a student’s performance on a vocabulary test as equaling exceeding that of 76 percent of a national sample of sixth-graders. Criterion-referenced interpretations can be made in various ways. For example, we can (1) describe the specific learning tasks a student is able to perform (e.g., counts from 1 to 1000), (2) indicate the percentage of tasks a student performs correctly (e.g., spells 65 percent of the words in the word list), or (3) compare the test performance to a set standard and make a mastery/non-mastery decision (e.g., answers correctly at least 80 percent of the items measuring identification of the main idea in paragraphs). Although a
standard of mastery can be used in making one type of criterion-referenced interpretation, it is not an essential element of criterion-referenced testing, as illustrated in the first two examples. Although the term percent was used in illustrating both types on interpretation, it was used in a distinctly different way each time. The norm-referenced interpretation indicated the student's relative standing in a norm group by noting the percentage of students in the group who obtained the same or a lower score (called a percentile score). The criterion-referenced interpretation focused on the percentage of items answered correctly (called a percentage-correct score). Although many types of scores are used in testing, the distinction between the percentile score and the percentage-correct score is a significant one because it illustrates the basic difference between a norm-reference interpretation and a criterion-referenced interpretation.

Strictly speaking, “norm referenced” and “criterion referenced” refer only to the method of interpreting the results. These distinct types of interpretation are likely to be most meaningful and useful, however, when tests (and other assessment instruments) are specifically designed for the type of interpretation to be made. Thus, it is legitimate to use the terms criterion referenced and norm referenced as broad categories for classifying tests and other assessment procedures.

Tests that are specifically built to maximize one type of interpretation are impossible to identify merely by examining the test itself (see “Comparison of Norm-referenced Tests and Criterion-referenced Tests”). It is in the construction and use of the tests that differences can be noted. An identifying feature of norm-referenced tests is the selection of items of average difficulty and the elimination of items that all students are likely to answer correctly. This procedure provides a wide spread of scores so that discrimination among students at various levels of achievement is possible. This is useful for decisions based on relative achievement, such as selection, grouping and relative grading. By contrast, criterion-referenced tests include items that are directly relevant to the learning outcomes to be measured, without regard to whether the items can be used to discriminate among students. No attempt is made to eliminate easy items or alter their difficulty. If the learning tasks are easy, then test items will be easy. The goal of the criterion-referenced test is to obtain a description of the specific knowledge and skills each student can demonstrated. This information is useful for planning both group and individual instruction.
These two types of tests are best viewed as the ends of a continuum, rather than as a clear-cut dichotomy. As shown in the following continuum, the criterion-referenced test emphasizes description of performance and the norm-referenced test emphasizes discrimination among individuals.

<table>
<thead>
<tr>
<th>Criterion-Referenced Test</th>
<th>Combined Type</th>
<th>Norm-Referenced Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description of Performance</td>
<td>Dual Interpretation</td>
<td>Discrimination among Individuals</td>
</tr>
</tbody>
</table>

In an attempt to capitalize on the best features of both, testers have attempted to make their norm-referenced tests more descriptive, thus allowing for both norm-referenced and criterion-referenced interpretations. Similarly, testers have added norm-referenced interpretations to tests that were specifically built for criterion-referenced. The use of dual interpretation with published tests seems to be an increasing trend that will move many tests more toward the center of the continuum. Although this involves some compromises in test construction and some cautions in test interpretation, the increased versatility may contribute to more effective test use.
2.0 INSTRUCTIONAL GOALS AND OBJECTIVES: FOUNDATION FOR ASSESSMENT

What types of learning outcomes do you expect from your teaching? Knowledge? Understanding? Applications? Thinking skills? Performance skills? Attitudes? Clearly, defining desired learning outcomes is the first step in good teaching. It is also essential to the assessment of student learning. Sound assessment requires relating the assessment procedures as directly as possible to intended learning outcomes.

Instructional goals and objectives play a key role in both the instructional process and the assessment process. They serve as guides for both teaching and learning, communicate the intent of the instruction to others, and provide guidelines for assessing student learning. These major purposes are illustrated in Figure 2.1.
Useful guidelines and materials specify the types of performance students are expected to demonstrate at the end of the instructional sequence (e.g., unit or course). Describing intended learning outcomes in performance terms is the main function of properly stated instructional goals and objectives. This clarification of what students should be learning and how the learning is to be expressed not only aids the teacher but also helps others understand the focus of the instruction.

Our interest is in the usefulness of instructional goals objectives for assessing student learning. Effective assessment depends as much on a clear description of what is to be assessed as on a determination of how to assess. Thus, before we develop or select tests and other assessment instruments to measure student learning, we need to clearly specify the intended learning outcomes. That is the main function of well-stated instructional goals and objectives.

### 2.1 INSTRUCTIONAL OBJECTIVES AS LEARNING OUTCOMES

Instructional goals and objectives are sometimes stated in terms of actions to be taken. Thus, we might have a statement like this:

> **Demonstrate to students how to use the microscope**

Although this statement clearly indicates what the teaching activity is, it is less clear about intended learning outcomes and does not point very explicitly to the type of student assessment that would be most appropriate. Literally, the objective will have been achieved when the demonstration has been completed – whether the students have learned anything or not. A better way to state objectives is in terms of what we expect students to be able to do at the end of instruction. After demonstrating how to use the microscope, for example, we might expect students to be able to demonstrate skill in using the microscope to identify features of a cell.

A statement such as this directs attention to the students and to the type of performance they are expected to exhibit as a result of the instruction. Thus, our focus shifts from the teacher to the student and from the learning experiences to the learning outcomes. This shift in focus makes clear the intent of our instruction and sets the stage for assessing student learning. Well-stated outcomes make clear the types of student performance we are willing to accept as evidence that the instruction has been successful.
When viewing instructional objectives in terms of learning outcomes, it is important to keep in mind that we are concerned with the *products* of learning rather than with the *process* of learning. *Figure 2.2* makes three important points about the role of instructional objectives in teaching-learning situations.

First, objectives establish direction, and when they are stated in terms of learning outcomes, they go beyond knowledge of the specific course content. Note the distinction between “study of” and “knowledge of” cell structures. The content (study of cell structure) is more aptly listed under process because it is the vehicle through which objectives (knowledge of parts of cell, and so on) are attained.

Second, consider the varying degrees of dependence that the products (“knowledge”, “skill” and “ability”) have on the course content. “Knowledge of parts of cell” is the most closely related to the specific content of a biology course. “Skill in using microscope” and “ability to write accurate reports of scientific observations”, relate to a greater variety of course content that could be used to achieve the same objectives.

The third point illustrated by the diagram is the degree to which objectives vary in complexity. The first learning outcome, “knowledge of parts of cell” is specific, is easily attained, and can be measured readily by a short answer or fixed-response paper-and-pencil test. The last learning outcome, “ability to write accurate reports of scientific observations,” is rather general, is unlikely to be attained completely in a single course, and requires judgmental analysis of student performances.

**FIGURE 2.2**

*The relationship of learning experience to learning outcomes*

<table>
<thead>
<tr>
<th>Student</th>
<th>Learning experience (process)</th>
<th>Learning outcome (product)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Study of cell structure of plants in laboratory</td>
<td>• Knowledge of parts of cell</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Skill in using microscope</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Ability to write accurate reports of scientific observations</td>
</tr>
</tbody>
</table>
2.1.1 Types of Learning Outcomes to Consider

Although the cognitive research perspective is useful in forcing attention on broader educational goals that need to be addressed in the development or selection of an assessment, those general goals need to be supplemented by more systematic thinking about the types of learning outcomes that would provide evidence that the goals are being achieved. It is useful, for this reason, to classify learning outcomes under a few general headings. Any such classification is necessarily arbitrary, but serves several useful purposes. It indicates types of learning outcomes that should be considered, it provides a framework for classifying those outcomes; and it directs attention toward changes in student performance in a variety of areas.

The following list of types of outcomes delineates the major areas in which instructional objectives might be classified. The more specific areas under each type should not be regarded as exclusive; they are merely suggestive of categories to be considered.

1. Knowledge
   1.1.1 Terminology
   1.1.2 Specific facts
   1.1.3 Concepts and principles
   1.1.4 Methods and procedures

2. Understanding
   2.1 Concepts and principles
   2.2 Methods and procedures
   2.3 Written material, graphs, maps and numerical data
   2.4 Problem situations

3. Application
   3.1 Factual information
   3.2 Concepts and principles
   3.3 Methods and procedures
   3.4 Problem-solving skills
4. Thinking skills
   4.1 Critical thinking
   4.2 Scientific thinking

5. General skills
   5.1 Laboratory skills
   5.2 Performance skills
   5.3 Communication skills
   5.4 Computational skills
   5.5 Social skills

6. Attitudes
   6.1 Social attitudes
   6.2 Scientific attitudes

7. Interests
   7.1 Personal interests
   7.2 Educational interests
   7.3 Vocational interests

8. Appreciations
   8.1 Literature, art and music
   8.2 Social and scientific achievements

9. Adjustments
   9.1 Social adjustments
   9.2 Emotional adjustments
2.1.2 Taxonomy of Educational Objectives

A useful guide for developing a comprehensive list of instructional objectives is the *Taxonomy of Educational Objectives*. This is a detailed classification of objectives, similar in form to the classification system used for plants and animals. It attempts to identify and classify all possible educational outcomes. The system first divides objectives into the following three major areas:

1. **Cognitive Domain**: Knowledge outcomes and intellectual abilities and skills

2. **Affective Domain**: Attitudes, interest, appreciation, and modes of adjustment

3. **Psychomotor Domain**: Perceptual and motor skills

Each of the three domains is subdivided into categories and subcategories. The major categories in the cognitive domain, for example, are knowledge, comprehension, application, analysis, synthesis, and evaluation. These categories begin with relatively simple knowledge outcomes and proceed through increasingly complex levels of intellectual ability. This pattern of classification is characteristic of all three of the domains.

The *Taxonomy* is primarily useful in identifying the types of learning outcomes that should be considered when developing a comprehensive list of objectives for classroom instruction. One need not use the terminology of the taxonomies when stating learning outcomes, but a review of the various taxonomy categories will aid in the development of a more complete list. The broad range of learning outcomes covered in the *Taxonomy* provides assurance that important types of learning are not overlooked.
GUIDELINES

Begin with a Simple Framework

Starting with a simple framework (Knowledge, Understanding, Application) will help move from factual information to more complex learning outcomes, as illustrated in the following examples:

K = Knowledge
U = Understanding
A = Application

Each of these categories can be expanded with skills and affective outcomes as needed.

Reading

K = Knows vocabulary
U = Reads with comprehension
A = Reads to obtain information to solve a problem

Writing

K = Knows the mechanics of writing
U = Understands grammatical principles in writing
A = Writes to communicate for a specific purpose

Math

K = Knows the number system and basic operations
U = Understands math concepts and processes
A = Uses mathematics to solve problems
Science

K = Knows terms and facts
U = Understands scientific principles
A = Applies principles to new situations

Social Studies

K = Knows factual information about social issues
U = Understands causes of social issues
A = Applies critical thinking skills to social issues

GUIDELINES

Stating Objectives as Learning Outcomes

Don’t state them in terms of:

1. teacher performance (e.g. “Teach students key concepts”)
2. learning process (e.g. “Student learns meaning of concepts”)
3. course content (e.g. “Student studies geometric figures”)
4. two objectives (e.g. “Student knows and understands concepts”)

State them in terms of student performance at the end of instruction.

1. “Knows the meaning of concepts”
   1.1 “Identifies a definition of the concept”
   1.2 “Identifies the concept that fits a given description”
   1.3 “Matches the concept to a picture“ (e.g. geometric figures)
   1.4 “Differentiates between the concept and a second concept”
   1.5 “Applies the concept to an every day situation”
2.1.3 Stating the Specific Learning Outcomes

As stated earlier, each general instructional objective must be defined by a sample of specific learning outcomes to clarify how students can demonstrate that they have achieved the general objectives. Unless the general objectives are further defined in this manner, they will not provide adequate direction for teaching or testing.

Statements of specific learning outcomes for a general objective will be easier to write and will more clearly convey instructional intent if each statement begins with an action verb that indicates definite, observable responses (e.g. identifies, solves, communicates). Such statements specify the types of student performance acceptable as evidence that the general instructional objective has been achieved. This assumes of course, that each specific learning outcome is directly relevant to the general objective it is defining. A statement like "Writes the textbook definition of a principle" would be appropriate for listing under "knows principles", but not under "Understands principles". For the latter objective, we would need a statement that goes beyond the recall of information, because understanding implies some novelty in the response. Here, a statement like "explains the principle in own words" or "applies the principle to solve an applied problem" would be more relevant.

A major problem in defining general instructional objectives is deciding how many specific learning outcomes to list under each objective. It is obvious that a fixed number cannot be specified. Simple knowledge and skill outcomes typically require fewer statements that complex ones. Because it is usually impossible or impractical to list all possible student responses for each general objective, the sample should be as representative as possible. In most cases there is not much advantage in listing more than seven or eight specific learning outcomes for each objective; four or five statements are probably more common. As a general guide, enough learning outcomes should be listed for each objective to show the typical performance of students who have satisfactorily achieved the objective.

The following general objectives and specific learning outcomes illustrate a satisfactory level of specificity for stating the intended learning outcomes.
1. Understand the meaning of concepts

1.1 Explains the concept in own words

1.2 Identifies the meaning of a concept in context

1.3 Differentiates between proper and improper instances of a concept

1.4 Distinguishes between two similar concepts on the basis of meaning

1.5 Uses a concept to explain an everyday event

2. Demonstrates skill in critical thinking

2.1 Distinguishes between fact and opinion

2.2 Distinguishes between relevant and irrelevant information

2.3 Identifies fallacious reasoning in written material

2.4 Identifies the limitations of given data

2.5 Formulated valid conclusions from given data

2.6 Identifies the assumptions underlying conclusions

In addition to illustrating the desired degree of specificity, these statements are good examples of content-free objectives. As noted earlier, both the general objectives and the specific learning outcomes should be kept free of specific content so that they can be used with various units of study. In stating our specific learning outcomes, we are attempting to describe what types of student performance represent each general objective – not what specific content the students are to learn.

Keeping the specific learning outcomes content-free is, of course, a matter of degree. In some cases all we can do is modify our statements so that they apply to a wider range of course material. The following statement illustrate ways to improve specific learning outcomes in this regard:

**EXAMPLES**

<table>
<thead>
<tr>
<th>Poor</th>
<th>Better</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinguishes between a square and a rectangle</td>
<td>Distinguishes among geometric shapes</td>
</tr>
</tbody>
</table>
If we used the first version of each of these specific learning outcomes, we would have to write new statements for each identification, comparison, or description we wanted our students to make. The better versions can be used with various areas of content, thus freeing us from the repetitious writing of objectives as new subject matter is considered. Specific learning outcomes should specify the types of reactions the students are to make to the content, not identify the content itself.

In some cases it may be desirable or necessary to consult reference books and other relevant materials for the types of performance that might represent an objective. When defining such complex outcomes as critical thinking, literary appreciation, and scientific attitude, for example, a review of the literature can be very useful. Although you may not find a detailed list of the specific components of each outcome, even general descriptions of the concepts will aid in defining relevant types of performance. In any event, resist the temptation to omit complex outcomes simply because they are difficult to define.

2.1.4 Clarification of Verbs Used in Specific Learning Outcomes

Because the action verb is a key element in stating the specific learning outcomes, the selection and clarification of these verbs play an important role in obtaining a clearly defined set of instructional objectives. Ideally we would like each verb 1) to convey clearly our instructional intent and 2) to specify precisely the student performance we are willing to accept as evidence that the general objective has been attained. Some verbs convey instructional intent well (e.g. identifies); others are more effective at specifying the student responses to be observed (e.g. labels, encircles, underlines). When it is necessary to choose between these two types of verbs, it is best to select those that most clearly convey instructional intent and then, if necessary, clarify further the expected student responses in one of the following ways:

1. Add a third level of specificity to the list of objectives
2. Define the action verbs used in the specific learning outcomes
3. Use sample test items to illustrate the intended outcomes

These procedures are probably most useful as guides to test construction and for communicating your intended learning outcomes to others.
The meaning of each specific learning outcome can be further clarified by listing some of the
tasks students are expected to perform to demonstrate achievement of the outcome.
This would provide three levels for each instructional objective as follows:

1. Comprehends the meaning of written material.

1.1 Identifies the main thought in a passage

1.11 Underlines the topic sentence

1.12 Selects the most appropriate title for the passage

1.13 Writes the main theme of the passage
3.0 VALIDITY AND RELIABILITY

When constructing or selecting assessments, the most important questions are: 1) to what extent will the interpretation of the scores be appropriate, meaningful, and useful for the intended application of the results? And 2) what are the consequences of the particular uses and interpretation that are made of the results?

Assessments take a wide variety of forms, ranging from the familiar multiple-choice or other type of fixed-response test to extended observations of performance. They also serve a variety of uses in the school, to plan instructional activities, or to communicate progress to students and parents; achievement tests might be used for predicting success in future learning activities or occupations; and appraisals of personal-social development might be used to better understand learning problems or to evaluate the effects of a particular school program. But regardless of the type of assessment used or how the results are to be used, all assessments should possess certain characteristics. The most essential of these are validity, reliability and usability.

Validity refers to the adequacy and appropriateness of the interpretations made from assessments, with regard to a particular use. For example, if an assessment is to be used to describe student achievement, we should like to be able to interpret the scores as a relevant and representative sample of the achievement domain to be measured. If the results are to be used to predict students’ success in some future activity, we should like our interpretations to be based on as good an estimate of future success as possible. If the results are to be used as a measure of students’ reading comprehension, we should like our interpretations to be based on evidence that the scores actually reflect reading comprehension and are not distorted by irrelevant factors. Basically then, validity is always concerned with the specific use of assessment results and the soundness of our proposed interpretations of those results. As well shall see later in the chapter, however, this does not mean that validation procedures can be matched to specific assessment uses on a one-to-one basis.
In recent years, our understanding of validation has also come to include and evaluation of the adequacy and appropriateness of the uses that are made of assessment results. This expanded view of validity leads to a focus on the consequences of particular uses of assessment results. For example, if a state-or district-mandated test led teachers to ignore important content not covered by the test, that consequence would need to be taken into account in judging the validity of that test use.

Reliability refers to the consistency of assessment results. If we obtain quite similar scores when the same assessment procedure is used with the same students on two different occasions, we can conclude that our results have a high degree of reliability from one occasion to another. Similarly, if different teachers independently rate student performances on the same assessment task and obtain similar ratings, we can conclude that the results have a high degree of reliability from one rater to another. Like validity, reliability is intimately related to the type of interpretation to be made. For some uses, we may be interested in asking how reliable our assessment results are over a given period of time and, for others, how reliable they are over different samples of the same behavior. In all instances in which reliability is being determined, however we are concerned with the consistency of the results, rather than with the appropriateness of the interpretations made from the results (validity).

The relation between reliability and validity is sometimes confusing to persons who encounter these terms for the first time. Reliability (consistency) of measurement is needed to obtain valid results but we can have reliability without validity. That is we can have consistent measures that provide the wrong information or are interpreted inappropriately. The target-shooting illustration in Figure 3.1 depicts the concept that reliability is a necessary but not sufficient condition for validity.
In addition to providing results that possess a satisfactory degree of validity and reliability, and assessment procedure must meet certain practical requirements. It should be economical from the viewpoint of both time and money; it should be easily administered and scored; and it should produce results that can be accurately interpreted and applied by available school personnel. These practical aspects of an assessment procedure all can be included under the heading of usability. The term usability, then, refers only to the practically of the procedure and says nothing about the other qualities present.

### 3.1 NATURE OF VALIDITY

When using the term validity in relation to testing and assessment, there are a number of cautions to be borne in mind.

1. **Validity** refers to the *appropriateness of the interpretation of the results* of an assessment procedure for a given group of individuals, not to the procedure itself. We sometimes speak of the 'validity of a test', for the sake of convenience, but it is more correct to speak of the validity of the interpretation and use to be made from the results.

2. Validity is a *matter of degree*; it does not exist on all-or-none basis. Consequently, we should avoid thinking of assessment results as valid or invalid. Validity is best considered in terms of categories that specify degree, such as high validity, moderate validity, and low validity.
3. Validity is always *specific to some particular use or interpretation*. Not assessment is valid for all purposes. For example, the results of an arithmetic test may have a high degree of validity for indicating computational skill, a low degree of validity for indicating arithmetical reasoning, a moderate degree of validity for predicting success in future mathematics courses, and essentially no validity for predicting success in art or music. Thus, when appraising or describing validity, it is necessary to consider the specific interpretation or use to be made of the results. Assessment results are never just valid; they have a different degree of validity for each particular interpretation to be made.

4. Validity involves an *overall evaluative judgment*. It requires an evaluation of the degree to which interpretations and uses of assessment results are justified by supporting evidence and in terms of the consequences of those interpretations and uses.

### 3.1.1 MAJOR CONSIDERATIONS IN ASSESSMENT VALIDATION

Four major considerations for validation are briefly described in Table 3.1. The strongest case for validity can be made when evidence is obtained regarding all four considerations in Table 3.1. That is, interpretations and uses of assessment results are likely to have greater validity when we have an understanding of (1) the assessment content and the specifications from which it was derived, (2) the relation of the assessment results to other significant measures, (3) the nature of the characteristic(s) being measured, and (4) the consequences of the uses and interpretations of the results. However, for many uses of a test or an assessment, it is not practical or necessary to have evidence dealing with all four considerations. For example, it is not practical to expect that a teacher would provide evidence that a classroom assessment designed to measure student learning is related to other significant measures. In this case, the primary concern would be content, but some of the analyses of the meaning of the scores (construct considerations) and possible effects on student motivation and learning (consequences considerations) would be relevant.
### TABLE 3.1 Major considerations in validation

<table>
<thead>
<tr>
<th>Consideration</th>
<th>Procedure</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content</strong></td>
<td>Compare the assessment tasks to the specifications describing the task domain under consideration</td>
<td>How well the sample of assessment tasks represents the domain of tasks to be measured</td>
</tr>
<tr>
<td><strong>Test-criterion Relationship</strong></td>
<td>Compare assessment results with another measure of performance obtained at a later date (for prediction) or with another measure of performance obtained concurrently (for estimating present status).</td>
<td>How well performance on the assessment predicts future performance or estimates current performance on some valued measures other than the test itself (called a criterion).</td>
</tr>
<tr>
<td><strong>Construct</strong></td>
<td>Establish the meaning of the assessment results by controlling (or examining) the development of the assessment, evaluating the relationships of the scores with other relevant measures, and experimentally determining what factors influence performance.</td>
<td>How well performance on the assessment can be interpreted as a meaningful measure of some characteristic or quality</td>
</tr>
<tr>
<td><strong>Consequences</strong></td>
<td>Evaluate the effects of use of assessment results on teachers and students. Both the intended positive effects (e.g. increased learning) and possible unintended negative effects (e.g. narrowing of instruction, drop out of school) need to be evaluated.</td>
<td>How well use of assessment results accomplished intended purposes and avoids unintended effects.</td>
</tr>
</tbody>
</table>

The essence of the content consideration in validation, then, is determining the adequacy of the sampling of the content that the assessment results are interpreted to represent. More formally, \textit{the goal in the consideration of content validation is to determine the extent to which a set of assessment tasks provides a relevant and representative sample of the domain of tasks about which interpretations of assessment results are made.} In classroom assessment, the domains of achievement tasks are determined by the curriculum and instruction, and assessment development involves (1) clearly specifying the domain of instructionally relevant tasks to be
used to measure student achievement and (2) constructing or selecting a representative set of assessment tasks.

Thus, to obtain a valid measure of learning outcomes, we proceed from the instruction (what has been taught) to the achievement domain (what is to be measured) and finally to the assessment itself (a representative sample of relevant tasks). As shown in Figure 3.2, content considerations in validation require a judgment that all three are in close harmony.

**FIGURE 3.2 Content considerations in the assessment of classroom achievement**

<table>
<thead>
<tr>
<th>Classroom Instruction</th>
<th>Achievement Domain</th>
<th>Achievement Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determines which intended learning outcomes (objectives) are to be achieved by students</td>
<td>Specifies and delimits a set of instructionally relevant learning tasks to be measured by an assessment</td>
<td>Provides a set of relevant assessment tasks designed to measure a representative sample of the tasks in the achievement domain</td>
</tr>
</tbody>
</table>

3.1.2 **Content Considerations on Assessment Development to Enhance Validity**

Content issues are typically considered during the development of an assessment. It is primarily a matter of preparing detailed specifications and then constructing an assessment that meets these specifications. Although there are many ways of specifying what an assessment should measure, one widely used procedure in constructing achievement tests uses a two-way chart called a *table of specifications*. We shall use a brief form of it here to help clarify the process of content validation in preparing classroom assessments.
**Table of Specifications.** The learning outcomes of a course or curriculum may be broadly defined to include both subject matter content and instructional objectives. The former is concerned with the topics to be learned and the latter with the types of performance students are expected to demonstrate (e.g., knows, comprehends, applies). Both of these aspects are of concern in defining content domain and assuring adequate sampling from it. We should like any assessment of achievement that we construct to produce results that represent both the content areas and the objectives we wish to measure, and the table of specifications aids in obtaining a sample of tasks that represent both.

**TABLE 3.2** Sample table of specifications.

<table>
<thead>
<tr>
<th>Content Area</th>
<th>Instructional Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Knows Concepts</td>
</tr>
<tr>
<td>Air pressure</td>
<td>4</td>
</tr>
<tr>
<td>Air temperature</td>
<td>4</td>
</tr>
<tr>
<td>Humidity and Precipitation</td>
<td>8</td>
</tr>
<tr>
<td>Wind</td>
<td>4</td>
</tr>
<tr>
<td>Clouds</td>
<td>8</td>
</tr>
<tr>
<td>Fronts</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>32</strong></td>
</tr>
</tbody>
</table>

A table of specifications, in a very simple form, is presented in Table 3.2 to show how such a table is used in test development. The percentages in the table indicate the relative degree of emphasis that each content area and each instructional objective is to be given in the test.
Thus, if a 25-item classroom test is to measure a representative sample of subject-matter content, 16 percent of the items (i.e. 4 items) should be concerned with air pressure, 12 percent with air, 24 percent with humidity and precipitation, 12 percent with the wind, 20 percent with clouds, and 16 percent with fronts. Similarly, if the test is to measure a representative sample of the instructional objectives, 32 percent of the items (i.e., 8 items) should measure knowledge of concepts, 48 percent should measure comprehension of concepts, and 20 percent should measure application of concepts. This, of course, implies that the emphasis on knowledge, comprehension, and application for each content area will follow the percentages in the table of specifications. For example, 4 percent of the test items, (i.e., a single item) concerned with air pressure should measure knowledge of concepts, 8 percent should measure comprehension of concepts, and 4 percent should measure application of concepts.

As noted earlier, the specifications describing the achievement domain to be measured should be in harmony with what was taught. Thus, the weights assigned in this table reflect the emphasis that was given during instruction. For example, comprehension outcomes received more emphasis than either knowledge or application outcomes in the instruction and therefore were given almost as much weight in the table as the other two categories combined. The table, then, indicates the sample of instructionally relevant learning tasks to be measured, and the more closely the test items correspond to the specified sample, the greater the likelihood of obtaining a valid measure of student learning.

The test items and other kinds of assessment tasks must function as intended if valid results are to be obtained. Test items and assessment tasks may function improperly if they contain inappropriate vocabulary, unclear directions, or some other defect. Similarly, tasks designed to measure understanding and application may measure only the simple recall of information if the solutions to the problems have been directly taught during instruction. In short, a host of factors can influence the intended function of the tasks and thus the validity of the assessment results.
3.1.3 FACTORS INFLUENCING VALIDITY

Numerous factors tend to make assessment results invalid for their intended use. Some are rather obvious and easily avoided. No teacher would think of measuring knowledge of social studies with a mathematics assessment. Nor would a teacher consider measuring problem-solving skills in third-grade arithmetic with an assessment designed for seventh graders. In both instances, the assessment results would obviously be invalid. The factors influencing validity are of this same general nature but much more subtle in character. For example, a teacher may overload a social studies test with items concerning historical facts, and thus the score are less valid as a measure of achievement in social studies. Or a third-grade teacher may select appropriate arithmetic problems for an assessment but use vocabulary in the problems and directions that only the better readers are able to understand. The arithmetic assessment then becomes, in part, a reading assessment and reduces the validity of the results for their intended use. These examples show some of the more subtle factors influencing validity to which the teacher should be alert, whether constructing classroom assessments or selecting published ones.

3.1.3 Factors in the Test of Assessment Itself

A careful examination of test items and assessment tasks will indicate whether the tests or assessment appears to measure the subject-matter content and the mental functions that the teacher is interested in assessing. However, any of the following factors can prevent the tests items or assessment tasks from functioning as intended and thereby lower the validity of the interpretations from the assessment results. The first four factors are equally applicable for assessments with tasks requiring extended student performances and traditional tests. The last six factors apply most directly to tests with fixed-choice or short-answer items that are scored right or wrong.
1. **Unclear directions.** Directions that do not clearly indicate to the student how to respond to the tasks and how to record the responses will tend to reduce validity.

2. **Reading vocabulary and sentence structure too difficult.** Vocabulary and sentence structure that are too complicated for the students taking the assessment will result in the assessment’s measuring reading comprehension and aspects of intelligence, which will distort the meaning of the assessment results.

3. **Ambiguity.** Ambiguous statements in assessment tasks contribute to misinterpretations and confusions. Ambiguity sometimes confuses the better students more than it does the poor students.

4. **Inadequate time limits.** Time limits that do not provide students with enough time to consider the tasks and provide thoughtful responses can reduce the validity of interpretations of results. Rather than measuring what a student knows about a topic or is able to do given adequate time, the assessment may become a measure of the speed with which the student can respond. For some content (e.g. a typing test), speed may be important. However, most assessments of achievement should minimize of speed on student performance.

5. **Inappropriate level of difficulty of the test items.** In norm-referenced test, items that are too easy or too difficult will not provide reliable discrimination among students and will therefore lower validity. In criterion-referenced tests, the failure to match the difficulty specified by the learning outcome will lower validity.

6. **Poorly constructed test items.** Test items that unintentionally provide clues to the answer will tend to measure the students’ alertness in detecting clues as well as mastery of skills or knowledge the test is intended to measure.

7. **Test items inappropriate for the outcomes being measured.** Attempting to measure understanding, thinking skills, and other complex types of achievement with test forms that are appropriate only for measuring factual knowledge will invalidate the results.
8. **Test too short.** A test is only a sample of the many questions that might be asked. If a test is too short to provide a representative sample of the performance we are interested in, its validity will suffer accordingly.

9. **Improper arrangement of items.** Test items are typically arranged in order of difficulty, with the easiest items first. Placing difficult items early in the test may cause students to spend too much time on these and prevent them from reaching items they could easily answer. Improper arrangement may also influence validity by having a detrimental effect on student motivation. This influence is likely to be strongest with young students.

10. **Identifiable pattern of answers.** Placing correct answers in some systematic pattern (e.g. T, T, F, F, or A, B, C, D, A, B, C, D) will enable students to guess the answers to some items more easily, and this will lower validity.

In short, any defect in the construction of the test or assessment that prevents it from functioning as intended will invalidate the interpretations to be drawn from the results. Much of what is written in the following chapters is directed toward helping teachers improve the validity of their interpretations of test scores and other assessment results.

### 3.1.3.1 Functioning of Tasks and Teaching Procedures

In assessing achievement, the ways in which tasks function cannot be determined merely be examining the form and content of the assessment. For example, the following task may appear to measure arithmetical reasoning if examined without reference to what the students have been taught:

**EXAMPLE :** Kassim learned that 60 percent of the graduates from his secondary school are admitted to community colleges, 30 percent are admitted to public four-year colleges or universities, and 10 percent are admitted to private four-year colleges or universities. Since these percentages add up to 100, Kassim reasons that he is sure to get admitted to some college or university. Explain, possibly with the use of a diagram, what is wrong with Kassim’s reasoning.
However, if the teacher has taught the solution to this particular problem before giving the assessment, the task now will measure no more than memorized knowledge. Similarly, assessments of understanding, critical thinking, and other complex learning outcomes will provide valid measures in these areas only if the tasks function as intended. If the students have previously been taught the solutions to the particular problems included in the assessment, or have been taught mechanical steps for obtaining the solutions, the assessment results cannot be considered valid indicators of the achievement of the more complex mental processes.

3.1.3.2 Factors in Administration and Scoring

The administration and scoring of an assessment may also introduce factors that have a detrimental effect on the validity of the interpretations from the results. In the case of teacher-made assessments, such factors as insufficient time, unfair aid to individual students who ask for help, cheating, and the unreliable scoring of student performances tend to lower validity. In the case of published tests, failure to follow the standard directions and time limits, giving students unauthorized assistance, and errors in scoring similarly contribute to lower validity. For all types of assessments, adverse physically and psychological conditions at the time of the assessment may also have a negative effect.

3.1.3.3 Factors in Student Responses

In some instances, invalid interpretations are caused by personal factors influencing a student’s response to the assessment situation rather than to any shortcomings in the test instrument or its administration. Some students may be bothered by emotional disturbances that interfere with their performance. Others may be frightened by the assessment situation and thereby are unable to respond normally, and still others may not be motivated to put forth their best effort. These and other factors that restrict and modify students’ in the assessment situation will obviously distort the results.
3.1.3.4 Nature of the Group and the Criterion

Validity is always specific to a particular group. An arithmetic test based on story problems, for example, may measure reasoning ability in a group at an early stage of instruction and a combination of simple recall of information and computation skill in a more advanced group. Similarly, scores on a science assessment may be accounted for largely by reading comprehension in one group and by knowledge of facts in another. What an assessment measures may be influenced by such factors as age, gender, ability level, educational background, and cultural background. Thus, in appraising reports of validity included in test manuals or other sources, it is in significant characteristics with the group of students we wish to assess determines how applicable the information is to our particular group.

Next to validity, reliability is the most important characteristic of assessment results. Reliability (1) provides the consistency that makes validity possible, and (2) indicates the degree to which various kinds of generalizations are justifiable. The practicality of the evaluation procedure is, of course, also of concern to the busy classroom teacher.

3.2 NATURE OF RELIABILITY

Reliability refers to the consistency of measurement; that is, how consistent test scores or other assessment results are from one measurement to another. Suppose, for instance, that Miss Jones has just given an achievement assessment to her students. How similar would the students’ scores have been had she assessed them yesterday, or tomorrow, or next week? How would the scores have varied had she selected a different sample of tasks? How much would the scores have differed had a different teacher scored it? These are the types of questions with which reliability is concerned. Assessment results merely provide a limited measure of performance obtained at a particular time. Unless the measurement can be shown to be reasonably consistent (that is, generalizable) over different occasions, different raters, or different samples of the same performance domain, we can have little confidence in the results.
We cannot expect assessment results to be perfectly consistent. There are numerous factors other than the quality being measured that may influence assessment results. If a single assessment is administered to the same group twice in close succession, some variation in scores can be expected because of temporary fluctuations in memory, attention, effort, fatigue, emotional strain, guessing and the like. With a longer time between tests, additional variation in scores may be caused by intervening learning experiences, changes in health, forgetting, and less comparable assessment conditions. If essay or other types of student performances are evaluated by different raters, some variation in scores can be expected due to less-than-perfect agreement among raters. If we use a different sample of tasks in the second assessment, still another factor is likely to influence the results. Individuals may find one assessment easier than the other because it happens to contain more tasks on topics with which they are familiar. Such extraneous factors as these introduce a certain amount of measurement error into all assessment results. Methods of determining reliability are essentially means of determining how much measurement error is present under different conditions. In general, the more consistent our assessment results are from one measurement to another, the less error there will be and, consequently, the greater the reliability.

The meaning of reliability, as applied to testing and assessment, can be further clarified by nothing the following general points:

1. **Reliability refers to the results obtained with an assessment instrument and not to the instrument itself.** Any particular instrument may have a number of different reliabilities, depending on the group involved and the situation in which it is used. Thus, it is more appropriate to speak of the reliability of the test scores or of the assessment results than of the test of the assessment.

2. **An estimate of reliability always refers to a particular type of consistency.** Assessment results are not reliable in general. They are reliable (or generalizable) over different periods of time, over different samples of tasks, over different raters, and the like. It is possible for assessment results to be consistent in one of these respects and not in another. The appropriate type of consistency in a particular case is dictated by the use to be made of the results.
For example, if we wish to know what individuals will be like at some future time, constancy of scores over time will be important. On the other hand, if we want to measure an individual’s current understanding of certain scientific principles we are apt to be more interested in the consistency of performance across different tasks designed to allow students to apply those principles. Thus, for different interpretations we need different analyses of consistency. Treating reliability as a general characteristic can lead to erroneous interpretations.

3. **Reliability is a necessary but not sufficient condition for validity.** As an assessment that produces totally inconsistent results cannot possibly provide valid information about the performance being measured. On the other hand, highly consistent assessment results may be measuring the wrong thing or may be used in inappropriate ways. Thus, low reliability indicates that a low degree of validity is present, but high reliability does not ensure a high degree of validity. In short, reliability merely provides the consistency that makes validity possible.

4. **Reliability is primarily statistical.** The logical analysis of an assessment will provide little evidence concerning the reliability of the scores. To evaluate the consistency of scores assigned by different raters, two or more raters must score the same set of student performances. Similarly, an evaluation of the consistency of scores obtained in response to different forms of a test or different collections of performance-based assessment tasks requires the administration of both test forms or collections of tasks to an appropriate group of students. Whether the focus is on inter-rater consistency or the consistency across forms or collections of tasks, consistency may be expressed in terms of shifts in the relative standing of persons in the group or in terms of the amount of variation to be expected in an individual’s score. Consistency in the first case is reported by means of a correlation coefficient called a reliability coefficient (see “Terminology” box) and in the second case is reported by means of the standard error of measurement. Both methods of expressing reliability are widely used and should be understood by persons responsible for interpreting assessment results.
3.2.1 DETERMINING RELIABILITY BY CORRELATION METHODS

In determining reliability, it would be desirable to obtain two sets of measures under identical conditions and then to compare the results. This procedure is impossible, of course, because the conditions under which assessment data are obtained can never be identical. As a substitute for this ideal procedure, several methods of estimating reliability have been introduced (American Psychological Association, 1985). The methods are similar in that all of them involve correlating two sets of scores, obtained either from the same assessment procedure or from equivalent forms of the same procedure. The correlation coefficient used to determine reliability is calculated and interpreted in the same manner as that used in determining statistical estimates of validity. The only difference between a validity coefficient and reliability and coefficient is that the former is based on agreement with an outside criterion and the latter is based on agreement between two sets of results from the same procedure.

<table>
<thead>
<tr>
<th>TERMINOLOGY</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Correlation Coefficient</strong> : A statistic that indicates the degree of relationship between any two sets of scores obtained from the same group of individuals (e.g., correlation between height and weight)</td>
</tr>
<tr>
<td><strong>Validity Coefficient</strong> : A correlation coefficient that indicates the degree to which a measure predicts or estimated performance on some criterion measure (e.g. correlation between scholastic aptitude scores and grades in school)</td>
</tr>
<tr>
<td><strong>Reliability Coefficient</strong> : A correlation coefficient that indicates the degree of relationship between two sets of scores intended to be measure of the same characteristic (e.g. correlation between scores assigned by to different raters or scores obtained from administrations of two forms of a test).</td>
</tr>
</tbody>
</table>
The chief methods of estimating reliability are shown in Table 3.3. Note that different types of consistency are determined by the different methods: consistency over a period of time, over different forms of the assessment, within the assessment itself, and over different raters. The reliability coefficient resulting from each method must be interpreted according to the type of consistency being investigated. Each of these methods of estimating reliability will be considered in further detail as we proceed.

3.2.1.1 Test-Retest Method

To estimate reliability by means of the test-retest method, the same assessment is administered twice to the same group of students with a given time interval between the two administrations (see Figure 3.3). The resulting assessment scores are correlated, and this correlation coefficient provides a measure of stability; that is, it indicates how stable the assessment results are over the given period of time. If the results are highly stable, those students who are high on one administration of the assessment will tend to be high on the other administration, and the remaining students will tend to stay in their same relative positions on both administrations. Such stability is indicated by a large correlation coefficient. Recall from our previous discussion of correlation coefficients that a perfect positive relationship is indicated by 1:00 and no relationship by 0.00. Measures of stability in the .80 range are commonly reported for standardized tests of aptitude and achievement over occasions within the same year.
### TABLE 3.3 Methods of estimating reliability

<table>
<thead>
<tr>
<th>Method</th>
<th>Type of Reliability Measure</th>
<th>Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-retest Method</td>
<td>Measure stability</td>
<td>Give the same test twice to the same group with any time interval between tests, from several minutes to several years</td>
</tr>
<tr>
<td>Equivalent forms method</td>
<td>Measure of equivalence</td>
<td>Give two forms of the test to the same group in close succession</td>
</tr>
<tr>
<td>(Test-retest with equivalents forms)</td>
<td>Measure of stability and equivalence</td>
<td>Give two forms of the test to the same group with increased time interval between forms</td>
</tr>
<tr>
<td>Split-half method</td>
<td>Measure of internal consistency</td>
<td>Give test once. Score two equivalent halves of test (e.g. odd items and even items); correct correlation between halves to fit whole test by Spearman-Brown formula</td>
</tr>
<tr>
<td>Kuder-Richardson method and coefficient Alpha</td>
<td>Measure of internal consistency</td>
<td>Give test once. Score total test and apply Kuder-Richardson formula</td>
</tr>
<tr>
<td>Inter-rater method</td>
<td>Measure of consistency of ratings</td>
<td>Give a set of student responses requiring judgmental scoring to two or more raters and have them independently score the responses</td>
</tr>
</tbody>
</table>

One important factor to keep in mind when interpreting measures of stability is the time interval between assessments. If the time interval is short, say a day or two, the constancy of the results will be inflated because students will remember the tasks and their responses to them from the first assessment. If the time interval is long, say about a year, the results will be influenced not only by the instability of the assessment procedure but also by actual changes in the students over that period of time. In general, the longer the interval between the first and second assessments, the more the results will be influenced by changes in the student characteristic being measured, and the smaller the reliability coefficient will be.
The best time interval between assessment administrations will depend largely on the use to be made of the results. Because college admissions test scores may be submitted as part of an application to college several years after the test was taken, stability over several years is quite important. But stability over a long period is neither important nor desirable for an assessment of performance on a unit in a course that is focused on student understanding of certain concepts and readiness to move on to new material. Thus, for some decisions we are interested in reliability coefficients based on a long interval between test and retest; for others, reliability coefficients based on a short interval may be sufficient. The important thing is to seek evidence of stability that fits the particular interpretation to be made.

The test-retest method is unlikely to be relevant for teacher-constructed classroom tests, because it is seldom possible or desirable to readminister the same assessment. But in choosing standardized test, stability is an important criterion. The test manual should provide evidence of stability, indicating the interval between tests and any unusual experience the group members might have had between testings. Other things (such as validity) being equal, a test shown to possess the type of stability needed to make sound decisions is the best test.

Stability is also an important consideration when using assessment results from school records. When using any assessment from permanent records, check the date of the assessment and any stability data available to determine whether the results are still dependable. If there is a doubt, and the decision is important, a reassessment is in order.

**FIGURE 3.1** Test-retest method (using same test forms)

<table>
<thead>
<tr>
<th>Test-Retest</th>
<th>September 25</th>
<th>October 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form A</td>
<td>Form A</td>
<td></td>
</tr>
<tr>
<td>1. hocides</td>
<td>1. hocides</td>
<td></td>
</tr>
<tr>
<td>2. hodie od</td>
<td>2. hodie od</td>
<td></td>
</tr>
<tr>
<td>3. kodesiciose</td>
<td>3. kodesiciose</td>
<td></td>
</tr>
<tr>
<td>4. kodesodico</td>
<td>4. kodesodico</td>
<td></td>
</tr>
</tbody>
</table>

Score = 82
Score = 86
3.2.1.2 Equivalent Forms Method

The equivalent forms method for estimating reliability uses two different but equivalent forms of an assessment (also called parallel or alternate forms). Equivalent forms are built to the same set of specifications (e.g. test content and difficulty but are constructed independently. The two forms of the assessment are administered to the same group of students in close succession, and the resulting assessment scores are correlated. This correlation coefficient provides a measure of the degree to which generalizations about student performance from one assessment to another are justified. Thus, it indicates the degree to which the two assessments are measuring the same aspects of behavior.

The equivalent forms method tells us nothing about the long-term stability of the student characteristic being measured. Rather, it reflects short-term constancy of student performance and the extent to which the assessment represents an adequate sample of the characteristic being measured. In assessing achievement, for example, thousands of tasks might be presented in a particular assessment, but because of time limits and other restricting factors, only some of the possible tasks can be used. The tasks included in the assessment should provide an adequate sample of the possible tasks in the area. The easiest way to estimate whether an assessment measures an adequate sample of the content is to construct versions of the assessment that are intended to cover the same domain of content and student skills and correlate the results. A high correlation indicates that the two assessments are providing similar results and, therefore, are probably reliable samples of the general area of content being measured.

The equivalent forms method of estimating reliability is widely used in standardized testing because most standardized tests have two or more forms available. In fact, a teacher should be suspicious of any standardized test that has two forms available and does not provide information about equivalence. The comparability of the results of the two forms cannot be assumed unless such evidence is presented.
The equivalent forms method is sometimes used with an interval between the administration of the two forms of the test (see Figure 3.2). Under these test-retest conditions, the resulting reliability coefficient provides a measure of stability and equivalence. This is a more rigorous test of reliability than the test-retest method or equivalent forms method with a short interval between forms, because the stability of the testing procedures, the constancy of the student characteristic being measured, and the representativeness of the sample of tasks included in the test all are taken into account. Consequently, this is generally recommended as the soundest procedure for estimating the reliability of test scores. As with the ordinary test-retest method, the reliability coefficient must be interpreted in light of the interval between the two forms of the test. For longer time periods, expect smaller reliability coefficients.

**FIGURE 3.2 Equivalent forms method (without and with a time interval)**

<table>
<thead>
<tr>
<th>Equivalent Forms</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>September 25</td>
<td>September 25</td>
<td></td>
</tr>
<tr>
<td>Form A</td>
<td>Form B</td>
<td></td>
</tr>
<tr>
<td>1. josodosh</td>
<td>1. kodsoodios</td>
<td></td>
</tr>
<tr>
<td>2. kodso-koshris</td>
<td>2. kodsoedi</td>
<td></td>
</tr>
<tr>
<td>3. jiodsosids</td>
<td>3. kodsoosdios</td>
<td></td>
</tr>
<tr>
<td>4. kodesosodis</td>
<td>4. kodesode</td>
<td></td>
</tr>
<tr>
<td>Score = 82</td>
<td>Score = 78</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test-Retest with Equivalent Forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 25</td>
</tr>
<tr>
<td>Form A</td>
</tr>
<tr>
<td>1. josodosh</td>
</tr>
<tr>
<td>2. kodso-koshris</td>
</tr>
</tbody>
</table>
3.2.1.3 Split-Half method

Reliability can also be estimated from a single administration of a single form of an assessment. The assessment is administered to a group of students in the usual manner and then is divided in half form scoring purposes. The split-half method is easy to implement with a traditional test or quiz consisting of say, ten or more items. To split the test into halves that are equivalent, the usual procedure is to score the even-numbered and the odd-numbered tasks separately (see Figure 3.3). This produces two scores for each student, which, when correlated, provide a measure of internal consistency. This coefficient indicates the degree to which consistent results are obtained from the two halves of the test, and may be thought of as the “half-length test reliability estimate”.

Although most often applied with traditional test containing a sizable number of items, the split-half method is also applicable to assessments consisting of a smaller number of tasks each, or which require a more extended period of time to complete. Some limited information can be obtained with as few as two tasks by correlating the scores obtained on one task with those obtained on another. Better information is provided as the number of tasks increases. With a small number of tasks, say four to eight, it is better to divide the tasks into two sets that are judged to be most comparable to each other rather than relying on the more mechanical odd-even method that is appropriate where a larger number of tasks or items are available.
FIGURE 3.3 Odd-even scoring for use of split-half method (odd and even scores are then correlated)

<table>
<thead>
<tr>
<th>Sum number of odd items correct</th>
<th>Sum number of even items correct</th>
<th>September 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items 1</td>
<td>Items 2</td>
<td>1. nwso rho</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>2. nworh orh</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>3. stpey</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>4. ta la de</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>5. fsi ley</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>6. odd iteets</td>
</tr>
<tr>
<td>Odd score = 40</td>
<td>Even score = 42</td>
<td>Total score = 82</td>
</tr>
</tbody>
</table>

As previously noted, the reliability coefficient is determined by correlating the scores of two half-assessments. To estimate the scores’ reliability based on the full-length assessment, the Spearman-Brown formula is usually applied:

\[
\text{Reliability on full assessment} = \frac{2 \times \text{correlation between half-assessments}}{1 + \text{correlation between half-assessments}}
\]

The simplicity of the formula can be seen in the following example, in which the correlation coefficient between scores on the assessment’s two halves is .60.:

\[
\text{Reliability on full assessment} = \frac{2 \times .60}{1 + .60} = \frac{1.20}{1.60} = .75
\]

This correlation coefficient of .75, then, estimates the reliability of a full assessment when the half-assessments correlated .60.
The split-half method is similar to the equivalent form method in that it indicates the extent to which the sample of tasks is a dependable sample of the content being measured. A high correlation between scores on the two halves of an assessment denotes the equivalence of the two halves and consequently the adequacy of the sampling. Split-half method is based on the administration of a single assessment. When two forms of an assessment are administered, even in close succession, more opportunity for inconsistency is introduced (e.g. differences from form to form in attention, speed of work, effort, fatigue, and assessment content). Because the equivalent forms method takes into account more sources of inconsistency, it provides more stringent evaluation of reliability.

3.2.1.4 Kuder-Richardson Method and Coefficient Alpha

Another method of estimating the reliability of assessment scores from a single administration is by means of formulas such as those developed by Kuder and Richardson. As with the split-half method, these formulas provide an index of internal consistency but do not require splitting the assessment in half for scoring purposes. One of the formulas, called Kuder-Richardson Formula 20, or KR-20, is applicable only in situations where student responses are scored dichotomously (zero or one), and therefore is most useful with traditional test items that scored as right or wrong. The KR-20 is based on the proportion of persons passing each item and the standard deviation of the total scores. The generalization of the KR-20 for assessments that have more than dichotomous, right-wrong scores (e.g. each task is scored on a 5-point scale), is called Coefficient Alpha. The computation of either KR-20 or Coefficient Alpha is rather cumbersome unless information is already available concerning the proportion passing or the standard deviations of scores for individual tasks. However, with the increasing availability of microcomputers and software for scoring and analyzing tests, it is now more feasible to obtain KR-20 and Coefficient Alpha estimates of reliability.
But here, our concern is with interpretation of the coefficients rather than the calculations. KR-20 and Coefficient Alpha estimates of reliability provide information about the degree to which the items or tasks in the assessment measure similar characteristics. For a test with relatively homogeneous content (e.g. an arithmetic computation test), the reliability estimate generally will be similar to that provided by the split-half method. Indeed the KR-20, or Coefficient Alpha estimate can be thought of as the average of all possible split-half coefficients for the groups tested. This is an advantage when considering as assessment with relatively homogeneous content which the estimate does not depend on the way in which the items are assigned to the two half-tests as in the split-half method. For assessment designed to measure more heterogeneous learning outcomes (e.g. an assessment covering ancient history, the Middle Ages, and modern history), however, the KR-20 or Coefficient Alpha estimate will be smaller than that provided by the split-half method and the latter method is to be preferred.

The simplicity of applying the split-half method, the KR-20 or the Coefficient Alpha has led to their widespread use in determining reliability. There are limitations, however, that restrict their value. First, they are not appropriate for speeded assessments – for assessments with time limits that prevent students from attempting every task. If speed is a significant factor in the assessment, reliability estimates will be inflated to an unknown degree. This poses no great problem in estimating the reliability of results from teacher-made assessments, because these are usually designed to allow students adequate time to complete all of the tasks. In the case of standardized tests, however, time limits are seldom so liberal that all students complete the test. Thus, measures of internal consistency reported in test manuals should be interpreted with caution unless evidence is also presented that speed of work is a negligible factor. For speeded tests, reliability obtained by the test-retest or equivalent forms method should be sought.
A second limitation of internal consistency procedures is that they do not indicate the consistency of student response from day to day. In this regard, they are similar to the equivalent forms method without a time interval. Only test-retest procedures indicate the extent to which test results are generalizable over different periods of time.

### 3.2.1.5 Inter-Rater Consistency

Judgment is required in scoring student responses to many types of assessments. This is obvious in the case of essay exams, but judgment is also required in scoring responses to open-ended mathematics problems or laboratory exercises in science. Whenever student work is judgmentally scored, it is reasonable to ask whether the same scores would be assigned by another equally qualified judge. Individual classroom teachers seldom have the luxury of having another teacher to independently score examples of student work from their classroom. But there is a growing need to evaluate inter-rater consistency as a result of the increasing numbers of state and district-mandated assessments that must be judgmentally scored.

Estimation of inter-rater consistency is relatively straightforward. Two or more raters must independently score the performances obtained for an appropriately selected sample of students. Consistency can be evaluated by correlating the scores assigned by one judge with those assigned by another judge. Consistency can also be evaluated by computing the proportion of times that students’ performances receive exactly the same scores from a pair of raters and the proportion that are within a single point of each other.

Suppose, the example, that two raters independently scored fifty student essays on a six-point scale. The results of the ratings are summarized in Table 4.2. Rater 1 assigned a score of 6 to five of the essays. For those five essays, Rater 2 assigned a score of 6 to three of them and scores of 4 and 5 to the other two. Thus, each entry in a cell of Table 4.2 show the number of essays that Rater 1 assigned with column of the table. The percentage of exact agreements is obtained by summing the counts where both raters assigned the same score
Percent exact agreement = 100 x \((3+7+5+4+2+3)/50\) = 48%

The percentage of times that raters agreed to within one point (all the counts between the two diagonal lines) would be computed in a similar fashion. For this example the percentage of agreement within one score point is 88 percent (100 times 44/50). One other indicator is the correlation between the two sets of scores, which for the data in the Table 4.2 is .80. The level of inter-rater consistency shown in Table 4.2 is high in comparison to levels of consistency generally achieved by independent raters. Even so, six of the fifty students would have their scores fluctuate by 2 points, depending on the person doing the scoring. One way to reduce the influence of raters is to have each performance independently scored by two or more raters and use the average rating. Double scoring is a common practice in situations where judgmentally scored performances have important consequences for individuals.

**TABLE 4.2** Scores assigned to fifty essays by two independent raters (Raters 1 and 2)

<table>
<thead>
<tr>
<th>Scores Assigned by Rater 2</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score of assigned scores</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Scores by Rater 1</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Scores by Rater 2</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>Column total</td>
<td>4</td>
<td>11</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>7</td>
<td>50</td>
</tr>
</tbody>
</table>
The percentage of agreement and correlation values indicate the degree to which the relative ordering of responses is consistent from one rater to another. It is possible, however, to have a high correlation, and even a high percentage of agreement, while still having important differences in the overall leniency of the two raters. Where there are disagreements, there is a strong tendency for one rater to consistently give a higher score than the other. Comparisons of the average score assigned by each rater provides a check on differences in leniency. For the example in Table 4.2 the average ratings assigned to the essays are reasonably similar (3.52 for Rater 1 and 3.60 for Rater 2).

Thus, there is not cause for concern in this case. The results shown in Table 4.3 (which might have been obtained if Rater 1 had been paired with Rater 3 instead of Rater 2), on the other hand, clearly indicate that Rater 3 is much more lenient that Rater 1, despite the fact that the percentage of exact agreement and the percentage of agreement within one point are the same (48 percent and 88 percent, respectively) for the date in Table 4.3 as they are for the date in Table 4.2.

The evaluation of inter-rater consistency is important to ensure that some students do not receive high scores as the result of rater leniency while others receive low scores because their work was scored by a stringent rater. Average scores assigned by raters to a common set of responses, percentage of agreement, and the correlation between scores assigned by rater pairs all contribute to an overall evaluation of the degree of consistency among different raters. Achieving a high degree of inter-rater consistency requires the development of consensus among raters regarding the types of performances that are valued. Agreed-upon scoring rubrics and training of raters to use those rubrics with examples of student work are generally required to achieve acceptable levels of inter-rater consistency and ensure that differences in the stringency of rating from one rater to another do not place some students at a disadvantage.
Achieving inter-rater consistency is important for judgmentally scored tasks, but it does not say anything about other types of consistency. For example, a high degree of inter-rater consistency does not guarantee consistency of ratings across tasks. There are other methods of evaluating various types of consistency or generalizability (e.g. across raters, across tasks and over time) simultaneously. Generalizability theory provides the foundation for those methods.

### 3.2.1.6 Comparing Methods
As stated earlier, each of the methods of estimating reliability provides different information concerning the consistency of test results. A summary of this information is presented in Table 3.4, which shows that most methods are concerned with only one or two types of consistency. As its name suggests, the inter-rater method only evaluates the degree of consistency of scores assigned to the same performances by different raters. The test-retest method, without a time interval, takes into account only the consistency of the assessment procedure and the short-term constancy of the response. If a time interval is introduced between the assessments, the constancy of the characteristics of the student from day to day also will be included. However, neither of the test-retest

### TABLE 4.3 Scores assigned to fifty essays by two independent raters (Raters 1 and 3)

<table>
<thead>
<tr>
<th>Score</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scores assigned by Rater 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>assigned by Rater 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>Column total</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>7</td>
<td>9</td>
<td>15</td>
<td>50</td>
</tr>
</tbody>
</table>
procedures provides information concerning the consistency of results over different samples of tasks, because both sets of scores are based on the same assessment.

The equivalent forms method without a time interval, the split-half method, and the Kuder-Richardson or Coefficient Alpha method all take into account the consistency of assessment procedures and the consistency of results over different samples of tasks. Only the equivalent forms method with an intervening time period between tests takes into account all three types of consistency, which is the reason that this measure of stability and equivalence is generally regarded as the most useful estimate of reliability.

**TABLE 3.4** Type of consistency indicated by each of the methods for estimating reliability

<table>
<thead>
<tr>
<th>Method of estimating reliability</th>
<th>Consistency of testing procedure</th>
<th>Constancy of student characteristics</th>
<th>Consistency over different samples of items</th>
<th>Consistency of judgment scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-retest (immediate)</td>
<td>X</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test-retest (time interval)</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equivalent forms (immediate)</td>
<td>X</td>
<td>*</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Equivalent forms (time interval)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Split-half</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Kuder-Richardson (Coefficient Alpha)</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Inter-rater</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

*short-term constancy of response is reflected in immediate retest, but day-to-day stability is not shown.*
3.2.2 Methods of Estimating Reliability

When examining the reliability coefficients of standardized tests, it is important to consider the methods used to obtain the reliability estimates. In general, the size of the reliability coefficient is related to the method of estimating reliability.

1. Test-retest method  May be larger than with the split-half method if the time interval is short. Coefficients become smaller as the time interval between tests is increased.

2. Equivalent forms method (without time interval)  Coefficients tend to be lower than with the split-half method of the test-retest method using a short time interval.

3. Equivalent forms method (with time interval)  Coefficients become smaller as the time interval between tests is increased.

4. Split-half method  (e.g. odd-even)  Provides an indication of the internal consistency of a test. Spuriously high estimates are produced for speeded tests.

5. Kuder-Richardson method (Coefficient Alpha)  Typically provides reliability estimates that are smaller than those obtained by the split-half method. These estimates are also inflated by speed.

6. Inter-rater method  Provides in indication of the degree to which similar scores are obtained regardless of who does the rating. Inter-rater consistency can be increased by using well-defined scoring rules and by careful training of raters.

3.2.2.1 How High Should Reliability Be?

The degree of reliability we demand in our educational assessments depends largely on the decision to be made. If we are going to use assessment results to decide whether to review certain areas of subject matter, we may be willing to use a teacher-made assessment of relatively low reliability. Our decision will be based on the scores of the total group, and variation in individual scores will not distort our decision too much. Even if we do err in our decision, no catastrophe will result. The worst that can happen is that the students will get an unnecessary review of material or they will be deprived of a review that may be beneficial to them. On the other hand, if we are going to use an assessment to decide whether to award a high school diploma or a college scholarship, we
should demand the most reliable measurement available. Such decisions have important consequences for the lives of the individuals involved.

It is not only a decision’s importance that matters but also whether it is possible to confirm or reverse the judgment at a later time. Decision making in education is seldom a single, final act. It tends to be sequential, starting with rather crude judgments and proceeding through a series of more refined judgments. In the early stages of decision making, low reliability might be quite tolerable because assessment results are used primarily as a guide to further information gathering. For example, on the basis of classroom assessments of questionable reliability, we might decide that some of our students are having such serious learning difficulties that they need special help. This decision can be confirmed or refuted by further assessment with more dependable measures. Opportunities for confirmation and reversal of judgments without serious consequences are almost always present in the early stages of educational decision making. Thus, the important thing when reliability is slow or unknown is not to treat the scores as if they were highly accurate. Make tentative judgments, seek confirming data, and be willing to reverse decisions when wrong.

Thus, when we ask how high reliability should be, several considerations must be taken into account. How important is the decision? Is it that can be confirmed or reversed at a later time? How far-reaching are the consequences of the action taken? For irreversible decisions that are apt to have great influence on the lives of individual students, we should make stringent demands on the reliability of the assessments we use. For lesser decisions, especially for those that can be later confirmed or reversed without serious consequences, we should be willing to settle for less reliable measures. Teacher-made tests commonly have reliabilities between .60 and .85, but they are useful for the types of instructional decisions typically made by teachers. Thus, the degree of reliability required depends largely on how confident we need to be about the decision being made. Greater confidence requires higher reliability.
3.3 **USABILITY**

In selecting assessment procedures, practical considerations cannot be neglected. Assessments are usually administered and interpreted by teachers with only a minimum of training in measurement. The time available for assessment is almost always limited, because assessment is in constant competition with other important activities for time in the school schedule. Likewise, the cost of assessment, although a minor consideration, is as carefully scrutinized by budget-conscious administrators as are other expenditures of school funds. These and other factors pertinent to the usability of assessment procedures must be taken into account when selecting assessment procedures. Such practical considerations are especially important when selecting published tests.

3.3.1 **Ease of Administration**

If the assessments are to be administered by teachers or others with limited training, ease of administration is an especially important quality to seek. For this purpose, directions should be simple and clear, subtests should be relatively few, and the time needed for administration of the assessment should not be too great. Administering a test with complicated directions and a number of subtests lasting but a few minutes each is a taxing chore for even an experienced examiner. For a person with little training and experience, such a situation is fraught with possibilities for errors in giving directions, timing and other aspects of administration that are like to affect results. Such errors of administration can have, of course, an adverse effect on the validity and reliability of the results.

<table>
<thead>
<tr>
<th>GUIDELINES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reliability Demands and Nature of the Decision</strong></td>
</tr>
<tr>
<td><em>High reliability is demanded when the:</em></td>
</tr>
<tr>
<td>- Decision is important</td>
</tr>
<tr>
<td>- Decision is final</td>
</tr>
<tr>
<td>- Decision is irreversible</td>
</tr>
<tr>
<td>- Decision is unconfirmable</td>
</tr>
<tr>
<td>- Decision concerns individuals</td>
</tr>
<tr>
<td>- Decision has lasting consequences</td>
</tr>
<tr>
<td><em>Example:</em> select or reject college applicants</td>
</tr>
</tbody>
</table>
Low reliability is tolerable when the:
- Decision is of minor importance
- Decision making is in early stage
- Decision is reversible
- Decision is confirmable by other data
- Decision concerns groups
- Decision has temporary effects

Example: Whether to review a classroom lesson

3.3.2 Time Required for Administration

With time for assessment at a premium, we always favor the shorter assessment, other things being equal. But in this case, other things are seldom equal, because reliability is directly related to the length of an assessment. If we attempt to cut down too much on the time allotted to assessment, we may reduce drastically the reliability or our scores.

For example, tests designed to fit a normal class period usually produce total test scores of satisfactory reliability, but their part-scores, obtained from the subtests, tend to be unreliable. If we want reliable measures in the areas covered by the subtests, we need to increase our testing time in each area. On the other hand, if we want a general measure in some area, such as verbal aptitude, we can obtain reliable results in 30 or 40 minutes, and there is little advantage in extending the testing time. A safe procedure is to allot as much time as is necessary to obtain valid and reliable results and no more. Between 20 and 60 minutes of testing time for each individual score yielded by a published test is probably a fairly good guide.

3.3.3 Ease of Interpretation and Application

In the final analysis, the success or failure of an assessment program is determined by the use made of the assessment results. If they are interpreted correctly and applied effectively, they will contribute to more intelligent educational decisions. On the other hand, if the assessment results are misinterpreted, misapplied, or not applied at all, they will be of little value and may actually be harmful to some individual or group.

Information concerning the interpretation and use of results of published test results is usually obtained directly from the test manual or related guides. Attention should be directed toward the clarity of score reports, the quality and relevance of norms and the comprehensiveness of the suggestions for applying the results to educational problems.
When the test results are to presented to students or parents, ease of interpretation and application are especially important.

3.3.4 Availability of Equivalent or Comparable Forms
For many educational purposes, equivalent forms of the same test are often desirable. Equivalent forms of a test measure the same aspect of behavior by using test items that are alike in content, level of difficulty, and other characteristics. Thus, one form of the test can substitute for the other, making it possible to test students twice in rather close succession without their answers on the first testing influencing their performance on the second testing. The advantage of equivalent forms is readily seen in mastery testing, when we want to eliminate the factor of memory while retesting students on the same domain of achievement. Equivalent forms of a test also may be used to verify a questionable test score. For example, a teacher may feel that a scholastic aptitude or achievement test score is spuriously low for a given student and may easily check this by administering an equivalent form of the test.
4.0 Testing and Assessment during Instruction

Tests and assessments given during instruction provide the basis for formative assessment. They are used to monitor learning progress, detect learning errors, and provide feedback to students and teacher. Teachers commonly call these formative tests learning tests, practice tests, quizzes, unit tests and the like. These tests and assessments typically cover some predefined segment of instruction (e.g. a chapter or particular set of skills) and thus encompass a rather limited sample of learning outcomes. The mix of types of tests items and more complex performance assessment tasks needs to be selected with care to ensure that the full range of critical instructional objectives is assessed. Ideally, the tests and assessments will be constructed in such a way that corrective prescriptions can be given for learning objectives that are yet to be achieved. When all students fail a set of items or perform poorly on an extended performance task, a group review may be applicable. When a small number of students have errors, alternative methods of study may be prescribed (e.g. reading assignments, practice exercises, and the like).

Persistent learning difficulties may require the use of diagnostic tests. For this type of testing, a number of test items are needed in each specific area, with some slight variation from item to item. In diagnosing students' difficulties in adding whole number, for example, it would be necessary to include addition problems containing various number combinations, some requiring carrying and some not, to pinpoint the specific types of errors being made. Diagnostic testing is a highly specialized area that has been somewhat neglected in educational measurement. There are some published diagnostic tests but these are primarily in the basic skills area. In other areas, teachers have to depend more heavily on the diagnostic features of formative tests or prepare their own diagnostic tests and assessments.
4.1 End-of-Instruction Testing and Assessment

At the end of a segment of instruction (e.g. unit or course), our main interest is in measuring the extent to which the intended learning outcomes have been achieved. Although these end-of-instruction tests and assessments are used primarily for summative assessment (e.g. to certify accomplishment or assign grades), they can also serve other functions. End-of-unit tests can be used for feedback to students, assigning of remedial work, and assessing instruction as well as for grading purposes. In fact, they can serve the functions of both formative and summative assessment, and in some cases serve as a pretest of the following unit (e.g. where the units are sequenced, as in math). End-of-course tests provide a broad survey of student learning over all of the intended outcomes of a course. In addition to their use in grading, they can also provide information for evaluating instructional effectiveness. A summary of the various types of classroom tests is presented in Table 4.1.
<table>
<thead>
<tr>
<th>Timing Function</th>
<th>Before instruction</th>
<th>During instruction</th>
<th>End of Instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Focus of measurement</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Timing</td>
<td>Readiness</td>
<td>Placement</td>
<td>Formative</td>
</tr>
<tr>
<td></td>
<td>Prerequisite entry skills</td>
<td>Course or unit objectives</td>
<td>Predefined segment of instruction</td>
</tr>
<tr>
<td><strong>Nature of sample</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Timing</td>
<td>Limited sample of selected skills</td>
<td>Broad sample of all objectives</td>
<td>Limited sample of learning tasks</td>
</tr>
<tr>
<td><strong>Item difficulty</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Timing</td>
<td>Typically has low level of difficulty</td>
<td>Typically has wide range of difficulty</td>
<td>Varies with the segment of instruction</td>
</tr>
<tr>
<td><strong>Time of administration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Timing</td>
<td>Beginning of course or unit</td>
<td>Beginning of course or unit</td>
<td>Periodically during instruction</td>
</tr>
<tr>
<td><strong>Use of results</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Timing</td>
<td>Remedy entry deficiencies or assignment to learning group</td>
<td>Instructional planning and advanced placement</td>
<td>Improve and direct learning through ongoing feedback</td>
</tr>
</tbody>
</table>
4.2 SELECTING APPROPRIATE TYPES OF ITEMS AND ASSESSMENT TASKS

It is common to make a distinction between classroom tests that consist of objective test items and performance assessments that require students to construct responses (e.g. write an essay) or perform a particular task (e.g. measure air pressure). Objective test items are highly structured and require the students to supply a word or two or to select the correct answer from a number of alternatives. They are called objective because they have single right or best answer that can be determined in advance. Performance assessment tasks, such as essay questions, permit the student to select, organize, and present the answer in essay form. Other types of performance assessment tasks may require the student to use equipment, generate hypotheses, make observations, construct something (e.g. a model), or perform for an audience (e.g. give a speech). For most performance assessment tasks there is not a single right or best response – there may be a variety of responses that are considered excellent. Expert judgment is required to score the performance. There is no conflict between these highly constrained objective items and the much less constrained performance assessment tasks. For some instructional purposes, objective items may be most efficient, whereas for others performance assessments may prove most satisfactory. Each approach should be used where most appropriate, with appropriateness determined by the learning outcomes to be measured and by the advantages and limitations of each approach.

4.2.1 The Objective Test Item

The great variety of different types of objective test items, can be classified into those that require the student to supply the answer and those that require the student to select the answer from a given number of alternatives. These two general classes are commonly further divided into basic types of objective test items, illustrated in the accompanying examples.
EXAMPLES  Supply Types

Short answer

What is the name of the author of Moby dick?  (Herman Melville)

What is the formula for hydrochloric acid?  (HCl)

What is the value of X in the equation 2X + 5 + 9?  (2)

Completion

Lines on a weather map joining points with the same barometric pressure are called (isobars).

The formula for ordinary table salt is (NaCl).

In the equation 2X + 5 = 9; X = (2)

Selection Types

True-false or alternative response

T  F  A virus is the smallest known organism
T  F  An atom is the smallest particle of matter
Yes  No  In the equation 2X + 5 + 9, X equals 3
Yes  No  Acid turns litmus paper red

Matching

(C)  1. And  A  Adjective
(D)  2. Dog  B  Adverb
(G)  3. Jump  C  Conjunction
(F)  4. She  D  Noun
(B)  5. Quickly  E  Preposition
          F  Pronoun
          G  Verb
Multiple-choice

In the equation $2X + 5 = 9$, $2X$ means

A 2 plus $X$.
B 2 minus $X$.
C 2 divided by $X$.
D 2 multiplied by $X$.

In which of the following sentences do the subject and verb disagree?

A When they win, they are happy
B Politics are hard to understand
C The majority is always right
D One or the other is to be elected.

In addition to these basic types of objective tests items, there are numerous modifications and combinations of types. However, there is little to be gained from listing all the possible variations, as many are unique to particular objectives or subject-matter areas. Some of the common variations used to measure understanding, thinking skills, and other complex learning outcomes will be illustrated later. An understanding of the principles of test construction, and the principles that apply to each of the specific types of objective tests items, should enable teachers to make adaptations that best fit their particular purposes.

The various types of objective test items have one feature in common that distinguishes them from performance assessment tasks: they present students with a highly structured task that limits the type of response they can make. To obtain the correct answer, students must demonstrate the specific knowledge, understanding, or skill called for in the item; they are not free to redefine the problem or to organize and present the answer in their own words. They must select one of several alternative answers or supply the correct word, number of symbol. This structuring of the problem and restriction on the method of responding contribute to objective scoring that is quick, easy and accurate. On the negative side, this same structuring makes the objective test
item inappropriate for measuring the ability to select, organize and integrate ideas. To measure such outcomes we must depend on performance assessment tasks.

Since written essays are the most commonly used form of performance assessment tasks it is worth focusing on essay questions before considering other types of performance assessment tasks. The essay question is commonly viewed as one item type. A useful classification, however, is one based on the amount of freedom of response allowed the student. This includes extended-response essay questions, in which students are given almost complete freedom in making their responses, and restricted-response essay questions, in which the nature, length, or organization of the response is limited. These types of essay questions are illustrated as follows:

**EXAMPLES**  
*Extended –response essay questions:*

Describe what you think the role of the federal government should be in maintaining a stable economy in Malaysia. Include specific policies and programs, and give reasons for your proposals.

*Restricted-response essay questions:*

State two advantages and two disadvantages of maintaining high tariffs on goods from other countries.

Extended-response essay questions permit students to decide which facts they think are most pertinent, to select their own method of organization, and to write as much as seems necessary for a comprehensive answer. Thus, such questions tend to reveal the ability to evaluate ideas, to relate them coherently, and to express them succinctly. To a lesser extent, they also reflect individual differences in attitudes, values, and creative ability.

Despite the apparent virtues of the extended-response essay question, they have two weaknesses: (1) they are inefficient for measuring knowledge of factual material because the questions are so extensive that only a small sample of content can be included in any one assessment, and (2) scoring is difficult and apt to be unreliable because the responses include an array of factual information of varying degrees of correctness, organized with varying degrees of coherence, and expressed with varying degrees of legibility and conciseness.
Restricted-response essay questions minimize some of the weaknesses of extended-response essay questions. Restricting the type of response makes it easier to measure knowledge of factual material and reduces somewhat the difficulty of the scoring. On the other hand, the more highly structured task presented by the restricted-response essay questions makes them less effective as a measure of the ability to select, organize and integrate ideas.

Neither extended-response essay questions nor restricted-response essay questions can serve all purposes equally. The type of essay question to use in a particular situation depends primarily on the learning outcomes to be measures and, to a lesser extent, on such practical considerations as the difficulty of scoring.

### 4.2.2 Other Types of Performance Assessment

There are many other types of performance assessment tasks. Examples include oral presentations; construction of graphs, diagrams or models, use of equipment or scientific instruments; typing; and playing a musical instrument. Tasks are selected so that the performance corresponds as closely as possible to an important instructional objective.

A distinction can be made between restricted-response performance tasks and extended-response performance tasks. These types are illustrated below.

**EXAMPLES**

**Extended-response performance task**

Prepare a weather report for another state and make an oral presentation of the report to the class using appropriate visual displays.

**Restricted-response performance task**

Measure and record the relative humidity.

The virtues and limitations of extended-and-restricted-response performance tasks are much the same as those listed for extended-and restricted-response essay questions. The freedom provided by the extended response performance task enables students to display such important skills as problem solving, planning organization, integration, and creativity. Questions if this type also provides an opportunity to observe student performance in more realistic contexts than are possible with objectives test items. On the other hand, they are time-consuming to administer and difficult to score.
Restricted-response performance tasks are generally easier to score and require less than extended-response performance tasks. However, they are generally less suited for measuring the higher-order skills measured by extended-response performance tasks.

**Comparative Advantages of Objective Test Items and Performance Assessment Tasks**

Both objective test items and performance assessment tasks can provide valuable evidence concerning student achievement. Each has advantages and limitations which make it more appropriate for some purposes than for others. A comparison of the relative merits of objective item tests and performance assessments is presented in Table 4.2.

In considering the comparative advantages of objective-item tests and performance assessments, we must be careful not to fall into either/or thinking that is, to use either objectives items or performance assessment tasks. It is frequently better to use both, with each measuring the particular learning outcomes for which it is best suited. This should also have a desirable influence on student learning, because in preparing for such tests, students must attend to the specific types of learning outcomes measured by objective items as well as the synthesis and integrated performance outcomes measured by performance assessment tasks.

**TABLE 4.2** Comparative advantages of objective tests and performance assessments

<table>
<thead>
<tr>
<th></th>
<th>Objective Test</th>
<th>Performance Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Learning outcomes</strong></td>
<td>Is efficient for measuring knowledge of facts. Some types (e.g. multiple-choice) can also measure understanding, thinking skills, and other complex outcomes. Inefficient or inappropriate for measuring ability to select and organize ideas, writing abilities, and some types of problem-solving skills.</td>
<td>Is inefficient for measuring knowledge of facts. Can measure understanding, thinking skills, and other complex learning outcomes (especially useful where originally of response is desired). Appropriate for measuring performance on tasks corresponding to important instructional objectives in realistic contexts.</td>
</tr>
<tr>
<td><strong>Preparation of questions</strong></td>
<td>A relatively large number of questions are needed for a test. Preparation is difficult and time-consuming.</td>
<td>Only a few weeks are needed for an assessment.</td>
</tr>
<tr>
<td><strong>Sampling of course</strong></td>
<td>Provides an extensive sampling of</td>
<td>Sampling of course content is</td>
</tr>
<tr>
<td>control content</td>
<td>course content because of the large number of questions that can be included in a test</td>
<td>usually limited because of the small number of tasks that can be included in an assessment.</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Control of student response</td>
<td>Complete structuring of task limits student to type of response called for. Prevents bluffing and avoids influence of writing skills, though selection-type items are subject to guessing.</td>
<td>Freedom to respond in own way enables students to display originality and guessing is minimized.</td>
</tr>
<tr>
<td>Scoring</td>
<td>Objective scoring</td>
<td>Judgmental scoring</td>
</tr>
<tr>
<td>Influence on learning</td>
<td>Usually encourages student to develop a comprehensive knowledge of specific facts and the ability to make fine discriminations among them. Can encourage the development of understanding, thinking skills, and other complex outcomes if properly constructed.</td>
<td>Encourages students to concentrate on larger units of subject matter, with special emphasis on the ability to organize, integrate and express ideas effectively. Essays may encourage poor writing habits if time pressure is a factor (it almost always is).</td>
</tr>
<tr>
<td>Reliability</td>
<td>High reliability is possible and is typically obtained with well-constructed tests.</td>
<td>Reliability is typically low, primarily because of inconsistent scoring and limited sample of tasks.</td>
</tr>
</tbody>
</table>

### 4.2.3 Selecting the Most Appropriate Types of Items and Tasks

A basic principle in selecting the type of test item and assessment task to use is: Select the item type that provides the most direct measure of the intended learning outcome. If, for example, the intended learning outcome is writing, naming, listing or speaking, the task should require the students to supply the answer. If the outcome involves the use of laboratory equipment to solve a problem, nothing short of an actual laboratory performance task will suffice. On the other hand, if the task calls for identifying a correct answer, a selection-type item should be used. In those cases where the specific learning outcome does not make clear which item type to use, selection-type items would be favored because of the greater control over the student’s response and the objectivity of the scoring. The following objective-item test was prepared for the weather.
unit discussed earlier. Note how each type of objective item provides a direct measure of the outcome it was designed to measure.

EXAMPLE Short-answer items

**Specific Learning Outcome**: Writes a definition of each term.

**Directions**: Write a one-sentence definition of each of the following terms.

1. Weather
2. Humidity
3. Occluded front

**Specific Learning Outcome**: Names the instrument used for measuring each weather element.

1. The instrument used to measure the amount of precipitation in a given locality is called a (an) ___________

**Specific Learning Outcome**: Lists the characteristics of a given weather phenomenon.

1. List three main characteristics of a hurricane

**Specific Learning Outcome**: Measures relative humidity

1. Use the appropriate instrument to measure the relative humidity and record the value obtained.
Multiple-Choice Items

Specific Learning Outcome: Identifies the units of measurement used in reporting each weather element on a weather map.

1. United States weather maps indicate air pressure in
   a. inches
   b. feet
   c. pounds
   d. millibars

True-False Items

Specific Learning Outcome: Distinguishes between correct and incorrect procedures for determining each weather element.

T  F  1. Dew point is determined by cooling a sample of air until it is free of moisture.

T  F  2. Ceiling if determined by using balloons that rise at know rates.

Matching Items

Specific Learning Outcome: Matches each weather instrument to the weather element it measures.

Directions: On the line to the left of each weather element in Column A, write the letter of the weather instrument in Column B that is used for measuring it. Each instrument in Column B may be used once, more than once, or not at all.

<table>
<thead>
<tr>
<th>Column A</th>
<th>Column B</th>
</tr>
</thead>
<tbody>
<tr>
<td>(B) 1. Air pressure</td>
<td>A. Anemometer</td>
</tr>
<tr>
<td>(E) 2. Air temperature</td>
<td>B. Barometer</td>
</tr>
<tr>
<td>(C) 3. Humidity</td>
<td>C. Hygrometer</td>
</tr>
<tr>
<td>(D) 4. Wind Velocity</td>
<td>D. Rain gauge</td>
</tr>
<tr>
<td></td>
<td>E. Thermometer</td>
</tr>
<tr>
<td></td>
<td>F. Wind vane</td>
</tr>
</tbody>
</table>
4.3 General Suggestions for Writing Test Items and Assessment Tasks

In preparing a set of test items or assessment tasks, there are some general rules that apply to all types of items or tasks. These will be listed here. The specific rules for writing each item or ask type will be described and illustrated in the following chapters.

1. **Use your test and assessment specifications as a guide.** The specifications describe the performance to be measured and the sample of learning outcomes to measure. Thus, they serve as an aid for selecting the types of items and tasks to prepare, for writing, and for determining how many items and tasks are needed for each subdomain of achievement.

2. **Write more items and tasks than needed.** Preparing more test items and assessment tasks than needed will permit the weaker items and tasks to be discarded during later review. It will also make it easier to match the final set of items and tasks to the specifications.

3. **Write the items and tasks well in advance of the testing date.** Setting the items and tasks aside for several days and then reviewing them with a fresh outlook will reveal any lack of clarity or ambiguity that was overlooked during their preparation. It is frequently surprising how many defects slipped through during the original writing.

4. **Write each test items and assessment task so that it calls forth the performance described in the intended learning outcome.** Both during writing and review, compare the task students are asked to perform to the learning outcome the task is designed to measure to make sure the two matches.

5. **Write each test item and assessment task so that the task to be performed is clearly defined.** Clarity is obtained by carefully formulating the question and instructions, using simple and direct language, using correct punctuation and grammar, and avoiding unnecessary wording.

6. **Write each item or task at an appropriate reading level.** Keep the reading difficulty and vocabulary level as simple as possible to prevent these factors from distorting the results. Students’ responses should be determined by the performance being measured, not by some factor the item or task was not designed to measure.
7. **Write each item or task so that it does not provide help in responding to other items or tasks.** Unless care is taken during writing, one item may provide information that is useful in answering another item. For example, a name, date, or fact inadvertently included in the stem of a multiple-choice item may be called for in a short-answer item in another part of the best.

8. **Write each item so that the answer is one that would be agreed upon by experts, or, in the case of assessment tasks, the responses judged excellent would be agreed upon by experts.** This rule is easy to satisfy when measuring factual knowledge but more difficult when measuring complex outcomes calling for the extended essays or other types of performance.

9. **Whenever a test item or assessment task is revised, recheck its relevance.** When reviewing items or tasks for appropriateness, clarity, difficulty and freedom from clues and bias, some revision is often needed. After revising an item or task, check to be sure that it still provides a relevant measure of the intended learning outcome. Even slight changes can sometimes modify the function of an item or a task.

4.4 **Focusing on Improving Learning and Instruction**

The ultimate purpose of testing and assessment is to improve student learning. As you construct classroom tests and assessments, keep in mind the extent to which it is likely to contribute, directly or indirectly, toward this end. Well-constructed classroom tests and assessments should increase both the quantity and quality of student learning.

1. **Tests and assessments can have a desirable influence on student learning if attention is paid to the breadth and depth of content and learning outcomes measured.** When we select a representative sample of content from all of the areas covered in our instruction, we are emphasizing to the students that they must devote attention to all areas. They cannot neglect some aspects of the course and do well on the tests. Similarly, when our tests measure a variety of types of learning outcomes, the students soon learn that a mass of memorized factual information is not sufficient. They must also learn to interpret and apply facts, develop deep conceptual understandings, draw conclusions, recognize assumptions, identify cause-and-effect relations, generate hypotheses, solve meaningful problems, and the like. This discourages the students from depending solely on
memorization as a basis for learning, and encourages them to develop the use of more complex mental processes.

2. **Constructing tests and assessments that measure a variety of learning outcomes should also lead to improved teaching procedures and, thus, indirectly to improved student learning.** As we translate the various learning outcomes into test items and assessment tasks, we clarify our notion of understandings, thinking skills, and other complex learning outcomes. This clarification enables us to plan the learning experiences of students more effectively and increases the degree to which we emphasize the understandings, thinking skills, and other complex learning outcomes in our teaching. A well-constructed test or assessment frequently leads to a review of teaching procedures and to the abandonment of those that encourage rote learning.

3. **Tests and assessments will contribute to improved teacher-student relations (with a beneficial effect on student learning) if students view the test and assessments as fair and useful measures of their achievement.** We can make fairness apparent by including a representative sample of the learning outcomes that have been emphasized during instruction, by writing clear directions, by making certain that the intent of each item or task is clear and that each item is free of any type of bias that would prevent a knowledgeable person from performing well, and by providing adequate time limits. Student recognition of usefulness, however, depends as much on what we do with the results of the test or assessment as on the characteristics of the instrument itself. We make the usefulness apparent by using the results as a basis for guiding and improving learning.
5.0 ASSEMBLING THE CLASSROOM TEST

The preparations of test items for use in a test is greatly facilitated if the items are properly recorded, if they are written at least several days before they are to be used, and if extra items are constructed.

5.1 Recording Test Items

When constructing the test items it is desirable to write each one on a separate index card, or, if a computer is used, in a form that makes it possible to easily address and retrieve individual items. In addition to the test item, the card should contain information concerning the instructional objectives, the specific learning outcome, and the content measured by the item. A space should also be reserved for item-analysis information, usually on the back of the card, to allow room to record the data each time the item is used.

Item cards provide flexibility. As the items are reviewed and edited, they can be eliminated, added, or revised with very little difficulty. The same holds true when arranging the items for the test: They can be arranged and rearranged merely by sorting the cards. The flexibility of this recording system also makes it easy to add the items to a computer item bank.

5.2 Reviewing Test Items and Assessment Tasks

No matter how carefully items or tasks have been prepared, defects inadvertently creep in during construction. As we concentrate on the clarity and conciseness of a question, a verbal clue slips in unnoticed. As we attempt to construct more challenging items we unwittingly introduce some ambiguity. As we rework a multiple-choice item to make the incorrect choices more plausible, the behavior called forth by the item is unintentionally modified. As we attempt to increase the authenticity of a task for problems faced outside the classroom, we introduce unintended reliance on access to resources that put students from families with limited
resources at an unfair disadvantage. In short, we focus so closely on some aspects of item or task construction that we overlook others. This results in an accumulation of unwanted errors that may distort the function of the item or task. “Such technical defects can be most easily detected by 1) reviewing the items and tasks after they have been set aside for a few days and 2) asking a fellow teacher to review and criticize them.

In reviewing test items and tasks, we should try to view them from the student’s viewpoint, as well as from that of the teacher. From these two vantage points, each item or task should be read carefully and its effectiveness judged. The following questions will help you analyze the quality of each item or task.

**i. Is the format appropriate for the learning outcome being measured?** If the learning calls for the definition of a term, for example, a supply-type item (e.g. short-answer item) would be appropriate, and a selection type item (e.g. multiple choice) would be clearly inappropriate. If the learning outcome calls for the ability to collect, organize, integrate and present information in the form of a coherent argument, then nothing short of a performance-based task will suffice. On the other hand, if the intended outcome was simply the identification of the correct definition, then a selection-type item would be adequate. Thus, the first step is to check whether the format is suitable for the type of student performance described in the testing and assessment plan. The action verb in the statement of each specific learning outcome (e.g. defines, describes, identifies) indicates which item format is more appropriate.

**ii. Does the knowledge, understanding, or thinking skill called forth by the item or task match the specific learning outcome and subject-matter content being measured?** When a table of specifications has been used as a basis for constructing items and tasks, this is merely a matter of checking to see whether the item or task is still relevant to the same cell in the table. If, for example, an item’s functioning content has shifted during construction, the item either should be modified so that it serves its original purpose or reclassified in light of the new purpose. In any case, the response called forth by an item or task should agree with the purpose for which it is to be used.

**iii. Is the point of the item or task clear?** A careful review of items and tasks often reveals ambiguity, inappropriate vocabulary, and awkward sentence structure that
were overlooked during their construction. Returning to items and tasks after they have been set aside for a few days provides a fresh outlook that makes such defects more apparent. The difficulty of the vocabulary and the complexity of the sentence structure must, of course, be judged in terms of the students' maturity level. At all levels, however, ambiguity should be removed. In its final form, each item or task should be so clearly worded that all students understand what is called for. The quality of student responses should be determined solely by whether they possess the knowledge, understanding or skill being measured.

iv. **Is the item or task free from excessive verbiage?** Often, items become excessively wordy because of awkward sentence structure or the inclusion of nonfunctional material. Some teachers justify the use of an item by including a statement or two concerning the problem’s importance. Others expand a simple question into an elaborate story situation to make the item more interesting. Although adding such nonfunctional material may be useful in some instances, items and tasks are generally more effective when the problem is stated as concisely as possible. When reviewing items the content of each item should be analyzed to determine the functional elements leading to the correct response. If there are any elements that the students may disregard entirely and still respond correctly, they probably should be removed. See “Reviewing and Revising Test Items”.

v. **Does the item have an answer that would be agreed upon by experts? How well would experts agree about the degree of excellence of task performances?** This is seldom a problem with factual material, which usually can be judged as correct or incorrect. It is more of a problem with selection-type items that ask for the best reason, the best method, or the best interpretation. The problem is greatest with tasks requiring extended performances where qualified judges may differ in their evaluation of performances. If experts agree on the best response, fine, but do not include items that require students to endorse someone’s unsupported opinion (even if it happens to be yours) and don’t evaluate performances on tasks simply in terms of your own personal preferences.

vi. **Is the item or task free from technical errors and irrelevant clues?** As noted earlier, an irrelevant clue for a selection-type item is any element that leads the poor achiever to the correct answer and thereby prevents the item from functioning as
intended. These include 1) grammatical inconsistencies, 2) verbal associations, 3) specific determiners (e.g. words such as always and never), and 4) some mechanical features, such as correct statements tending to be longer than incorrect one. Most of these clues can be removed merely by trying to detect them during the item review. They somehow seem more obvious after the items have been set aside for a while.

vii. **Is the item or task free from racial, ethnic, and sexual bias?** A final check should be made to make certain that the vocabulary and problem situation in each item or task would be acceptable to the members of all groups and would have a similar meaning to them. An effort should be made to remove any type of stereotyping, such as always portraying minorities in subservient roles, women in homemaking roles, and the like. A judicious and balanced use of different roles for minorities and males and females should contribute to more effective assessment.

When items or tasks have been revised and those to be included in the test or assessment have been tentatively selected, the following questions should be asked:

1. Does the set of items and tasks measure a representative sample of the learning outcomes and course content included in the assessment plan?

2. Are there enough items or tasks for each interpretation to be made?

3. Is the difficulty of the items and tasks appropriate for the measurement purpose and for the students for whom the test or assessment is intended?

4. Are the test items free from overlapping so that the information in one does not provide a clue to the answer in another?

The first question can be answered by comparing the final selection of items and tasks with the table of specifications or other assessment plan. Answer to the last three are determined by reviewing the items and tasks in each content area and as a total set. Affirmative answers to these questions mean the items and tasks are ready to be assembled for administration.
5.3 Arranging Items in the Test

There are various methods of grouping items in an achievement test, and the method will vary somewhat with the use to be made of the results. For most classroom purposes, the items can be arranged by a systematic consideration of 1) the types of items used 2) the learning outcomes measured, 3) the difficulty of the items, and 4) the subject matter measured.

First and foremost, the items should be arranged in sections by item type. That is, all true-false items should be grouped together, then all matching items, then all multiple-choice items and so on. This arrangement requires the fewest sets of directions; it is the easiest for the students because they can retain the same mental set throughout each section; and it greatly facilitates scoring. When two or more item types are included in a test, there is also some advantage in keeping the simpler item types together and placing the more complex ones in the test, as follows:

i. True-false or alternative-response items.

ii. Matching items

iii. Short-answer items

iv. Multiple-choice items

v. Interpretive exercises

vi. Restricted-response essay questions.

It is not expected that all item types will appear in the same test. Seldom are more than a few types used, but this is the general order.

5.4 Preparing Directions for the Test or Assessment

Teachers sometimes devote considerable time and attention to the construction and assembly of test items or a challenging performance-based assessment and then dash off directions with very little thought. In fact, many teachers include no written directions with their tests, assuming either that the items are self-explanatory or that the students are conditioned to answering the types of items used in the test. Some teachers also use oral directions, but they frequently leave much to be desired. Whether written, oral or both, the directions should include at least the following points:
i. Purpose of the test or assessment

ii. Time allowed for completing the test or performing the task

iii. Basis for responding

iv. Procedure for recording the responses

v. What to do about guessing for selection type-test items.

vi. What to do scoring open-ended or extended responses

The amount of detail for each of these points depends mainly on the students’ age level, the comprehensiveness of the test or assessment, the complexity of the items or tasks, and the students’ experience with the testing or assessment procedure used. Using new item types and separate answer sheets, for example, requires much more detailed directions than do familiar items requiring students merely to circle or underline the answer.

i. **Purpose of the Test or Assessment.** The purpose of the test or assessment is usually indicated when the test is announced or at the beginning of the semester when the evaluation procedures are described as a part of the general orientation to the course. Should there be any doubt whether the purpose is clear to all students, however, it could be explained again at the time of testing or assessment. This is usually done orally. The only time a statement of the purpose of the test or assessment needs to be included in the written directions is when it is to be administered to several sections taught by different teachers. Then a written statement of purpose ensures greater uniformity.

ii. **Time Allowed for Completing the Test or Performing the Task.** It is helpful to tell the students how much time they will have for the whole test or performance task and how to distribute their time among the parts. When essay questions are included, it is also good to indicate approximately how much time should be allotted to each question. This enables the students to use their time most effectively and prevents less able students from spending too much time on questions that are particularly difficult for them.
Classroom tests or assessments of achievement should generally have liberal time allowances. Except for special computational skills, speed is not important. Our main concern is the level of achievement each student has attained. Were it not for practical considerations like the length of class periods and the pressure of other school activities, there would be no need for any time limits with most classroom achievement tests or assessment tasks.

**iii. Basis for Responding.** The directions for each section of the test should indicate the basis for selecting or supplying the answers. With true-false, matching, and multiple-choice items, this part of the directions can be relatively simple. For example, a statement like ‘select the choice that best completes the statement of answers the question’ might be sufficient for multiple-choice items. When interpretive exercises are used, however, more detailed directions are necessary because the basis for the response is much more complex. The directions must clearly indicate the type of interpretation expected.

**iv. Procedure for Recording Responses.** Answers may be recorded on the test form itself or on separate answer sheets. If the test is short, the number of students taking the test is small, or the students are relatively young, answers are generally recorded directly on the test paper. For most other situations, separate answers sheets are preferred because they reduce the time needed for scoring, and they make it possible to use the test papers over again. The latter feature is especially useful when the test is to be given to students in different sections of the same course.

Directions for recording the answer on the test paper itself can be relatively simple. With selection items, it is merely a matter of instructing the students to circle, underline, or check the letter indicating the correct answer. For students in the primary grades, it is usually better to ask them to mark the answer directly by drawing a line under it. With supply items, the directions should indicate where to put the answer and the units in which it is to be expressed if the answer is numerical.
v. **What to Do about Guessing for Selection-Type Items.** When selection-type items are used, the directions should tell students what to do when they are uncertain of the answer. Should they guess or omit the item? If no instructions are given on this point, the bold students will guess freely, whereas others will answer only those items of which they are fairly certain. The bold students will select some correct answers just by lucky guesses, and thus their scores will be higher than they should be. On the other hand, if the students are instructed “do not guess” or “answer only those items of which you are certain”, the more timid students will omit many items they could answer correctly. Such students are not very certain about anything, which prevents them from responding even when they are reasonably sure of the answers. With these directions, the bold students will continue to guess, although possibly not quite so wildly.

vi. **Reproducing the Test**

In preparing the test materials for reproduction, it is important that the items be spaced and arranged so that they can be read, answered, and scored with the least amount of difficulty. Cramming too many test items onto a page is poor economy. What little paper is saved will not make up for the time and confusion that result during the administration and scoring of the test.

<table>
<thead>
<tr>
<th>GUIDELINES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Helping Students Prepare for Tests and Assessments</strong></td>
</tr>
<tr>
<td><strong>General Preparation</strong></td>
</tr>
<tr>
<td>1. Suggest ways of studying</td>
</tr>
<tr>
<td>2. Give practice tasks like those to be used</td>
</tr>
<tr>
<td>3. Teach test-taking skills</td>
</tr>
<tr>
<td>4. Teach how to write well-organized essay answers</td>
</tr>
<tr>
<td>5. Stress the value of tests and assessments for improving learning</td>
</tr>
<tr>
<td><strong>Preparation for Each Test or Assessment</strong></td>
</tr>
<tr>
<td>1. Announce in advance when the test or assessment will be given</td>
</tr>
<tr>
<td>2. Describe the conditions of administration (e.g. one hour, closed book)</td>
</tr>
<tr>
<td>3. Describe the length and the types of items or tasks to be used (e.g. twenty multiple-choice, three essay items, one extended-response performance task).</td>
</tr>
</tbody>
</table>
4. Describe the content and type of performance to be covered (a table of specifications is useful for this).
5. Describe how the test or assessment will be scored and how the results will be used.
6. Give the students sample items and tasks similar to those to be used (use a short practice test or present items orally and discuss responses).
7. Relieve anxiety by using a positive approach in describing the test or assessment and its usefulness.

5.5 ADMINISTERING AND SCORING CLASSROOM TESTS AND ASSESSMENTS

The same care that went into the preparation of the test or assessment should be carried over into its administration and scoring. Here we are concerned with 1) providing optimum conditions for obtaining the students’ responses and 2) selecting convenient and accurate procedures for scoring the results.

5.5.1 Administration

The guiding principle in administering any classroom test or assessment is that all students must be given a fair chance to demonstrate their achievement of the learning outcomes being measured. This means a physical and psychological environment conducive to their best efforts and the control of factors that might interfere with valid measurement.

Physical conditions such as adequate work space, quiet, proper light and ventilation, and comfortable temperature are sufficiently familiar to teachers to warrant little attention here. Of greater importance, but frequently neglected, are the psychological conditions influencing results. Students will not perform at their best if they are tense and anxious during testing. Some of the things that create excessive test anxiety are:

1. Threatening students with tests if they do not behave.
2. Warning students to do their best “because this test is important”.
3. Telling students they must work fast in order to finish on time.
4. Threatening dire consequences if they fail.
The antidote to test anxiety is to convey to the students, by both work and deed, that the test and assessment results are to be used to help them improve their learning. They also should be reassured that the time limits are adequate to allow them to complete the test or assessment tasks. This, of course, assumes that the test and assessment results will be used to improve learning and that the time limits are adequate.

The time of testing can also influence the results. If tests are administered just before the “big game” or the “big dance”, the results may not be representative. Furthermore, for some students, fatigue, the onset of illness, or worry about a particular problem may prevent maximum performance. Arranging the time of testing accordingly and permitting its postponement when appropriate can enhance the validity of the results.

Actual administration is relatively simple, because a properly prepared test or assessment is practically self-administering. Oral directions, if used, should be presented clearly. Any sample problems or illustrations put on the blackboard should be kept brief and simple. Beyond this, suggestions for administration consist mainly of things to avoid.

i. **Do not talk unnecessarily before letting students start working.** When a teacher announces that there will be “a full forty minutes” to complete the test and then talks for the first ten minutes, students feel that they are being unfairly deprived of testing time. Besides, just before a test is no time to make assignments, admonish the class, or introduce next week’s topic. Students are mentally set for the test and will ignore anything not pertaining to the test for fear it will hinder their recall of information needed to answer the questions. Thus, the well-intentioned remarks merely increase anxiety toward the test and create hostility toward the teacher.

ii. **Keep interruptions to a minimum.** At times, a student will ask to have an ambiguous item clarified, and it may be beneficial to explain the item to the entire group at the same time. Such interruptions are necessary but should be kept to a minimum. All other distractions outside and inside the classroom should, of course, also be eliminated when possible. It is sometimes helpful to hang a “Do not disturb-TESTING” sign outside the door.
iii. **Avoid giving hints to students who ask about individual items.** If the item is ambiguous, it should be clarified for the entire group, as indicated earlier. If it is not ambiguous, refrain from helping the student answer it. Refraining from giving hints to students who ask for help is especially difficult for beginning teachers. But giving unfair aid to some students (the bold, the apple polishers, and so on) decreases the validity of the results and lowers class morale.

iv. **Discourage cheating, if necessary.** When there is good teacher-student rapport and the students view tests as helpful rather than harmful, cheating is usually not a problem. Under other conditions, however, it might be necessary to discourage cheating by special seating arrangements and careful supervision. Receiving unauthorized help from other students during a test has the same deleterious effect on validity and class morale as does receiving special hints from the teacher. We are interested in students’ doing their best; but for valid results, their scores must be based on their own unaided efforts. See “Steps to Prevent Cheating”.

<table>
<thead>
<tr>
<th>GUIDELINES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Steps to Prevent Cheating</strong></td>
</tr>
<tr>
<td>1. Take special precautions to keep the test secure during preparation, storage, and administration.</td>
</tr>
<tr>
<td>2. Have students clear off the tops of their desks (for adequate work space and to prevent use of notes).</td>
</tr>
<tr>
<td>3. If scratch paper is used (e.g. for math problems), have it turned in with the test.</td>
</tr>
<tr>
<td>4. Proctor the testing session carefully (e.g. walk around the room periodically and observe how the students are doing).</td>
</tr>
<tr>
<td>5. Use special seating arrangements, if possible (e.g. leave an empty row of seats between students).</td>
</tr>
<tr>
<td>6. Use two forms of the test and give a different form to each row of students (for this purpose, use the same test but simply rearrange the order of the items for the second form).</td>
</tr>
<tr>
<td>7. Prepare tests that students will view as relevant, fair, and useful.</td>
</tr>
<tr>
<td>8. Create and maintain a positive attitude concerning the value of tests for improving learning.</td>
</tr>
</tbody>
</table>
5.5.2 Scoring the Test

If the students’ answers are recorded on the test paper itself, a scoring key can be made by marking the correct answers on a blank copy of the test. Scoring then is simply a matter of comparing the columns of answers on this master copy with the columns of answers one each student’s paper. A strip key, which consists merely of strips of paper on which the columns of answers are recorded, may also be used if more convenient. These can easily be prepared by cutting the columns of answers from the master copy of the test and mounting them on strips of cardboard cut from manila folders.

When separate answer sheets are used, a scoring stencil is most convenient. This is a blank answer sheet with holes punched where the correct answers should appear. The stencil is laid over each answer sheet, and the number of answer checks appearing through the holes are counted. When this type of scoring procedures is used, each test paper should also be scanned to make certain that only one answer was marked for each item. Any item containing more than one answer should be eliminated from the scoring.

As each test paper is scored, mark each item that is answered incorrectly. With multiple-choice items, a good practice is to draw a red line through the correct answer of the missed items rather than through the student’s wrong answers. This will indicate to the students those items missed and at the same time will indicate the correct answers. Time will be saved and confusion avoided during discussion of the test. Marking the correct answers of missed items is especially simple with a scoring stencil. When no answer check appears through a hole in the stencil, a red line is drawn across the hole.

5.6 Determining Item and Task Effectiveness

The effectiveness of each test item can be determined by analyzing student responses to it. Item analysis is generally associated with a norm-referenced perspective. This is natural, because the results of an item analysis can be used to select items of desired difficulty that best discriminate between high and low achieving students. Selection on these grounds is not relevant from a criterion-referenced perspective. From both perspectives, however, the results of an item analysis can be useful in identifying faulty items and can provide information about student misconceptions and topics that need additional work.
Item analysis is usually designed to answer questions such as the following:
1. Did the item function as intended?
2. Were the best items of appropriate difficulty?
3. Were the test items free of irrelevant clues and other defects?
4. Was each of the disasters effective (in multiple-choice items)?

Answers to all but Question 2 are relevant in constructing future tests based on either a norm-referenced or a criterion-referenced perspective. The answer to Question 2 is only relevant when planning future norm-referenced tests; however, it is relevant in instructional planning regardless of perspective.

Answers to such questions are of obvious value in selecting or revising items for future use. The benefits of item analysis are not limited to the improvement of individual test items, however. There are a number of fringe benefits of special value to classroom teachers. The most important of these are the following:

5.6.1 Computing Item Difficulty. The difficulty of a test item is indicated by the percentage of students who get the item right. Hence, we can compute item difficulty \( P \) by means of the following formula, in which \( R \) equals the number of students who got the item right, and \( T \) equals the total number of students who tried the item.

\[
P = \frac{R}{T} \times 100
\]

Applying this formula to the item-analysis, say 14 students, our index of item difficulty is 70 percent, as follows:

\[
P = \frac{14}{20} \times 100 = 70 \text{ percent}
\]

In computing item difficulty from item-analysis data, our calculation is based on the upper and lower groups only. We assume that the responses of students in the middle group follow essentially the same pattern. This estimate of difficulty is sufficiently accurate for classroom use and is easily obtained because the needed figures can be taken directly from the item-analysis data.

Note that because our item analysis is based on ten in the upper group and ten in the lower group, all we need to do to obtain item difficulty is to divide the number getting it right by two \((14/2 = 7)\), move the decimal point one place to the right \((70)\), and add the percent sign \((70\%)\). In other words, 14 out of 20 is the same as 7 out of 10, which is 70
percent. If 13 students were to get the item right, item difficulty would be 6.5 out of 10 (13/2 = 6.5), or 65 percent. This may seem a bit confusing at first, but once your grasp the idea, you can compute item difficulty very quickly. As noted earlier, the ease of interpreting item statistics is one of the advantages of using ten in each group. If more (or fewer) than ten are used, the formula for computing item difficulty is the same, but it is much more difficult to compute the results mentally.

5.6.2 Computing Item Discriminating Power. As we have already stated, an item discriminates positively if more students in the upper group than the lower group get the item right. Positive discrimination indicates that the item is discriminating in the same direction as the total test score. Because we assume that the total test score reflects achievement of desired objectives, we would like all of our test items to show positive discrimination.

The discriminating power of an achievement test item refers to the degree to which it discriminated between students with high and low achievement. Item discriminating power (D) can be obtained by subtracting the number of students in the lower group who get the item right \((R_L)\) from the number of students in the upper group who get the item right \((R_U)\) and dividing by one-half the total number of students included in the item analysis \((.5T)\). Summarized in formula form, it is:

\[
D = \frac{R_U - R_L}{.5T}
\]

Applying this formula to the item-analysis data, we obtain an index of discriminating power of 0.60, as follows:

\[
D = \frac{10 - 4}{10} = 0.60
\]

This indicates approximately average discriminating power. An item with maximum positive discriminating power is one in which all students in the upper group get the item right and all the students in the lower get the item wrong. This results in an index of 1.00, as follows:

\[
D = \frac{10 - 0}{10} = 1.00
\]
An item with no discriminating power is one in which an equal number of students in both the upper and lower groups get the item right. This results in an index of .00 as follows:

\[ D = \frac{10 - 10}{10} = .00 \]

When out item analysis is based on 10 in the upper group and 10 in the lower group, the index of discriminating power, like item difficulty, can be computed easily and quickly. All we need to do is subtract the number in the lower group getting it right from the number in the upper group getting it right \((10 - 4 = 6)\), move the decimal point one to the left \((.6)\), and add a zero after it \((.60)\). With ten in each group the index of discrimination is essentially the difference between the numbers getting it right in the two groups with the decimal point moved one to the left. The zero is added simply because the index of discrimination is usually carried to two decimal places. With more than ten in each group, we could not make these simple mental calculations but would have to resort to use of the formula.

5.6.3 Evaluating the Effectiveness of Distracters. How well each distracter is operating can be determined by inspection, and so there is no need to calculate an index of effectiveness, although the formula for discriminating power can be used for this purpose. In general, a good distracter attracts more students from the lower group than the upper group. Thus, it should discriminate between the upper and lower groups in a manner opposite to that of the correct alternative. An examination of the following item-analysis data will illustrate the ease with which the effectiveness of distracters can be determined by inspection. Alternative A is the correct answer.

<table>
<thead>
<tr>
<th>Alternatives</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>Omits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper 10</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Lower 10</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

First, note that the item discriminates positively, because five in the upper group and three in the lower group got the item right. The index of discriminating power is fairly low however \((D=.20)\), and this may be partly due to the ineffectiveness of some of the distracters. Alternative B is a poor distracter because it attracts more students from
upper group than from the lower group. This is most likely due to some ambiguity in the statement of the item. Alternative C is evidently not a plausible distracter because it attracted no one. Alternative D is functioning as intended, for it attracts a larger proportion of students from the lower group. Thus, the discriminating power of this item can probably be improved by removing any ambiguity in the statement of the item and revising or replacing Alternatives B and C. The specific changes must, of course, be based on an inspection of the test item itself; item-analysis data merely indicate poorly functioning items, not the cause of the poor functioning.

In some cases, an examination of the test item will reveal no obvious error in the structure of the item and it may be best to try it with a second group. The number of cases involved is so small that considerable variation in student response can be expected from one group to another. A casual comment by the teacher, or some other classroom event, may cause students to select or reject a particular alternative.

5.6.4 Cautions in Interpreting Item-Analysis Results

Item analysis is a quick, simple technique for appraising the effectiveness of individual test items. The information from such an analysis is limited in many ways, however, and must be interpreted accordingly. The following are some of the major cautions to observe.

i. Item discriminating power does not indicate item validity. In our description of item analysis, we use the total test score as a basis for selecting the upper group (high achievers) and the lower group (low achievers). This is the most common procedure because comparable measures of achievement are usually not available. Ideally, we would examine each test item in relation to some independent measure of achievement. However, the best measure of the particular achievement we are interested in assessing is usually the total score on the achievement test we have constructed because each classroom test is related to specific instructional objectives and course content. Even standardized tests in the same content area are usually inadequate as independent criteria, because they are aimed at more general objectives than those measured by a classroom test in a particular course.

Using the total score from our classroom test as a basis for selecting high and low achievers is perfectly legitimate as long as we remember that we are using an internal criterion. In doing so, our item analysis offers evidence concerning the internal consistency of the test rather than its validity. That is, we are determining how effectively each test item is measuring whatever the
whole test is measuring. Such item-analysis data can be interpreted as evidence of item validity only when the validity of the total test has been proven or can be legitimately assumed. This is seldom possible with classroom tests so we must be satisfied with more limited interpretations of our item analysis data.

ii. **A low index of discriminating power does not necessarily indicate a defective item.** Items that discriminate poorly between high and low achievers should be examined for the possible presence of ambiguity, clues, and other technical defects. If none is found and the items measure an important learning outcome, they should be retained for future use. Any item that discriminates in a positive direction can contribute to the measurement of student achievement, and low indexes of discrimination are frequently obtained for reasons other than technical defects.

To summarize, a low index of discriminating power should alert us to the possible presence of technical defects in a test item but should not cause us to discard an otherwise worthwhile item. A well-constructed achievement test will, of necessity, contain items with low discriminating power; to discard them would result in a less, rather than more, valid test.

iii. **Item-analysis data from small samples are highly tentative.** Item analysis procedures focus our attention so directly on a test item’s difficulty and discriminating power that we are commonly misled into believing that these are fixed, unchanging characteristics. This, of course, is not true. Item-analysis data will vary from one group to another, depending on the students’ level of ability, educational background, and type of instruction they have had. Add to this the small number of students available for analyzing the items in our classroom tests, and the tentative nature of our item-analysis data becomes readily apparent. If just a few students change their responses, our indexes of difficulty and discriminating power can be increased or decreased by a considerable amount.

The tentative nature of item-analysis data should discourage us from making fine distinctions among items on the basis of indexes of difficulty and discriminating power. If an item is discriminating in a positive direction, all of the alternatives are functioning effectively, and it has no apparent defects, it can be considered satisfactory from a technical standpoint. The important question then is not how high the index of discriminating power is, but, whether the
item measures an important learning outcome. In the final analysis, the worth of an achievement test item must be based on logical, rather than statistical, considerations.

When used with norm-referenced classroom test, item analysis provides us with a general appraisal of the functional effectiveness of the test items, a means for detecting defects, and a method for identifying instructional weaknesses. For these purposes, the tentative nature of item-analysis data is relatively unimportant. When we record indexes of item difficulty or discriminating power on item cards for future use, we should interpret them as rough approximations only. As such, they are still superior to our unaided estimates of item difficulty and discriminating power.
CHAPTER 6

ELEMENTARY STATISTICS

6.0 ORGANIZING AND DISPLAYING SCORES

When test scores are obtained for a group of students they are usually in haphazard order as shown in Table 6.1.

After careful inspection of the scores on the midterm exam for the 24 students we see that scores in the 70s and 80s are fairly common, that only 3 students had scores in the 90s, and that one student had a score less than 50. Similar statements might be made after inspecting the scores on the final exam. We might even note that there is a tendency for the scores on the final to be slightly higher than those on the midterm and that there is a tendency for students with high scores on the midterm to receive relatively high scores on the final, though the relationship is far from perfect. For example, Student E, who had the higher score on the midterm, had the fourth higher score on the final, while Student W, who had the fifth highest score on the midterm, had the highest score on the final.

From our detailed inspection of the scores in Table A.1, we could actually get a rough idea of all three of the statistical concepts that we are concerned with here.

<table>
<thead>
<tr>
<th>Student</th>
<th>Midterm</th>
<th>Final</th>
<th>Student</th>
<th>Midterm</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>78</td>
<td>85</td>
<td>M</td>
<td>65</td>
<td>80</td>
</tr>
<tr>
<td>B</td>
<td>67</td>
<td>71</td>
<td>N</td>
<td>92</td>
<td>93</td>
</tr>
<tr>
<td>C</td>
<td>88</td>
<td>78</td>
<td>O</td>
<td>53</td>
<td>69</td>
</tr>
<tr>
<td>D</td>
<td>74</td>
<td>71</td>
<td>P</td>
<td>65</td>
<td>75</td>
</tr>
<tr>
<td>E</td>
<td>97</td>
<td>91</td>
<td>Q</td>
<td>83</td>
<td>76</td>
</tr>
<tr>
<td>F</td>
<td>84</td>
<td>88</td>
<td>R</td>
<td>79</td>
<td>74</td>
</tr>
<tr>
<td>G</td>
<td>57</td>
<td>76</td>
<td>S</td>
<td>45</td>
<td>63</td>
</tr>
<tr>
<td>H</td>
<td>65</td>
<td>68</td>
<td>T</td>
<td>95</td>
<td>80</td>
</tr>
</tbody>
</table>
6.1 Simple Ranking

For some uses, it may be sufficient to arrange a set of scores in order of size and to assign a rank to each score. This will indicate the relative position of each score in the group. Ordinarily, the largest score is given a rank of 1, the second largest a rank of 2, and so on until all scores are ranked. The midterm scores from Table A.1 have been rearranged in order of size and assigned ranks to illustrate the procedures. The results are presented in Table 6.2.

Note that when two or more students have the same score (the two 81s, the two 74s, and the three 65s) the average of the ranks for those students is given to each. Thus, the two students who received scores of 81 on the midterm would have been ranked 8 and 9. They are both given the average of these two ranks (8.5) because there is no basis for giving one of the students a rank of 8 and the other one a rank of 9.

Using Table A.2, it is obviously much easier to find the higher and lowest midterm scores, the number of students with scores above 90 or below 70, and to see that scores in the 70s are obtained by students who rank close to the middle of the class than it was with Table A.1. It is also easy to see that half the students score above 75 and half of them below 75. Creating a rank order list for the final exam scores would make it easy to see similar characteristics of those scores. For example, it would make it easier to see that while nine students had scores below 70 on the midterm, only five had scores below 70 on the final.
TABLE 6.2 Ranking test scores

<table>
<thead>
<tr>
<th>Midterm</th>
<th>Rank</th>
<th>Midterm Rank</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>97</td>
<td>1</td>
<td>74</td>
<td>13.5</td>
</tr>
<tr>
<td>95</td>
<td>2</td>
<td>74</td>
<td>13.5</td>
</tr>
<tr>
<td>92</td>
<td>3</td>
<td>70</td>
<td>15</td>
</tr>
<tr>
<td>88</td>
<td>4</td>
<td>67</td>
<td>16</td>
</tr>
<tr>
<td>85</td>
<td>5</td>
<td>65</td>
<td>18</td>
</tr>
<tr>
<td>84</td>
<td>6</td>
<td>65</td>
<td>18</td>
</tr>
<tr>
<td>83</td>
<td>7</td>
<td>65</td>
<td>18</td>
</tr>
<tr>
<td>81</td>
<td>8.5</td>
<td>Tied ranks</td>
<td>62</td>
</tr>
<tr>
<td>81</td>
<td>8.5</td>
<td>8 and 9</td>
<td>58</td>
</tr>
<tr>
<td>79</td>
<td>10</td>
<td>57</td>
<td>22</td>
</tr>
<tr>
<td>78</td>
<td>11</td>
<td>53</td>
<td>23</td>
</tr>
<tr>
<td>76</td>
<td>12</td>
<td>45</td>
<td>24</td>
</tr>
</tbody>
</table>

With only 24 scores there is not a great need to go beyond a rank order display of scores. With more scores, however, it is often helpful to construct grouped frequency distributions and graphs of the score distributions. Grouped frequency distributions of the midterm and final scores are shown in Table 6.3. Note that the scores have been grouped into class intervals, the number of scores falling in each interval has been tallied, and the tallies have been counted to obtain the frequency, or number of scores in each interval. Thus, there were two students with midterm scores in the interval 95-99, one student with a score in the interval 90-94, and so on. The total number (N) is the sum of the numbers in the frequency column. In the finished table the tally column is usually omitted as it has been for the final exam scores.

To construct a grouped frequency distribution it is conventional to choose a class interval that is an odd number and that will result in approximately 10 to 12 intervals. This makes the midpoint of each interval an integer. For our example the class interval is 5 (e.g. 95, 96, 97, 98 and 99 in the highest interval) and the midpoint of each interval is an integer (e.g. 97 is the midpoint of the highest interval). The lower bound of each interval starts with a multiple of the width of the interval (e.g. the lowest interval starts with 45, the next with 50, and so on). All class intervals should be the same size.
TABLE 6.3 Frequency distributions of midterm and final scores

<table>
<thead>
<tr>
<th>Class interval</th>
<th>Tally</th>
<th>Frequency</th>
<th>Class interval</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>95-99</td>
<td>//</td>
<td>2</td>
<td>95-99</td>
<td>1</td>
</tr>
<tr>
<td>90-94</td>
<td>/</td>
<td>1</td>
<td>90-94</td>
<td>3</td>
</tr>
<tr>
<td>85-89</td>
<td>//</td>
<td>2</td>
<td>85-89</td>
<td>3</td>
</tr>
<tr>
<td>80-84</td>
<td>////</td>
<td>4</td>
<td>80-84</td>
<td>4</td>
</tr>
<tr>
<td>75-79</td>
<td>///</td>
<td>3</td>
<td>75-79</td>
<td>4</td>
</tr>
<tr>
<td>70-74</td>
<td>///</td>
<td>3</td>
<td>70-74</td>
<td>4</td>
</tr>
<tr>
<td>65-69</td>
<td>////</td>
<td>4</td>
<td>65-69</td>
<td>3</td>
</tr>
<tr>
<td>60-64</td>
<td>/</td>
<td>1</td>
<td>60-64</td>
<td>1</td>
</tr>
<tr>
<td>55-59</td>
<td>//</td>
<td>2</td>
<td>55-59</td>
<td>1</td>
</tr>
<tr>
<td>50-54</td>
<td>/</td>
<td>1</td>
<td>50-54</td>
<td>0</td>
</tr>
<tr>
<td>45-49</td>
<td>/</td>
<td>1</td>
<td>45-49</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N = 24</td>
<td></td>
<td>N = 24</td>
</tr>
</tbody>
</table>

6.2 Graphic Presentations of Frequency Distributions

A frequency distribution presents test data in a clear, effective manner, and it is satisfactory for most classroom purposes. But if we want to study the distribution of scores more carefully or to report the results to others, a graphic representation may be more useful. The two most commonly used graphs are the histogram (or bar graph) and the frequency polygon (or line graph). Both graphs are presented in Figure 6.1, based on the midterm scores in Table 6.3. The scores are shown along the base line, or horizontal axis, and are grouped into the same class intervals used in Table 6.3. The vertical axis, to the left of the graphs, indicated the number of students earning each score and thus corresponds to the frequency column in Table 6.3.
The histogram presents the data in the form of rectangular columns. The base of each column is the width of the class interval, and the height of the column indicates the frequency, or the number of students falling within that interval. It is as if each student earning a score within a given class interval were standing on the shoulders of the student beneath, to form a human column.

**FIGURE 6.1** Histogram and frequency polygon plotted from midterm exam scores in Table 6.3

The *frequency polygon* is constructed by plotting a point at the midpoint of each class interval at a height corresponding to the number of students, or frequency, within that interval and then joining these points with straight lines. As can be seen in Figure 6.1, the frequency polygon and histogram are simply different ways of presenting the same data. In actual practice we would use only one of the graphs, the choice being somewhat arbitrary.

Histograms or frequency polygons allow us to see the shape of the distribution of scores as well as some of the features we have seen before. With a small number of students, such as the 24 in our example, the shape of the distribution is often jagged, going up and down and up again as you go from left to right. With a large number of scores, however, distributions of scores for many students on standardized tests often appear smooth and bell shaped, not unlike the shape of a normal distribution, and in many cases normal distribution is assumed or used as an approximation.
### Measures of Central Tendency

A measure of central tendency is simply an average or typical value in a set of scores. We all are familiar with the *arithmetic average* obtained by adding all of the scores in a set and dividing this sum by the number of scores. In statistics, this type of average is called the mean and is represented by the letter M (or $\bar{X}$). Two other commonly used measures of central tendency are the *median* (represented by $Mdn$ or $P_{50}$) and the *mode*. The median is the midpoint of a set of scores, that is, the point on either side of which half the scores occur. The mode (fashion) is the score that occurs most frequently. Because the mean, median, and mode are different types of averages, the word *average* should be avoided when describing data. Preciseness requires that the specific type of average be indicated.

The method of determining each measure of central tendency will be described next and is illustrated in Table 6.4.

**TABLE 6.4 Measures of central tendency**

<table>
<thead>
<tr>
<th>Midterm Score (X)</th>
<th>Final Score (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>97</td>
<td>96</td>
</tr>
<tr>
<td>95</td>
<td>94</td>
</tr>
<tr>
<td>92</td>
<td>93</td>
</tr>
<tr>
<td>88</td>
<td>91</td>
</tr>
<tr>
<td>85</td>
<td>88</td>
</tr>
<tr>
<td>84</td>
<td>87</td>
</tr>
<tr>
<td>83</td>
<td>85</td>
</tr>
<tr>
<td>81</td>
<td>81</td>
</tr>
<tr>
<td>81</td>
<td>80</td>
</tr>
<tr>
<td>79</td>
<td>80</td>
</tr>
<tr>
<td>78</td>
<td>80</td>
</tr>
<tr>
<td>76</td>
<td>78</td>
</tr>
<tr>
<td><strong>Median = 75</strong></td>
<td><strong>Median = 77</strong></td>
</tr>
<tr>
<td>74</td>
<td>76</td>
</tr>
<tr>
<td>74</td>
<td>76</td>
</tr>
<tr>
<td>70</td>
<td>75</td>
</tr>
<tr>
<td>67</td>
<td>74</td>
</tr>
<tr>
<td>65</td>
<td>72</td>
</tr>
<tr>
<td>65</td>
<td>71</td>
</tr>
<tr>
<td>65</td>
<td>71</td>
</tr>
<tr>
<td>62</td>
<td>69</td>
</tr>
<tr>
<td>58</td>
<td>68</td>
</tr>
<tr>
<td>57</td>
<td>67</td>
</tr>
<tr>
<td>53</td>
<td>63</td>
</tr>
<tr>
<td>45</td>
<td>58</td>
</tr>
<tr>
<td><strong>Sum of X =</strong></td>
<td><strong>Sum of Y = 1,873</strong></td>
</tr>
</tbody>
</table>

50% of scores
6.3.1 The Mean (\( M \) or \( X \))

The mean, or arithmetic average, is the most widely used measure of central tendency. Because it is calculated by adding a series of scores and then dividing this sum by the number of scores, the computation can be represented by the following formula:

\[
\bar{X} = \frac{\text{Sum of all scores}}{\text{Number of scores}} \\
\bar{X} = \frac{\sum X}{N}
\]

Where

\( \Sigma \) = the sum of

\( X \) = any score

\( N \) = number of scores

Applying this formula to the scores in Table A.6, produces a mean of 73.92 for the midterm and 78.04 for the final. The mean takes into account the value of each score, and so one extremely high or low score could have an appreciable effect on it.

6.3.2 The Median (\( Mdn \) or \( P_{50} \))

The median is a counting average. It is determined by arranging the scores in order of size and counting up to (or down to) the midpoint of the set of scores. If the number of scores is even (as in Table A.6) the median will be halfway between the two middlemost scores. When the number of scores is odd, the median is the middle score.

The median is a point that divides a set of scores into equal halves, and thus the same number of scores falls above the median as below the median, regardless of the size of the individual scores. Because it is a counting average, an extremely high or low score will not affect its value.
6.3.3 The Mode

The mode is simply the most frequent or popular score in the set and is determined by inspection. In Table A.6 the mode is 65 for the midterm because the largest number of persons made that score. The mode is the least reliable type of statistical average and is frequently used merely as a preliminary estimate of central tendency. A set of scores sometimes has two modes and is called bimodal.

6.3.4 Measures of Variability

A set of scores can be more adequately described if we know how much they spread out above and below the measure of central tendency. For example, we might have two groups of students with a mean test score of 70 but in one group, the span of scores is from 60 to 80, and in the other the span is from 50 to 100. These represent quite spreads of performance. We can identify such differences by numbers that indicate how much scores spread out in a group. These are called measures of variability, or dispersion. The three most commonly used measures of variability are the range, the quartile deviation, and the standard deviation.

6.4 The Range

The simplest and crudest measure of variability is the range, calculated by subtracting the lowest score from the higher score. In the preceding example, the range of scores in the first group is 20 points and in the second, 40 points. The range provides a quick estimate of variability but is undependable because it is based on the position of the two extreme scores. The addition or subtraction of a single score can change the range significantly. In the preceding example, the ranges of the two groups would become equal if we added two students to the first group: one with a score of 50 and another with a score of 100. It is obvious that a more stable measure of variability would be desirable.
For our example with the midterm and final exam scores, however, the smaller range for the final (96-58 = 38) than for the midterm (range = 97-45=52) accurately reflects the smaller variability in scores on the final exam that we have previously noted and that will be quantified by the other measures of variability.

6.5 COEFFICIENT OF CORRELATION

The next statistical measure that we shall consider is the correlation coefficient. The meaning of the correlation coefficient and its use in describing the validity and reliability of test scores can be found in Chapters 3 and 4. Basically, a coefficient of correlation expresses the degree of relationship between two sets of scores by numbers ranging from -1.00 to +1.00. A perfect positive correlation is indicated by a coefficient of +1.00 and a perfect negative correlation by a coefficient of -1.00. A correlation of .00 indicates no relationship between the two sets of scores. Obviously, the larger the coefficient (positive or negative), the higher the degree of relationship expressed.

6.6 LINEAR REGRESSION COEFFICIENTS

Correlation coefficients provide an indication of the strength of relationship between two sets of scores. In many situations, however, we are interested in using one set of scores to predict the scores that would be obtained on another measure. For example, we might want to predict the grades that students will obtain from their test scores. Or we might want to predict scores on an achievement test from scores obtained on an aptitude test. In the example used in this appendix, we can show how linear regression is used to make such predictions.

The linear regression equation for predicting $Y$ (the final) from $X$ (the midterm) is:

$$\hat{Y} = a + b (X)$$

Where $a$ and $b$ are the regression coefficients. The caret (^) over the $Y$ indicates that is the predicted $Y$ score rather than the one that is actually achieved. The coefficient $b$ is known as the slope, and the coefficient $a$ is known as the intercept. They can be computed using the following formulas.

$$b = r \left( \frac{SD_Y}{SD_X} \right)$$
and
$$a = \bar{Y} - b (\bar{X})$$
where
r is the relation coefficient,
SD$_X$ and SD$_Y$ are the standard deviations of X and Y, and
$X$ and $Y$ are the means of $X$ and $Y$, respectively.

For the midterm and final scores the coefficients are
\[
b = \frac{(0.7441)(10.0411)}{(13.5355)} = 0.5520
\]
And
\[
a = 78.0417 - (0.5520)(73.9167) = 37.24
\]

Hence, the predicted scores on the final from scores on the midterm are given by the following equation.

\[
\hat{Y} = 37.24 + 0.552 \times (X)
\]

Using this formula for Students A, B and C who had midterm scores of 78, 67 and 88 respectively, yields the following predicted scores on the final.

Student A : $\hat{Y} = 37.24 + 0.552 \times (78) = 80.3$

Student B : $\hat{Y} = 37.24 + 0.552 \times (67) = 74.2$

Student C : $\hat{Y} = 37.24 + 0.552 \times (88) = 85.8$

The actual final exam score for Students A, B, and C are 85, 71 and 78, respectively. Thus, Students A and B did somewhat better on the final than predicted while Student C did somewhat worse than predicted. The difference between the actual Y scores and the predicted ones is known as the error of prediction. The errors of prediction get smaller in comparison to the standard deviation of $Y$ as the correlation increases. Thus, the higher the correlation, the more accurate the prediction will be.

Of course, if we already know the scores on $Y$ there is no need to predict them, other than to see who did better or worse than would have been predicted from previous performance. A more common use of the regression equation for making predictions occurs when the equation that is obtained for one group of students (e.g. the freshman class of 1995) is used to predict the performance of another group (e.g. the applicants for freshman class of 1996).
example, we could use the regression of freshman grades on test scores obtained for the 1995 freshman class to predict the grade that 1996 applicants would obtain if they are admitted.
7.0 TEST ANXIETY

Sarason (1959), a major early contributor to theory and research on test anxiety, accurately observed that ‘we live in a test-conscious, test-giving culture in which the lives of people are in part determined by their test performance’ (p. 26). It is therefore not surprising that test anxiety is a pervasive problem on the college campus. Many students are so disturbed by test anxiety that they must seek professional assistance to help them cope with its debilitating effects. Nearly 40 years ago, Brown (1938), called attention to the seriousness of the problem of test anxiety for college students. In commenting on the causes of two students suicides at the University of Chicago, Brown notes that “one of these was definitely due to worry over an approaching examination and the other presumably was ……These incidents show that students are taking their examinations more and more seriously and that the emotional reactions of the students before examinations is an important problem.” (Brown, 1938, pp. 11-12).

Since World War II, psychologists and counselors have become increasingly concerned with understanding the nature of test anxiety and the development of effective methods for its treatment. It has been repeatedly demonstrated that people who are high in test anxiety experience decrements in performance in evaluating situations. Anxious individuals perceive testing situations as personally threatening and respond to them with intense emotional reactions. Evaluative situations evoke task irrelevant, self-centered, worry response that interfere with effective performance on cognitive-intellectual tasks.

Test anxiety is an unpleasant feeling or emotional state that has physiological and behavioral concomitants and that are experienced in formal testing or other evaluative situations. When test anxiety is experienced, a variety of cognitive and attentional processes are called into play that interfere with effective and successful task performance. Current theoretical formulations try to explain the nature and functioning of these cognitive and attentional variables.
Test stimuli are those which the individual associates with evaluation. These may be immediate events, such as a teacher’s remark, “We will have a quiz today,” or “John, how would you answer that?”. Test stimuli may also be related to the future, such as the decision to major in a premedical program with the knowledge that, four years later, one will be faced with MCAT exams and personal interviews at the medical schools. Test stimuli are conditioned stimuli. Their meaning to the individual depends on prior experience. Thus, what is a stimulus of test anxiety for one person may be a neutral event for another.

Interpretation of test stimuli depends on the nature of one’s prior experience with these stimuli. They may be perceived as having interesting or positive meaning or as being threatening or neutral. Some individuals can approach evaluation as a positive event. For example, one may approach a test with the mature view that I will either succeed or not in meeting my expectations in this instance, and I will learn and grow from whatever happens.

This is in marked contrast to those whose fear of failure causes them never to set out to achieve, to focus narrowly and intently on one area of achievement at the expense of failing to develop in other areas, or to bungle through challenging tasks. It is important to remember that the interpretation of test stimuli is, by definition, an interpretation based on one’s own past history. Thus, fear of failure and other negative interpretation of test stimuli are not fear of failing to carry out the operations required at the time. Rather, these negative interpretations involve “plugging in” to old ideas, such as, “If I fail at this, my life will be less worthwhile”, or “I will have fulfilled my father’s views that I am not worth anything”, or “no one will respect me”.

A-state reactions vary depending on interpretations of prior experience and on the nature of the test stimulus. The A-state reaction may consist of heightened arousal, vigilance and a sense of enthusiasm, or it may include fear and worry, confusion, illness, anger lowering of self-esteem and other negative events.

Cognitive reappraisal refers to the way in which an individual responds to this or her A-state. These responses may be constructive, defensive, avoidant or a combination of these kinds of responses.
Coping, avoidance and defensiveness refer to the nature of the feelings, approaches and outcomes that may occur; for example, the task may be successfully or unsuccessfully completed. It may be consciously and confidently addressed, fearfully approached, avoided or blundered through. The nature of one’s performance may be fully acknowledged, denied, or blamed on someone else. The individual may feel good, bad, unaware, or indifferent about the task and the performance.

As this general model in Figure 1 suggest, there are adaptive test anxiety processes that healthy people experience every day, and there are maladaptive processes to which we refer when we speak of high anxiety. People who experience high anxiety in test situations are very sensitive to cues that suggest the imminence of testing and interpret testing situations as a serious threat to their well-being. The resulting A-state response is a powerful, unpleasant, and disruptive emotional reaction. The cognitive reappraisal typically involves considering a number of unconstructive ways to deal with the test and the anxiety. The coping behaviors followed are less constructive and effective than is desirable and may be accomplished by defensive and avoidant behavior.

Thus, we must distinguish between two meanings of the term, test anxiety, anxiety as a state and anxiety as trait (Spielberger, 1966). State anxiety is a transitory state that occurs when an individual perceives stimuli of (real or imagined) test and responds with certain emotions and behavior. Trait anxiety refers to a relatively stable personality characteristic, the disposition to perceive as threatening a wide range of the stimuli that are associated with tests and the tendency to respond to these with extreme A-state reactions.

Evaluation situations that threaten self-esteem evoke higher levels of A-state response in high A-trait individuals than low A-trait individual. Differential level of A-state reaction is related to level of performance in intellectual tasks. In contrast, physical danger does not evoke such differential A-state responses; rather, it evokes a similar increase in A-state for high and low A-trait persons.
7.1 Theories of the Effect of Test Anxiety

The generally accepted current explanation of the negative effects of test anxiety is that ineffective cognitive strategies and attentional deficits cause poor task performance. Children with low level of anxiety appear to become deeply involved in evaluative task, but highly anxious children do not. Highly anxious children seem to experience attentional blocks, extreme concern with autonomic and emotional self-cues, and cognitive deficits such as misinterpretation of information. The highly anxious child’s attentional and cognitive deficits are likely to interfere with learning and responding in evaluative situations and result in lowered performance.

The conceptualizations suggest the importance of cognitive factors as mediating influences in the effects of test anxiety on children’s learning and performance. These cognitive factors influence the perception of a situation as evaluative or not (Sarason, 1978). The cognitive activities considered important in the mediation of test anxiety are generally conceptualized as attentional in nature (Sarason, 1975, 1978; Wine, 1871). These mechanisms influence stimulus reception and interpretation as well as overt behavior (Sarason, 1975, 1978). Hence, attention deficits in high-anxious individuals have been a major concern in testing or evaluative situations.

Some researchers (e.g. Liebert & Morris, 1967; Morris & Liebert, 1970) have examined the highly anxious person’s attention to self-stimuli as opposed to task stimuli. Worry is conceptualized as cognitive concern over performance in a task; emotionality is viewed as an automatic arousal aspect of test anxiety. The adverse effects of test anxiety are presumed to be the result of a division of attention between concern over task performance, on the one hand, and the physiological aspects of arousal, on the other hand. The highly anxious person attends more to the autonomic aspects of arousal and less to the task than does the person with low anxiety. This division of attention results in poorer test performance for persons than persons who are less anxious. The primary concept from this perspective is worry, because it results in a cognitive concern about one’s ability relative to others and about the consequences of failure. This concern replaces attending to and working at the task at hand.
Worry is defined as cognitive concern over task performance, and emotionality as the autonomic arousal aspect of anxiety indicate that worry is the more stable, enduring component of test anxiety; whereas, self-report of emotionality has a more transient quality and is confined to evaluating situations. Worry affects cognitive performance and performance expectations. Emotionality does not relate consistently to these variables. Thus it is the cognitive, self-preoccupied worry component of test anxiety that interferes most directly with task performance. The higher self-reported levels of emotionality on the part of highly anxious persons also probably reflect greater attention to internal events as opposed to externally directed, task focused attention.

According to attentional theory, it should be possible to negate the deleterious effects of test anxiety by helping the child focus attention more directly on the task. Research also indicates that providing task-relevant strategies helps highly anxious children better attend to evaluative tasks and improve their performance. Other research (Sarason, 1972) indicates that providing task-oriented instructions, cues about expected performance, task-effective models, and memory supports facilitates the performance of highly anxious persons in evaluative situations.

As described previously, the cognitive and attentional deficits associated with high test anxiety are partly the result of parental and other adult reactions to the child’s success and failure in evaluative situations. Hill (1972) has placed special emphasis on the developing child’s success and failure experience in explaining why some children become highly test-anxious. Children with low test anxiety generally have a history of success in school and other evaluative situations and experience generally positive interactions with adults in evaluative settings. As a result, they develop a relatively higher motivation toward success and learn to rely on their own evaluations of performance for guidance in problem solving. Since highly anxious children have a generally poorer history of success in school and other evaluative situations and have experienced somewhat more punitive interactions with evaluative adults both parents and teachers, they develop problem-solving strategies that indicate motivation to avoid failure and criticism rather than to approach success. Highly anxious children, then, are apt to develop a high dependence on adults for evaluation of their performance and for direction in problem solving.
7.1.1 Measuring Test Anxiety

A number of questionnaires have been developed to measure test anxiety. These include the following:

1. Test Anxiety Questionnaires (Mandler & Sarason, 1952)
2. Test Anxiety Scale (Sarason, 1978)
3. Worry and Emotionality questionnaire (Morris & Liebert, 1968)
4. Achievement Anxiety Test (Alpert & Haber, 1960), and
5. State Trait Anxiety Inventory

7.1.2 State-Trait Anxiety Inventory

The State Trait Anxiety Inventory (STAI) is by far the most widely used measure of one of mankind’s most pervasive questions. Traditionally, anxiety has been viewed as having two distinct forms. The state form of anxiety consists of the transitory feelings of fear or worry, which most of us experience from time to time. The trait form of anxiety is the relatively stable tendency of an individual to respond anxiously to a stressful situation. The STAI is a 40-item measure which provides information about a person’s levels of state and trait anxiety. The construction of the STAI by Charles Spielberger and his colleagues began in 1964 and was published in 1970.

The STAI were designed for high school and college students and for adults, but all of the items were written below the sixth-grade reading level. As a result, the test could also be used with junior high schools students. Developed as a self-administered test, the STAI may be given either individually or in groups. Although there is no time limit, most subjects will complete both scales in less than 10 minutes.

When the full 40-item STAI is administered, as recommended in the manual, two scores will be obtained. One score will reflect the person’s current level of state anxiety and can range from 20 to 80 with higher scores reflecting more anxiety. The other score indicates the person’s general level of trait anxiety and also can range from 20 to 80 with higher scores indicating more anxiety.
Scoring the STAI is also straightforward. A subject’s score on each scale is simply the sum of the responses to the 20 items on that scale. The only complication is that the responses to the 10 anxiety-absent items on the State-Anxiety Scale and the 9 anxiety-absent items on the Trait-Anxiety Scales need to be identified (e.g. 1-4, etc) before they are summed.

The interval consistency of the Trait-Anxiety Scale, ranged from 0.89 to 0.91 across male and female samples of working adults, military recruits, and college and high school students. For the State-Anxiety Scale this range from 0.86 to 0.95. This reliability was measured using samples of 202 male and 22 female high school students and 1728 Air Force recruits (Chaplin, 1984).

7.1.3 Test Anxiety and Examination Performance

The study of test anxiety was started by Madler and Sarason (1952) now famous investigator of anxiety and learning. Sarason and Madler believed that the testing situation evokes both learned task drives and learned anxiety drives. Some of the anxiety drives are task relevant while others are task irrelevant. The learned task drives and the task-relevant anxiety drives facilitate test performance, while the task-irrelevant anxiety drives decreases test performance. The task-irrelevant anxiety is suffered by the highly anxious person during examination, resulting in a lowered performance.

The anxious student is commonly described as one who knows the course material but, because of anxiety, is unable to demonstrate his or her knowledge on an examination (Paul and Eriksen, 1964). Therefore, a significant inverse relationship between test anxiety scores and examination scores is expected.

Sarason (1957) conducted a study to find the relationship between test anxiety, general anxiety, and intellectual performance of 305 liberal arts undergraduates at Yale University. Most of the subjects were administered the Test Anxiety Questionnaire (TAQ) and the General Anxiety Questionnaire (GAQ). He found that TAQ scores tended to correlate negatively with measures of academic achievement, although with increase
in number of years in college the negative correlation disappeared. Highly anxious subjects performed at a significantly lower level than did less anxious subjects.

Doctor and Altman (1969) conducted a study to find the correlation between worry and emotionality test anxiety. Their subjects were 159 sophomores in a general psychology course at the University of Kentucky. The subjects were tested with the Test Anxiety Inventory developed by Liebert and Morris (1968), before and after the final examination. Results of their findings showed that emotionality scores dropped significantly following completion of the examination irrespective of initial level of test anxiety or performance expectancy. As hypothesized, worry was more highly correlated with expectancy of success ratings than emotionality, but the prediction that worry scores would not change following the examination was replicated only for the relatively less worried subjects. Highly worried subjects evidenced significant decrements in pre-to post examination assessments of worry, suggesting that these scores had, perhaps, been incremented by the perceived aversion to the test-taking situation. The worry component of test anxiety was more highly associated with actual performance on the final examination than emotionality with highly worried subjects performing significantly poorer than less worried subjects. Emotionality level was relatively low. It was pointed out that differential changes in worry at the high and low values of this variable had not been observed in previous studies (Spiegler, Morris & Liebert, 1968).

Deffenbacher (1977) did a study to find the relationship between worry and emotionality to performance on the Miller analogies test using students at the University of Oregon. In the experiment 52 males and 27 females completed the Worry-Emotionality Inventory just prior to taking the Miller Analogies Test. Worry and emotionality were significantly correlated (r=0.69).

State test anxiety, worry and emotionality were inversely related to performance, but partial correlations indicated that only worry was correlated with performance when the common variance between worry and emotionality was partially out. Highly worried students performed less well than less worried students. The effects of emotionality were nested within worry levels. At low levels of worry, emotionality was unrelated to performance, but at high levels of worry, high emotionality was associated most
negatively with performance. That is, the negative effects of emotionality were nested within the upper range of worry. No sex differences were found. The results were interpreted as supporting the conceptual distinction between worry and emotionality. Similarly, Spielberger et al. (1978) found a negative correlation between grade point average and worry scale of the test anxiety inventory.

In summary, the research on test anxiety and examination performance supports the position that test anxiety consists of two components, worry and emotionality. Further evidence (Deffenbacher, 1977) shows that high worry is the concern and effect causing poor performance of students in evaluative measures. The research suggests that highly anxious students perform poorly as compared with less anxious students on evaluating measures. Most of these students were in liberal arts and none in science. It has also been criticized that most of the studies are poor and the instruments can be easily falsified. So these studies leave some doubt as to their validity. In spite of the criticisms, many studies have shown the same results. Knowing the negative effects of test anxiety in evaluative measures, suggests the concern of neutralizing them in the complex school system? Almost every facet of school curriculum requires an evaluation. In fact, evaluation is a part of the teaching and learning process (Grondlund, 1985). As mentioned earlier in the theory of test anxiety, the negative effects of test anxiety can be lessened if evaluations measures are not stressed. Wine (1979) made several suggestions for altering the evaluative character of the school, include:

* The teacher’s role should be redefined. Rather than the traditional heavily evaluative character of the teacher’s interactions with children, a shift is proposed to a task-oriented, information-giving role model, showing excitement and involvement in specific subject matter.

* Feedback to the subjects must be informative, immediate, and directed to specific tasks rather than global.

* The teacher, as a social-evaluative figure, should remove himself or herself as much as possible from the test-anxious child’s environment.
* Task materials should be made as interesting and engaging as possible and should be presented with task-oriented, curiosity and attention-eliciting instructions, with a minimal evaluative component.

With better understanding of the relationship between test anxiety and evaluative measure, teachers and teacher educators should be able to reduce student anxiety and improve students' evaluative performance.

### 7.2 Statement of the Problem

This study investigates relationships among student testing, anxiety, attitude, and achievement in chemistry courses at college level. It considers the effect of conventional testing and frequent testing on student achievement in chemistry. Also it attempts to determine if frequent testing results in lower test anxiety. If further attempts to determine if frequent testing results in improving student attitudes toward science.

The purpose of the study is to investigate the following:

1. Which of the two approaches (i.e. frequent vs. conventional) to testing in a general chemistry for college students at University Technology of Malaysia would result in greater achievement in chemistry?

2. The effect of these approaches of testing on students test anxiety.

3. The effect of these approaches of testing on students' attitude toward science.

### 7.3 The Research Hypotheses

The objectives of the study were to obtain data that could be used to test the following null hypotheses:

1. There is no significant difference in achievement in general chemistry between the control and experimental groups.

2. There is no significant difference in achievement in general chemistry between students with high test anxiety in the control and experimental groups.
3. There is no significant difference in achievement in general chemistry between students with low test anxiety in the control and experimental groups.

4. There is no significant difference in achievement in general chemistry between students with positive attitudes toward science in the control and experimental groups.

5. There is no significant difference in achievement in general chemistry between students with negative attitudes toward science in the control and experimental groups.

6. There is no significant difference in achievement between students with high and low test anxiety irrespective of the control and experimental groups.

7. There is no significant difference in achievement between students with positive and negative attitudes toward science irrespective of the control and experimental groups.

8. There is no significant difference in mean test anxiety scores between the control and experimental groups.

9. There is no significant difference in mean attitude toward science scores between the control and experimental groups.

A 0.05 level of significance was used as a basis for rejecting the hypotheses.

7.4 Selection of Sample

The sample consisted of 278 students who were taking first year Chemistry at the University Technology of Malaysia. The sample included males and females. All the students in the sample were given an achievement test in Chemistry, Malay Version of the State Trait Anxiety Inventory (Speilberger, 1977) and Science Attitude Questionnaire (Sumner, 1973) as pretests. The Science Attitude Questionnaire and State Trait Anxiety Inventory were given at the same time followed by a Chemistry achievement test on the same day. A fifteen-minute rest was provided between administration of the instruments. All the pretests were given during the first week of the semester.

The pretests were marked and students’ responses were tabulated. The mean scores on each test were calculated. The mean scores on the State Trait Anxiety Inventory and Science
Attitude Questionnaire were used to assign students to high-low test anxiety and positive-negative attitude toward science for the purpose of assigning students to groups. The mean score for the test anxiety was 50.67. The mean score for the attitude toward science was 174. A score 51 and above was considered as high test anxiety while 50 and below was considered as low test anxiety. An attitude toward science score of 174 and above was considered as positive attitude toward science while 173 and below was considered as negative attitudes toward science. On this basis the samples were separated into high-low anxiety and positive-negative attitudes toward science.

Within the group of high test anxiety, positive or negative attitude toward science, pairs of high test anxiety and positive attitude toward science students were identified. Students in the matched pairs were randomly assigned to group H (+) in the control group and one to group H (+) in the experimental group. Similar procedure was done for assigning groups of H (-) in the control and experimental groups. This procedure was repeated until the group of high test anxiety, positive or negative attitude were exhausted, resulting in 32 students randomly assigned to the groups H (+) in the control and experimental groups.

Within the group of low test anxiety, positive or negative attitude toward science, pairs of low test anxiety and positive attitude toward science students were identified. Students in the matched pairs were randomly assigned to group L (+) in the control group and one to group L (+) in the experimental group. Similar procedure was one for assigning groups of L (-) in the control and experimental groups. This procedure was repeated until the group of low test anxiety, positive or negative attitude were exhausted, resulting in 51 students randomly assigned to the groups L (-) in the control and experimental groups. Also there were 27 students randomly assigned to groups L (-) in the control and experimental groups.

From Table 1, it can be seen that the high test anxiety-positive attitude group in the experimental and control groups had 32 students each. The high test anxiety-negative attitude in the experimental and control groups had 29 students each. The low test anxiety-positive attitude in the experimental and control groups had 51 students each. The low test anxiety-negative attitude in the experimental and control groups had 27 students each. Thus the control and experimental groupings were comprised of both high and low anxiety students. Also the experimental and control groups included students of positive and negative attitudes toward science. Table 1 shows the distribution of the groupings.
TABLE 1 Table of Groupings

<table>
<thead>
<tr>
<th>Anxiety</th>
<th>Control</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>High</td>
<td>(+)</td>
<td>(-)</td>
</tr>
<tr>
<td>Low</td>
<td>(+)</td>
<td>(-)</td>
</tr>
<tr>
<td>Attitude</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+)</td>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>(-)</td>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>Number of students</td>
<td>32</td>
<td>29</td>
</tr>
</tbody>
</table>

According to Cohen’s (1969) sample size table, the minimum sample size required when \( r = 0.35 \), \( (1 - \beta) = 0.80 \) and significance level of 0.05, was 27 subjects per cell. Each cell exceeded the minimum sample size.

7.5 Research Variables

The dependent variables in the study were student achievement in Chemistry, measures of trait anxiety and attitude toward science. These variables were measured by administering selected items from the American Chemical Society Cooperative Examinations in Chemistry, State Trait Anxiety Inventory (Spielberger, 1977) and Science Attitude questionnaire (Sumner, 1978).

The independent variable in the study was frequent testing. This was accomplished by administering achievement tests in Chemistry at two-week intervals throughout the treatment.

7.6 Criterion Instruments

The criterion instruments used in the study were as follows: 1) selected items from the American Chemical Society Cooperative Examinations in Chemistry 2) State-trait Anxiety Inventory, and 3) Science Attitude Questionnaire. Each of these instruments were administered as pretests at the beginning of the semester and as a posttest at the end of the semester.

7.7 Achievement in Chemistry

The test items used to measure achievement in chemistry were selected from the American Chemical Society Cooperative Examinations and translated into the Malay language. Permission was sought to use selected items. The items were selected from the American Chemical Society Cooperative Examination in General Chemistry form 1981B (Brief Test) and the American Chemical Society Cooperative Examination in General Chemistry, Form 1981. A
A total of 130 items from the two examinations were presented to a Delphi panel consisting of eight Chemistry instructors from the Chemistry Department at the University Technology of Malaysia. Out of 130 items, 60 items were selected by the panel for inclusion in the instrument.

7.8 Scale and Scoring

All items in the American Chemical Society Cooperative Examinations in Chemistry were multiple-choice and were scored using an interval scale. Each item had a one point value. The same format, scale and scoring were used for the selected items.

7.9 Content Validity

Content validity was determined by a Delphi panel of eight Chemistry instructors from the Chemistry Department at University Technology of Malaysia. A Delphi technique was used to construct the instrument. The first draft of this instrument called for a judgment about determining whether or not each item measured one of the content objectives of the Chemistry course. Each item was judged and either accepted or rejected. Items were selected on the basis of agreement of 7 or more panel members.

On the second round, each panel member, isolated from other members, received the selected proposed list of items and was asked to evaluate the proposed list of items as to which items clearly measured the stated objectives of the chemistry course.

Using the Delphi procedure, sixty items were identified by the panel as clearly measuring the stated objectives of the chemistry course. These items were selected to be included in the final version of the chemistry achievement test.

7.10 Reliability

Reliability of the instrument was determined by using Kuder Richardson formula 20 (Sax, 1974).

\[
KR_{20} = \frac{n}{n - 1} \left( \frac{SD^2 - \sum pg}{SD^2} \right)
\]

Where \( n \) = the number of items on the test.
SD² = the variance of scores (the standard deviation squared)

p = the difficulty level of each item or the proportion of the group that responded correctly

q = the proportion that missed the item (1-p)

Students’ scores on the chemistry achievement posttest were used to determine the reliability. The reliability was found to be 0.86. A value of 0.75 was set as the stated value of acceptance. This level of acceptance was recommended by Borg (1984).

7.11 Test Anxiety

The instrument for measuring test anxiety was a translated version of the State Trait Anxiety Inventory developed by Speilberger (1977). Permission was sought for using the test. All items in the original instrument were translated into the Malay language. Although the state and trait were given in the questionnaires only the trait questionnaires were used for the analysis. The method of scoring the items are shown on Appendix D page 162.

7.12 Validation of the Translation

Following translation, the items for an experimental form were presented for evaluation by eight subject matter and language experts. Agreement among expert evaluators was interpreted as providing evidence of the content validity of the translated scale. An item was accepted when all evaluators judged the item to be satisfactory translated. When the translators and evaluators disagreed concerning the translation of a particular item, two or more different translations of such items were included in the preliminary form of the scale. The evaluators were then given the new version of the instrument; they re-evaluated the items until every item in the final instrument was judged to have been satisfactorily translated. Items in the final version were judged to be good or excellent by all.

7.13 Attitude toward Science

This instrument for measuring attitude toward science was a translated version of the science Attitude Questionnaire (Sumner, 1978). Permission was sought for using the test. All items in the original instrument were translated into the Malay language. The method of scoring the items were shown on Appendix F page 177.
7.14 Reliability
Reliability of the instrument was determined by using the Hoyt and Stunkard (1952) method. This method provided a straightforward solution to the problem of estimating the reliability coefficient for unrestricted scoring items. The formula for calculating the reliability was previously included. The scores from pretest and posttest attitude toward science were used to determine the reliability of the instrument. The reliability of the pretest and posttest were found to be 0.82 and 0.85 respectively. A value of 0.65 and above was judged as the level of acceptance in agreement with the recommendation by Borg (1984).

7.15 Research Design and Procedures
One-hundred-thirty-eight subjects in the experimental group received then multiple-choice questions once every two weeks. While 139 subjects in the control group received twenty multiple-choice questions once a month. During the whole semester, the experimental group received a total of three tests. Time required for testing and the amount of contents were adjusted evenly between subjects in the control and experimental group so as not to be biased. The tests were administered on Saturday for both groups. Students were also told the value of the test with respect to the course grading system.

7.16 Testing the Hypotheses
There is no significant difference in achievement in General Chemistry between the control and experimental groups. Hypotheses 1 was tested using a three-way analysis of covariance. The means and standard deviations of student scores in the experimental and control groups on the general chemistry achievement pretest and posttests are provided in Table 5, while the results of the three-way analysis of covariance with (pretest achievement in chemistry scores) as the covariates are presented in Table 6.
Table 5 Mean and standard deviation of the pretest and post test chemistry score in the control and experimental groups (N=278 students).

<table>
<thead>
<tr>
<th>Grouping</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest score</td>
<td>11.96</td>
<td>4.61</td>
</tr>
<tr>
<td>Post test score</td>
<td>28.14</td>
<td>4.62</td>
</tr>
<tr>
<td>Experimental Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest score</td>
<td>12.00</td>
<td>4.60</td>
</tr>
<tr>
<td>Post test score</td>
<td>35.55</td>
<td>7.69</td>
</tr>
</tbody>
</table>

Table 6 Three-way analysis of covariances for chemistry achievement for treatment, anxiety and attitude toward science.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>Sum of squares</th>
<th>Mean Square</th>
<th>F-ratio</th>
<th>Sign. level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chem. Pretest</td>
<td>3</td>
<td>1369.8321</td>
<td>1369.8321</td>
<td>38.782</td>
<td>0.0000</td>
</tr>
<tr>
<td>Main effect</td>
<td>1</td>
<td>3852.1930</td>
<td>1317.3977</td>
<td>37.298</td>
<td>0.0000</td>
</tr>
<tr>
<td>Treatment</td>
<td>1</td>
<td>3787.4191</td>
<td>3787.4191</td>
<td>107.228</td>
<td>0.0000</td>
</tr>
<tr>
<td>Anxiety</td>
<td>1</td>
<td>159.1247</td>
<td>159.1247</td>
<td>4.505</td>
<td>0.0347</td>
</tr>
<tr>
<td>Attitude</td>
<td>1</td>
<td>15.8693</td>
<td>15.8693</td>
<td>0.449</td>
<td>0.5104</td>
</tr>
<tr>
<td>2-factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td>3</td>
<td>55.5949</td>
<td>18.5316</td>
<td>0.525</td>
<td>0.6657</td>
</tr>
<tr>
<td>Trmt x Anxiety</td>
<td>1</td>
<td>1.0901</td>
<td>1.0901</td>
<td>0.031</td>
<td>0.8626</td>
</tr>
<tr>
<td>Trmt x Att</td>
<td>1</td>
<td>15.01876</td>
<td>15.01876</td>
<td>0.425</td>
<td>0.5219</td>
</tr>
<tr>
<td>Anxiety x Att</td>
<td>1</td>
<td>38.0684</td>
<td>38.0684</td>
<td>1.078</td>
<td>0.3001</td>
</tr>
<tr>
<td>Residual</td>
<td>270</td>
<td>9536.7290</td>
<td>35.3212</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>277</td>
<td>14914.349</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5 shows that the posttest adjusted mean scores in general chemistry of students in the experimental and control groups were 35.55 and 28.14 respectively. From Table 6, it can be seen that the students in the experimental group achieved significantly higher in general chemistry that did students in the control group. Base on these findings, Hypothesis 1 was rejected at 0.05 level of significance.
Hypotheses 2
There is no significant difference in achievement in General Chemistry between high test anxiety students in the control and experimental groups.
Hypotheses 2 was tested using a one-way analysis of covariance. Table 7 shows the mean and standard deviation of students score in general chemistry achievement test of high anxiety students in experimental and control groups. Table 8 shows the results of the one-way analysis of covariance with pretest achievement in chemistry scores serving as the covariates.

Table 7  Mean and standard deviations of chemistry scores of high test anxiety students in the control and experimental groups (N=122 students)

<table>
<thead>
<tr>
<th>High anxiety</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Control Group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Score</td>
<td>11.49</td>
<td>4.25</td>
</tr>
<tr>
<td>Post Test Score</td>
<td>27.20</td>
<td>4.60</td>
</tr>
<tr>
<td><strong>Experimental Group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Score</td>
<td>12.13</td>
<td>4.07</td>
</tr>
<tr>
<td>Post Test Score</td>
<td>34.67</td>
<td>8.18</td>
</tr>
</tbody>
</table>

Table 8  One-way analysis of covariance of high anxiety students in the control and experimental groups.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>Source of squares</th>
<th>Mean square</th>
<th>F-ratio</th>
<th>Sign level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre Chem.</td>
<td>1</td>
<td>1073.0236</td>
<td>1073.0236</td>
<td>29.011</td>
<td>0.0000</td>
</tr>
<tr>
<td>High anxiety</td>
<td>1</td>
<td>1511.0247</td>
<td>1511.0247</td>
<td>40.853</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>119</td>
<td>4401.4272</td>
<td>36.9867</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>121</td>
<td>6985.4754</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7 reveals that the posttest mean scores in general chemistry of high test anxiety students in the control and experimental groups were 34.67 and 27.20 respectively. From Table 8, it can be seen that the high anxiety students in the treatment group achieved significantly higher in general chemistry than did high anxiety students in the control group. Based on these findings, hypothesis 2 was rejected at 0.05 level of significance.
Hypothesis 3
There is no significant difference in achievement in General Chemistry between low test anxiety student in the control and experimental groups.
Hypothesis 3 was tested using a one-way analysis of covariance. Table 9 shows the mean, and standard deviation of students score in General Chemistry achievement test of low test anxiety students in the control and experimental groups. Table 10 shows the result of the one-way analysis of covariance with pretest achievement in chemistry scores serving as the covariates.

TABLE 9  Mean and standard deviations of chemistry scores of low test anxiety students in the control and experimental groups (N=156 students)

<table>
<thead>
<tr>
<th>Low test anxiety</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Control Group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Score</td>
<td>12.33</td>
<td>4.87</td>
</tr>
<tr>
<td>Post Test Score</td>
<td>28.88</td>
<td>4.52</td>
</tr>
<tr>
<td><strong>Experimental Group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Score</td>
<td>11.91</td>
<td>4.98</td>
</tr>
<tr>
<td>Post Test Score</td>
<td>36.23</td>
<td>7.27</td>
</tr>
</tbody>
</table>

TABLE 10 One-way analysis of covariance of low test anxiety students in the control and experimental groups.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>Source of squares</th>
<th>Mean square</th>
<th>F-ratio</th>
<th>Sign level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates</td>
<td>1</td>
<td>48.59985</td>
<td>448.59985</td>
<td>13.441</td>
<td>0.0003</td>
</tr>
<tr>
<td>Low anxiety</td>
<td>1</td>
<td>2193.5552</td>
<td>2193.5552</td>
<td>65.725</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>1</td>
<td>5106.3257</td>
<td>33.37468</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>155</td>
<td>7748.4808</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9 reveals that the posttest mean scores in general chemistry of students in the control and experimental groups were 36.23 and 28.88 respectively. From Table 10, it can be seen that the low test anxiety students in the treatment group achieved significantly higher in general chemistry than the low test anxiety students in the control group. Based on these findings hypothesis 3 was rejected at 0.05 level of significance.
Hypothesis 4
There is no significant difference in achievement in General Chemistry between positive attitude toward science students in the control and experimental groups. Hypothesis 4 was tested using a one-way analysis of covariance. Table 11 shows the mean and standard deviation of students scores in General Chemistry achievement test of positive attitude toward science students in the control and experimental groups. Table 12 shows the result on the one-way analysis of covariance with pretest achievement in chemistry scores serving as covariates.

Table 11 Mean and standard deviation of chemistry score of positive attitude toward science students in the control and experimental groups (N=166 students).

<table>
<thead>
<tr>
<th>Positive Attitude</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Score</td>
<td>12.52</td>
<td>4.65</td>
</tr>
<tr>
<td>Post Test Score</td>
<td>28.01</td>
<td>4.57</td>
</tr>
<tr>
<td>Experimental Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Score</td>
<td>12.14</td>
<td>4.79</td>
</tr>
<tr>
<td>Post Test Score</td>
<td>35.59</td>
<td>7.37</td>
</tr>
</tbody>
</table>

TABLE 12 One-way analysis of covariance of positive attitude students in the control and experimental groups.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>Source of squares</th>
<th>Mean square</th>
<th>F-ratio</th>
<th>Sign level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates</td>
<td>1</td>
<td>430.61581</td>
<td>430.6158</td>
<td>12.137</td>
<td>0.0006</td>
</tr>
<tr>
<td>Positive</td>
<td>1</td>
<td>2466.3815</td>
<td>2466.3815</td>
<td>69.517</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>163</td>
<td>5783.0508</td>
<td>35.47884</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>165</td>
<td>8680.0482</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 11 shows that the posttest mean scores in general chemistry of students with positive attitude toward science in the control and experimental groups were 35.59 and 28.01 respectively. From Table 12, it can be seen that students with positive attitude toward science in the experimental group achieved significantly higher in general chemistry than students with positive attitude toward science in the control group. Based on these findings, hypothesis 4 was rejected at 0.05 level of significance.
Hypothesis 5
There is no significant difference in achievement in General Chemistry between negative attitude toward science students in the control and experimental groups.
Hypothesis 5 was tested using a one-way analysis of covariance. Table 13 shows the mean and standard deviation of students scores in General Chemistry achievement test of negative and attitude toward science students in the control and experimental groups. Table 14 shows the result of the one-way analysis of covariance with pretest achievement in chemistry scores serving as the covariates.

TABLE 13 Mean and standard deviation of chemistry score of negative attitude toward science students in the control and experimental groups (112 students)

<table>
<thead>
<tr>
<th>Negative Attitude</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Control Group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Score</td>
<td>11.36</td>
<td>4.55</td>
</tr>
<tr>
<td>Post Test Score</td>
<td>28.32</td>
<td>8.08</td>
</tr>
<tr>
<td><strong>Experimental Group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Score</td>
<td>11.61</td>
<td>4.23</td>
</tr>
<tr>
<td>Post Test Score</td>
<td>35.23</td>
<td>4.87</td>
</tr>
</tbody>
</table>

TABLE 14 One-way analysis of covariance of negative attitude students in the control and experimental groups.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>Source of squares</th>
<th>Mean square</th>
<th>F-ratio</th>
<th>Sign level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates</td>
<td>1</td>
<td>1153.2175</td>
<td>1153.2175</td>
<td>32.973</td>
<td>0.0006</td>
</tr>
<tr>
<td>Negative attitude</td>
<td>1</td>
<td>1267.9274</td>
<td>1267.9274</td>
<td>36.252</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>109</td>
<td>3812.2748</td>
<td>34.9749</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>111</td>
<td>6233.4196</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 13 reveals that the posttest mean score in general chemistry of students with negative attitude toward science in the control and experimental groups were 35.23 and 28.32 respectively. From Table 14, it can be seen that students with negative attitude toward science in the experimental group achieved significantly higher in general chemistry than student with negative attitude toward science in the control group. Based on these findings, hypothesis 5 was rejected at 0.05 level of significance.
Hypothesis 6
There is no significant difference in achievement between high and low anxiety students irrespective of the control and experimental groups.
Hypothesis 6 was tested using a three-way analysis of covariance. Table 15 show the mean and standard deviation of high and low anxiety students irrespective of the control and experimental groups. Table 6 shows the result of the three-way analysis of covariance with pretest achievement in chemistry scores serving as covariates.

**TABLE 15** Mean and standard deviation of high and low test anxiety students irrespective of the control and experimental groups (N=278 students)

<table>
<thead>
<tr>
<th>Groupings</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High Anxiety</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Score</td>
<td>11.81</td>
<td>4.16</td>
</tr>
<tr>
<td>Post Test Score</td>
<td>30.93</td>
<td>7.60</td>
</tr>
<tr>
<td><strong>Low Anxiety</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Score</td>
<td>12.12</td>
<td>4.91</td>
</tr>
<tr>
<td>Post Test Score</td>
<td>32.56</td>
<td>6.95</td>
</tr>
</tbody>
</table>

Table 15 reveals that the posttest mean scores in general chemistry of the high test anxiety and low anxiety test students, were 30.93 and 32.56 respectively. From Table 6, it can be seen that the low test anxiety students achieved significantly higher in general chemistry than high test anxiety students. Based on this findings, hypothesis 6 was rejected at 0.05 level of significance.

**TABLE 16** Mean and standard deviation of positive and negative attitude toward science students irrespective of the control and experimental groups (N=278 students)

<table>
<thead>
<tr>
<th>Groupings</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive Group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Score</td>
<td>12.32</td>
<td>4.71</td>
</tr>
<tr>
<td>Post Test Score</td>
<td>31.89</td>
<td>7.25</td>
</tr>
<tr>
<td><strong>Negative Group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Score</td>
<td>11.48</td>
<td>4.37</td>
</tr>
<tr>
<td>Post Test Score</td>
<td>31.78</td>
<td>7.49</td>
</tr>
</tbody>
</table>
Table 16 shows that the posttest mean scores in general chemistry of positive attitude toward science and negative attitude toward students were 31.89 and 31.78 respectively. From Table 16, it can be seen that students with positive attitude toward science did not achieve significantly higher in general chemistry than negative attitude toward science students. Based on these findings, hypothesis 7 was accepted at 0.05 level of significance.

Hypothesis 8

Table 17 shows the mean and standard deviation of test anxiety score of students in the control and experimental groups. Table 18 shows the result of the one-way analysis of covariance with pretest in test anxiety score serving as covariates.

TABLE 17 Mean and standard deviations of the test anxiety score of students in the control and experimental groups (N=278 students)

<table>
<thead>
<tr>
<th>Groupings</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Group</td>
<td>51.34</td>
<td>5.49</td>
</tr>
<tr>
<td>Pretest Score</td>
<td>51.44</td>
<td>6.81</td>
</tr>
<tr>
<td>Post Test Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative Group</td>
<td>51.51</td>
<td>5.32</td>
</tr>
<tr>
<td>Pretest Score</td>
<td>45.53</td>
<td>5.75</td>
</tr>
<tr>
<td>Post Test Score</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE 18 One-way analysis of covariance of test anxiety score between the control and experimental groups.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>Source of squares</th>
<th>Mean square</th>
<th>F-ratio</th>
<th>Sign level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates</td>
<td>1</td>
<td>1058.810</td>
<td>10518.810</td>
<td>429.178</td>
<td>0.0006</td>
</tr>
<tr>
<td>Treatment</td>
<td>1</td>
<td>2412.5960</td>
<td>2412.5960</td>
<td>98.436</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>275</td>
<td>6740.0368</td>
<td>24.50923</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>277</td>
<td>19671.442</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 17 shows that the posttest mean scores of students test anxiety in the control and experimental groups were 45.53 and 51.44 respectively. From Table 18, it can be seen that students in the experimental group did score significantly lower in test anxiety than students in the control group. Based on these findings, hypothesis 8 was rejected at 0.05 level of significance.

**Hypothesis 9**

There is no significant difference in mean attitude toward science score between the control and experimental groups.

Hypothesis 9 was tested using a one-way analysis of covariance. Table 19 shows the mean and standard deviation of attitude toward science score of the control and experimental groups. Table 20 shows the result of one-way analysis of covariance of attitude toward science score between experimental and control groups with pretest in attitude toward science serving as covariates.

**TABLE 19** Mean and standard deviation of attitude toward science score between the control and experimental groups (N=278 students).

<table>
<thead>
<tr>
<th>Groupings</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive Group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Score</td>
<td>193.22</td>
<td>24.87</td>
</tr>
<tr>
<td>Post Test Score</td>
<td>184.60</td>
<td>23.07</td>
</tr>
<tr>
<td><strong>Negative Group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Score</td>
<td>194.31</td>
<td>25.98</td>
</tr>
<tr>
<td>Post Test Score</td>
<td>198.19</td>
<td>26.98</td>
</tr>
</tbody>
</table>

**TABLE 20** One-way analysis of covariances attitude toward science between the control and experimental groups.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>Source of squares</th>
<th>Mean square</th>
<th>F-ratio</th>
<th>Sign level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates</td>
<td>1</td>
<td>92424.266</td>
<td>92424.266</td>
<td>8.619</td>
<td>0.0006</td>
</tr>
<tr>
<td>Treatment</td>
<td>1</td>
<td>11430.885</td>
<td>11430.885</td>
<td>1.066</td>
<td>0.3028</td>
</tr>
<tr>
<td>Residual</td>
<td>275</td>
<td>294831.3</td>
<td>10723.023</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>277</td>
<td>3052685.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 19 reveals that posttest mean scores of students attitude toward science in the control and experimental groups were 198.19 and 184.60 respectively. From Table 20 it can be seen that students in the experimental group did not score significantly higher in attitude toward science than students in the control group. Based on these findings, hypothesis 9 is accepted at 0.05 level of significance.

7.17 Summary of the Findings

In summary, the findings of these analysis were as follows:

1. The experimental group who received frequent testing were found to perform significantly higher in chemistry achievement than did the control group who did not receive frequent testing.
2. The high test anxiety students in the experimental group achieved significantly higher in chemistry than the test anxiety students in the control group.
3. The low test anxiety students in the experimental group achieved significantly higher in chemistry than the low test anxiety students in the control group.
4. Students with positive attitude toward science in the experimental group achieved significantly higher in chemistry than students with positive attitude toward science in the control group.
5. Students with negative attitude toward science in experimental group achieved significantly higher in chemistry than students with negative attitude toward science in the control group.
6. There was a significant difference in performance between the high and low test anxiety students. The low test anxiety students performed significantly higher in chemistry achievement than the high test anxiety students.
7. There was no significant difference in performance between students with positive and those with negative attitude toward science.
8. There was significant difference in mean test anxiety score between the experimental and control groups. The experimental group had a posttest mean test anxiety score that was lower than the mean test anxiety score of students in the control group.
9. There was no significant difference in students mean attitude toward science scores between the experimental and control groups. The experimental group had a mean attitude toward science score that was slightly higher than the mean attitude toward science score.
that was slightly higher than the mean attitude toward science score of the control group but it was not significantly different.

7.18 CONCLUSIONS

i. Chemistry Achievement
From the findings it was concluded that frequent testing does significantly enhance achievement in chemistry, as compared with conventional testing. The study also showed that, without taking treatment into consideration, the students with lower test anxiety achieved significantly higher in chemistry than those with high test anxiety. Similarly, without taking treatment into consideration, the study showed that student with positive attitudes toward science did not perform significantly higher than students with negative attitudes toward science. The study also showed that frequent testing coupled with lower anxiety or positive attitudes resulted in significantly higher achievement in chemistry. In other words, frequent testing resulted in significant gains in achievement in chemistry. Further, having a good attitude toward science was not enough for a significant increase achievement in chemistry.

ii. Test Anxiety
From the findings on test anxiety it was concluded that students exposed to frequent testing were able to significantly reduce their anxiety as compared to students under conventional testing. Since students with lower anxiety achieved significantly higher in chemistry than students with high anxiety, it can be concluded that frequent testing promotes achievement indirectly by lowering anxiety.

iii. Attitudes toward Science
Results showed that frequent testing does increase chemistry students’ attitudes toward science, but the increase was not significant. Students with positive attitudes toward science tended to perform better than students with negative attitudes toward science, although the differences were not significant. Exposing the students to frequent testing should improve their performance. The posttest scores also revealed that students in the control group seemed to have their attitudes toward science lowered at the end of the semester.
The conclusions from the findings are limited to the University Technology of Malaysia and the similar groups that may exist.

7.19 Discussion of the Findings

i. Chemistry Achievement

Frequent testing does significantly enhance achievement in chemistry, as compared with conventional testing. This finding reaffirmed earlier studies (Fitch, 1951; Kulp 1934; Marso 1970; Pikumas 1965; Sumprer, 1982). Frequent testing may enhance achievement for the following reasons:

1. Frequent testing can help the students indirectly see the structure of the course (Standalee, 1960).
2. The instructor is able to continually analyze restructure when necessary, and improve his or her teaching methods within a given unit of study (Stokes, 1973).
3. Smaller amounts of materials have to be studied for the frequent testing as compared with conventional testing.

The study also showed that the students with lower test anxiety achieved significantly higher in chemistry than those with high test anxiety. The findings also reaffirmed the earlier findings by Deffenbacher (1977), Doctor and Altman (1969), Liebert and Morris (1967), Sarason (1972), and Stoops (1978). There are several possible reasons as to why high anxious persons do not perform well in examinations. Two of these follow:

1. High anxious persons, under stress, experience cognitive interference and preoccupation that make time pass slowly and results in poor performance (Stoops, 1978).

2. Worry, as a component of test anxiety, has been found to be significantly, negatively related to both performance expectancy and examination performance.

In this study, student with positive attitudes toward science did not perform significantly better than students with negative attitudes toward science. Several studies in the past showed the same result (Baker, 1985; Kahle, 1982; Wilson 1983). Attitude toward science was not a strong factor causing higher or lower achievement in science. (Al-Shargi, 1987).
ii. **Test Anxiety**
In this study students exposed to frequent testing were able to significantly reduce their test anxiety as compared with students under conventional testing. A possible explanation for this is that frequent testing may help to divide the course or task at hand, thus lessening the problem of study and stress.

iii. **Attitudes toward Science**
This study showed that frequent testing did increase chemistry students’ attitudes toward science, although the increase was not significant. A possible reason for this improvement is that, as frequent testing promotes higher achievement, achievement acts to improve the attitudes (Wilson, 1983).

The study also showed that the control group students seemed to have their attitudes toward science lowered at the end of the semester. As these students were not exposed to any form of treatment, other factors could have been working to lower their attitude toward science. These factors could be the instructors, school system, curriculum, or environmental. This finding raises questions for future research.
CHAPTER 8

8.0 Critical Issues in Assessment
Some of the critical issues that transpire from assessment in our school system are as follows:

i. Learning, assessment achievement in school system
ii. Anxiety in examination
iii. Stress to the students and teachers
iv. Unfairness in assessment
v. Little free time

i. Learning, Assessment and Achievement in School System
The correct model in our school system should be:

Learning → Assessment → Achievement

Does our school system follow this model? The answers is no, then what model do we follow? Due to exam-oriented curriculum, our school system whether we like it or not, does not follow the above model. Sad to say, we are following this model:

Assessment → Achievement → Learning

In other words, we assess, the students first then find their achievement from there, then we conclude whether the students are learning or not.

ii. Anxiety in Examination (Assessment)
Due to the exam-oriented system, the students are subjected to the high anxiety. Normally, in announcing or declaring the results of UPSR, PMR and SPM, we only take care the successful ones but not the failures. We praise the successful ones but ignore the failures. We do not even look why they failed. Do we give them fair assessment? Well, they are many unanswered questions.

iii. Stress to the Students and Teachers
Normally, during the exam year for UPSR, PMR and SPM, students and teachers are put the stressful situation. For students, they have little free time except study, drill and practice to prepare for the upcoming examination. Teachers are stressed in preparing their students for the examination. A lot of preparation and work. According to study conducted by Dr. Salwa Mohamed (daughter of Faiza Rice) in 1996, 80% of teachers in Johor were under stress, some have even consulted doctor/psychiatrist to help them overcoming the stressful situation.
iv. **Unfairness in Assessment**

The assessment (exam) in our school system is subjecting the students to an unfair treatment. In SPM assessment, 2 years of work is being tested in 2½ hours of exam or assessment. The question is, is it fair to the students? In PMR, 3 years of work is being tested or assessed in 2½ hours of examination; is it fair? In UPSR, almost 6 years of work or learning is being tested in 1½ hours of objective questions.

v. **Little Free Time**

During the exam year in Form 5, Form 3 and Standard 6, students have little free time except thinking or studying for the examination. That is why, some students say Form 4 is honeymoon year because no examination. The terminology of honeymoon year should not be introduced or used. Students should study and play all year around. Learning, playing or indulged in co-curriculum should be part and parcel of the school system.

8.1 **Transformation to the School-based Assessment**

Do we have solutions to the above problems? Yes, we do, by changing the assessment system to the school-based assessment. Even though there are many positive sides in the school-based assessment but the government is not able to change fully to the school-based assessment.

The main reason is the society is so engrossed to the old system that they are skeptical to the new system. As a result, only partial assessment implementation of the school-based is being introduced. UPSR – full school-based. PMR – partial and SPM is still with the old system. School-based assessment is planned, formed, administered, examined, recorded and reported by the school teachers involving students, parents and external organization. School-based assessment is conducted in two forms: formative assessment during teaching and learning and summative assessment at the end of the semester or term.

8.1.1 **Some of the objectives of school-based assessment**:

1. Lessening the emphasis of public assessment
2. Continuous assessment
3. Improve the students’ learning
4. More holistic assessment
5. An effective way tackling human capital
6. Value added improvement to the present system of assessment in UPSR and PMR.
8.1.2 Some of the positive functions of the school-based assessment:
1. Carrying out diagnostic test to the individual and finding prescriptions for the remedial.
2. Supervising the development and growth of the students.
3. Carrying out the evaluation that suite to the local situation.
4. Making learning more enjoyable and meaningful.
5. Clarifying the achievement/grade.
6. Preparation of the feedback on the students’ ability, proceed to the higher step, grading and other feedback.

8.1.3 Some of the positive characteristics of school-based assessment:
   i. Holistic
   ii. Integrated
   iii. Fairness
   iv. Flexible
   v. Based on standard
   vi. Part of the learning and teaching

8.2 Test Anxiety and Achievement
8.2.1 Test Anxiety: What is test anxiety?
Too much anxiety about a test is commonly referred to as test anxiety. It is perfectly natural to feel some anxiety when preparing for and taking a test. In fact, a little anxiety can jump start your studying and keep you motivated. However, too much anxiety can interfere with your studying. You may have difficulty learning and remembering what you need to know for the test. Further, too much anxiety may block your performance during the test. You may have difficulty demonstrating what you know during the test.

8.2.2 How do I know if I have test anxiety?
You probably have test anxiety if you answer YES to four or more of the following:
1. I have a hard time getting started studying for a test.
2. When studying for a test, I find many things that distract me.
3. I expect to do poorly on a test no matter how much or how hard I study.
4. When taking a test, I experience physical discomfort such as sweaty palms, and upset stomach, a headache, difficulty breathing, and tension in my muscles.
5. When taking a test, I find it difficult to understand the directions and questions.
6. When taking a test, I have difficulty organizing my thoughts.
7. When taking a test, I often “draw a blank”.
8. When taking a test, I find my mind wandering to other things.
9. I usually score lower on a test than I do on assignments and papers.

8.2.3 What causes it?
All anxiety is a reaction to anticipating something stressful. Like other anxiety reactions, test anxiety affects the body and the mind. When you’re under stress, your body releases the hormone adrenaline, which prepares it for danger (you may hear this referred to as the ‘fight or flight’ reaction). That’s what causes the physical symptoms, such as sweating, a pounding heart, and rapid breathing. These sensations might be mild or intense.

Focusing on the bad things that could happen also fuels test anxiety. For example, someone worrying about doing poorly might think thoughts like, “What if I forget everything I know?” or “What if the test is too hard?”. Too many thoughts like these leave no mental space for thinking about the test questions. People with test anxiety can also feel stressed out by their physical reaction and think things like “What if I throw up” or “Oh no, my hands are shaking”.

Just like other types of anxiety, test anxiety can create a vicious circle: the more a person focuses on the bad things that could happen, the stronger the feeling of anxiety becomes. This makes the person feel worse and, because his or her head is full of distracting thoughts and fears, it can increase the possibility that the person will do worse on the test.

8.2.4 Who’s like to have test anxiety?
People who worry a lot or who are perfectionists are more likely to have trouble with test anxiety. People with these traits sometimes find it hard to accept mistakes they might make or to get anything less than a perfect score. In this way, even without meaning to, they might really pressure themselves. Test anxiety is bound to thrive in a situation like this.

Students who aren’t prepared for tests but who care about doing well are also likely to experience test anxiety. If you know you’re not prepared, it’s a no-brainer to realize that you’ll be worried about doing poorly. People can feel unprepared for tests for several
reasons: they may not have studied enough, they may find the material difficult, or perhaps they feel tired because didn’t get enough sleep the night before.

8.2.5 What you can do?
Test anxiety can be real problem when someone is so stressed out over a test that he or she can’t get past the nervousness to focus on the test questions and do his or her best work. Feeling ready to meet the challenge, though, can keep test anxiety at a manageable level.
If you are adequately prepared but still overreact and panic, you are suffering from test anxiety. We will review the steps that will help you to overcome its effects. These steps are:
i. Before the test
ii. Thinking straight about your test
iii. Taking care of your basic needs
iv. Getting ready
v. Facing the test
vi. During the test
vii. After the test

i. BEFORE THE TEST

• PREPARATION. Preparation is a key element for reducing anxiety. The higher your level of preparation, the lower your level of anxiety. Moreover, getting ready for your test increases your self-confidence.
• NO CRAMMING. Avoid cramming for a test. This is an ineffective way of studying. If you cram the night before you might be able to pass some parts of your test, but you will remember nothing afterwards (and in most cases that information will be included in your final). Trying to learn weeks worth of material the day before the test does not work either. Usually this is not a good time to learn much because you feel anxious. You feel pressured, and probably guilty, for studying at the last minute, therefore you cannot concentrate very well. Please do not tell me that when you study ahead of time you do worse than when you study they day before. Years of research on how to study are against you! If this happens to you it is because you are either, studying in advance
without learning, or you have developed the negative habit of learning under pressure. Both are ineffective ways of learning and both can easily create anxiety.

- **REVIEW ALL THE INFORMATION.** Study from your book notebook, and any other materials used class. Combine their information. Work on mastering the main, as well as specific concepts presented in your class.

- **ASK YOURSELF QUESTIONS.** This method is well explained in the Study Skills screen of the Counseling Centre program. When studying, transform the headings into question, and answer them using the different sources of information used in class. Ask yourself what kind of questions your instructor may ask you. Try to answer them too. Moreover, ask your instructor for samples of previous test, and practice with them.

- **USE FLASH CARDS.** Yes, you can use this type of help to organize your study. This kind of help will allow you to allocate your time in an effective way. You will be able to determine what you already know, and spend more time reviewing those materials that need more studying.

### ii. THINKING STRAIGHT ABOUT YOUR TEST

**CHANGE THE WAY YOU THINK ABOUT STUDYING.** Changing the way you think about studying can improve your performance. Studying-and grades are not a measure of your self-worth. You may be investing too much of your personal definition on studying as an insurmountable task. These kinds of beliefs are very effective in creating anxiety and stress. And these reactions can reduce, in turn, your capacity to concentrate, and learn. Confirming that studying is an impossible task for you! *(Does the concept “vicious circle” ring a bell?)*

- **PUT YOUR TEST IN PERSPECTIVE.** A test is only a test. Keep in mind that there will be others. This will help you remove part of the emotional charge we put on our tests, reducing your stress, and allowing you to study better.

- **ELIMINATE NEGATIVE SELF-TALK.** Avoid thinking of yourself in a negative way. Avoid getting entangled in negative aspects related with studying. Focus on what needs to be done and do it. You will be surprised how much time student spend doing everything else but studying. And negative thoughts are an example of “everything else”.

- **INVEST IMT IN PLANNING.** Plan ways to improve your studying. Evaluate your plan accordingly with your performance. Plan ways to keep what you did right; and plan ways to improve what can be improved.
• **PUT YOUR GRADE IN PERSPECTIVE.** Your grade is not necessarily a reflection of your preparation. Most of my students believe that the success of a test anxiety reduction programs should be measured by the grades obtained. The reality is that your grades will not improve immediately. It will take time and more than one test to see that kind of results. Therefore, your performance should be evaluated against what you did. If you had a good plan, and you stick to it, that is what really counts; even if the grade was not as high as you would have liked it to be. You might have improved significantly, but the test may have been more difficult than expected. The reverse is also possible, you may have failed your plan and still get a good grade. (e.g. the test included those questions you knew all about). Again, you should use more than your actual grade to evaluate your performance.

• **DEVELOP REASONABLE EXPECTATIONS.** Take your test one at a time. Set realistic goals. Show as much as you know as you can. Hope for a result that matches the stage of development you have reached at this point. Unrealistic expectations will only lead to frustration, which, in turn, will become a good excuse to give up.

iii. **TAKING CARE OF YOUR BASIC NEEDS**

• **KEEP IN MIND THAT YOU ARE MORE THAN A TEST TAKER.** Students concerned about tests usually neglect other aspects of themselves. Do not forget that taking a test is only one of the important things in your life. You should also care for your biological, emotional, psychological, and social needs.

• **“MENS SANA IN CORPORE SANA”**. “Healthy mind in healthy body”. Exercise. Stay in good shape. Eat consciously. Keep up with your recreational and social activities. All contribute to your well-being and capacity to buffer test anxiety.

• **REMEMBER THAT “FOOD FOR THOUGHT” IS ONLY A LITERARY EXPRESSION.** More often than not you will see students abusing food) e.g. cookies or beverages (e.g. coffee). It looks as if they believe that this will help them in studying or taking tests. In fact, the result is often the opposite. A stomach ache will keep you from concentrating. Caffeine may give you the jitters. You may end up feeling light-headed.

• **DOSIFY YOUR STUDY.** Study for short periods of time (see the Study skills screen for more information). Follow a moderate pace. Do not forget that your mind can take more of the same forever. Take breaks.
• **REST THE NIGHT BEFORE THE TEST.** Distract your mind with activities other than studying. Rest. Get plenty of sleep. A refreshed mind will allow you to do your best. An overly tired mind will function at its best. (This is the reason why studying overnight usually does not pay off).

• **DO NOT ABUSE YOURSELF.** Once you feel you know what you need to know, quit studying, and do something relaxing. The only reason why keep you keep studying way after you are reasonably prepared is your lack of confidence. Be patient. Learning when to stop takes time. Accomplishing it boosts your sense of self-confidence and self esteem.

**iv. GETTING READY**

• **FACE THE DAY OF THE TEST WITH PRIDE.** Take responsibility for your actions. If you studied enough, be proud of yourself. What really matters at this point is not the potential grade, but the fact that you did what you were supposed to do. This is an accomplishment in itself.

• **EAT A SENSIBLE BREAKFAST.** Do not abuse food before the test. Some students use food as a way to reduce anxiety. Indulging in food on the day of the test may backfire on you, impairing your performance by making you feel physically uncomfortable (or sick) during the test.

• **RELAX DURING THE HOUR BEFORE THE TEST.** Do something relaxing the hour before the test. It is too late to try to learn what you did not learn before. Last minute cramming will cloud what you have learned before. It will also undermine your confidence.

• **ARRIVE AT THE CLASSROOM EARLY, BUT DO NOT STAY THERE.** Arrive at the classroom early if you want to select a good seat (e.g. a seat away from distractions). Then, go out of the room and use the remaining time to walk and relax.

• **AVOID “STRESS-CARRIERS”**. Politely avoid classmates who produce anxiety and affect your disposition to the test. Do not let them scare, stress or upset you.

• **BRING A “STRESS-SAVER” WITH YOU.** Bring a magazine or newspaper to read if waiting for the test stresses you.

• **USE PHYSICAL RELAXATION.** Learn and use tensing and relaxing techniques to fight off the tension and anxiety.
v. FACING THE TEST

- **CHECK YOUR INTERNAL STATE.** How are you? How is your anxiety level? If it is high or moderately high, take some time to relax. Even though it takes time away from your test, relaxing increases your chances to do a more efficient job, saving your time instead.

- **COACH YOURSELF.** Sometimes students get anxious after finding out that they do not know the answer to the first or second question. Tell yourself that you are going to do your best. Tell yourself that you are going to answer the questions you know first, then the questions you are not really sure about, and, finally, the questions you do not know. Follow your plan!

- **REVIEW YOUR TEST.** Before you begin answering the questions, review the entire test working on the easiest questions first.

vi. DURING THE TEST

- **OUTLINE ANSWERS ON ESSAY QUESTIONS.** Develop a short outline of your answers for essay questions. This will help you to organize your answer, avoid irritating repetitions, and skip circular arguments.

- **GIVE SHORT ANSWERS FOR SHORT-ANSWER QUESTIONS.** Answer short and to the point. Use specific terms and ideas. If you cannot remember a technical term, describe it in your own words.

- **READ OPTIONS CAREFULLY.** Read all the options of multiple choice questions. Eliminate the most obvious. Use qualifying words such as “always” or “only” to eliminate others. If unsure, rely on your first hunch, then mark the question with an asterisk or a star and move on. If you have time at the end, go back and review your marked questions.

- **WEAR YOUR WATCH.** Do not rush through the test. Keep track of the time. Pace yourself. If you are running out of time, concentrate on those questions which you can answer. Make sure you match the number of the questions with the number of your answer on the Scantron.

- **DO NOT GET STUCK.** Do not get stuck on one question. Skip it and solve the next one. Go back to the question after you finish answering those you can. Remember that you do not get points for trying. *RELAX YOUR TENSION*. If your tension is hampering your capacity to do your best, tense and relax your body and needed during the test. This exercise releases your tension. Breathing deeply, in and out, also helps to...
release anxiety. For more information on relaxation, see the Relaxation, Breathing Techniques and Biofeedback screens in this program.

- **ASK QUESTIONS.** Ask for more information if you are not sure about a question in your test. Asking your instructor a question can also help to distract you and reduce your anxiety. *TALK TO YOURSELF*. If your anxiety continues, tell yourself phrases like "I can be anxious later, now I am going to continue my test". Use any type of internal dialogue (nobody else need to hear you) that can help you do better in your test.

- **IF WORSE COME TO WORSE, USE ANY OTHER LEGAL TRICK TO DISTRACT YOURSELF.** If anxiety continues, use any acceptable way to distract yourself from it. Request permission to go to the bathroom or get a drink, etc. If nothing else works, sharpen your pencil.

vii. **AFTER THE TEST**

- **REWARD YOURSELF.** Whether you did well or not, reward yourself for taking, and surviving your test! You deserve it.

- **LATER ON.** Evaluate your study plan. Were you prepared for it? Were you able to control your anxiety and relax. Find out what you did right and repeat it the next time. Find out what needs more work. Do not dwell on your mistakes. You are supposed to make some. Use them as a guide for what needs to be improved, and work on improving them.

- **THEN..** Develop an improved plan and begin a studying for your next test!

**CONCLUSIONS**

We have been exposed to general ideas of assessment and the delabilating effect of test anxiety and stress.

So, we as teachers/lecturers should be responsible enough when we intend to assess our students. As parents, we should not put too much stress to our kids and try to cooperate with teachers on many aspects of their education.
BIBLIOGRAPHY


CURRICULUM VITAE

Professor Dr. Sulaiman bin Yamin was born in Sungai Kapal, Pengerang, Johor. He attended his primary education at Sek. Keb. Tambatan Rendah and Temenggong Abdul Rahman School (an English medium school). He finished his secondary education at Maktab Sultan Abu Bakar (formerly known as English College, Johor Bahru).

He enrolled his first degree at Universiti Kebangsaan Malaysia in 1974 to do chemistry with Education. Upon graduation with chemistry degree (Hons.) and diploma education in 1978, he was appointed as an officer at Curriculum Development Center, Ministry of Education. He stayed for a year at curriculum development center before joining Universiti Teknologi Malaysia as an assistant lecturer.

He went to pursue his Master Degree in Chemistry Education at Oregon State University, USA in September 1980 and graduated in July 1981. He continued working at UTM until 1984 where he won scholarship to do PhD in Science Education at the same university (Oregon State University, USA). He graduated in 1988 with PhD in science education.

While in UTM, he lectured in chemistry, chemistry education subjects, statistics or research and methodology of research. He is considered as an expert in Data Analysis for Science Social. He has graduated; under him, several doctors and numerous masters. He was appointed as external examiners in PhD from USM and UKM for several years.

He was responsible for the expansion of post graduate program in Faculty of Education in UTM. He was promoted to Associate Professor in 1991 (UTM) and later professor in 2002 (UTM). During his tenure with UTM, he has produced modules and books on statistics for research, methodology for research and science education. He held several posts in UTM - namely Deputy Dean, head of departments and senate member. He served as PALAPES staff and holding a rank of Captain from 1991 to 1998.

Professor Dr. Sulaiman bin Yamin joined UTHM as contract professor in July 2007 and holds a post of Deputy Dean for post graduate and research. At present, he has 7 PhD students and 4 master students. He has two grants namely incentive grant for post graduate and ERGS grants. He is also appointed by MQA to be an evaluator for private colleges and universities in COPA and COPIA.