

---

**Institute of Education and Behavioral Sciences  
Department of Psychology**

**Educational Measurement and Evaluation  
(Psyc 3097)**

**APRIL, 2016**

---

# UNIT 1: The Concepts of Measurement and Evaluation

## 1.1. DEFINITIONS OF TERMS

In measurement and evaluation, there are common terms that are used similarly or differently by different scholars in the field. These terms include test, measurement, assessment, and evaluation.

In the following section, we will see the definitions of these terms.

### ***Test***

Say a word “test,” and what most persons think of will probably be school examinations, college entrance examinations, or employment tests involving writing or marking answers. This constitutes a narrow view of the world of testing. The definition given below is broader and more inclusive. *An educational and psychological test is defined as a systematic procedure for observing (i.e., getting information) and describing one or more characteristics of a person with the aid of either a numerical scale (measurement such as test scores) or a category (qualitative means).*

Common terms in measurement and evaluation are

- Test
- Measurement
- Assessment
- Evaluation

#### a) Test

A test is a task or a series of tasks or questions that students must answer. It is used to get information regarding the extent to which the students have mastered the subject matter taught and the attainment of instructional objectives.

A test is a particular form of measurement. It is a formal, systematic procedure to gather information about student's behavior or performance.

#### ***b) Measurement***

*Measurement* is concerned with systematic collection, quantification, and ordering of information. It is a process of quantifying or assigning a number to performance according to explicit rules. In other words, we assign numbers to any behavior, characteristics, property or attribute based on agreed upon rules. For instance, if a person tells you the size of a table he measure is 45, what do you understand by this number? What could be your next question to the person? If a student said I score 25, what further information would you need? In both case you need the units. In the first instance, 45 what? Centimeters, meters, millimeters, inches,...? What is the unit? In the latter case, 25 out of what maximum score? Out of 100,

50, 25? Thus assigning number to the length of the table (i.e., the attribute of the table) and to the behavior (i.e., performance) of the student is not enough. We need a rule (for example, the units).

Measurement can take many forms, ranging from the application very elaborate and complex electronic devices, to paper – and- pencil exams, to rating scales or checklists.

### *c) Assessment*

Assessment is a general term that includes all the different ways teachers gather information in their classroom. It includes observations, oral questions, paper-and-pencil tests, homework, laboratory work, research paper, and the like. It is a process of collecting, synthesizing, and interpreting information to aid in decision-making (Nitko, 1996; and Airasian, 1996). Assessment is concerned with the totality of the educational setting and is the more inclusive term, that is, it subsumes measurement and evaluation. Assessment focuses not only on the nature of the learner, but also on what is to be learned and how (Payne, 1997).

### *d) Evaluation*

Evaluation is the process of making judgment about student's performance, instruction, or classroom climate. It occurs after assessment information has been collected, synthesized and thought about, because this is when the teacher is in a position to make informed judgments (Airasian, 1996).

Evaluation includes both quantitative and qualitative descriptions of student behavior plus value judgment concerning the desirability of that behavior. Measurement is limited to quantitative descriptions of student behavior. It doesn't include qualitative descriptions nor does it imply judgments concerning the worth or value of the behavior measured. The following simple formula shows the relationship between measurement and evaluation.

Evaluation = Quantitative description of students' behavior (measurement) and/or qualitative description of students' behavior (non-measurement) plus value judgments.

Thus, evaluation may or may not be based on measurement (or tests) but when it is, it goes beyond the simple quantitative description of students' behavior. Evaluation, then, involves judgment.

From instructional standpoint, **evaluation** may be defined as a systematic process of determining the extent to which instructional objectives are achieved by the learners. There are two important aspects of this definition. First, evaluation implies a systematic process, which omits causal, uncontrolled observation of students. Second, evaluation always assumes that instructional objectives have been previously identified. Without previously determined objectives, it is almost impossible to judge the nature and extent of students' learning progress.

---

## 1.2 ROLE OF EVALUATION IN TEACHING

### 1. *Instructional Management Decisions*

Instructional management decisions include

- planning instructional activities (deciding what to teach to students),
- placing students into learning sequences,
- monitoring students' progress,
- diagnosing students' learning difficulties,
- motivating students for learning,
- providing students and parents with feedback about achievements,
- evaluating teaching effectiveness, and
- assigning grades to students.

The role of evaluation in the instructional process can be best represented by the following diagram.

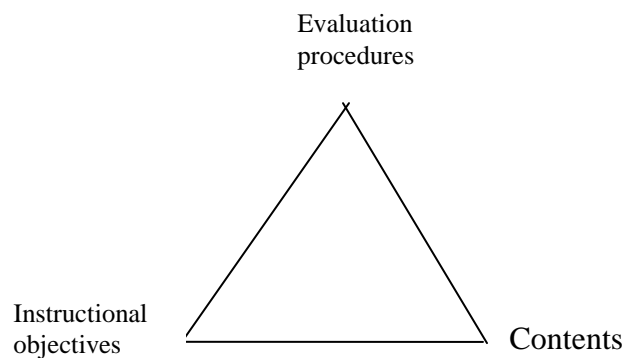


Figure 1.1 A diagram representing the relationship of the evaluation, instructional objectives, and course contents

### 2. *Selection Decisions*

An institution or organization decides that some persons are acceptable while others are not for certain jobs or vacancies; those not acceptable are rejected and no longer are the concern of the institution, or organization. This rejection and the elimination of those rejected from immediate institutional concern feature is central to a selection decision (Cronbach and Gleser, 1965, cited in Nitko, 1996). An educational institution often uses assessments to provide part of the information on which to base selection decision. For example, college admissions are often selection decisions: some candidates, who fulfill the selection requirements, are admitted but those who do not fulfill the criteria are not; those not admitted are no longer the college's concern.

***Selection decisions*** are decisions made on who will be accepted or rejected by an institution.

When an institution uses an assessment procedure for selection, it is important for it to show that the candidate's results on these assessments reflect a relationship to success in the program or job for which the institution is selecting persons. If the data do not show that those assessment results can distinguish effectively between those candidates likely to succeed, the assessments procedures should be improved or eliminated. Because the data they provide are not valid. Hence, it may be illegal to continue to use assessment results that bear no relationship to success on the job.

### **3. *Placement decision***

Placement decisions differ from selection is that in selection decisions rejection is possible and the institution is not concerned about what happens to those rejected, whereas in placement decisions persons are assigned to different levels of the same general type of instruction, education or work, and no one is rejected (Cronbach, 1990, Cronbach and Gleser, 1965, cited in Nitko 1996). Suppose that a school places students according to their ability level: Section A for gifted (high achieving students), Section B for average students, and Section C for slow learners. Slow learners who cannot be placed (put) in the gifted students section, must be placed at other educational level.

**Placement decisions** are decisions made after applicants are selected. The purpose is to assign them to different levels or types of categories.

That is, students with low reading readiness test scores, for example, cannot be sent home. They must be placed into appropriate educational levels and taught to read. So in placement decision the institution or the school is responsible for all individual learners. It cannot reject them. Placement decisions involve “vertical” grouping within a single job, program or subject, as has been seen above. The subject taught to gifted students might be more complex than that for average learners, and the content to be taught for slow learners could be much simpler than that for gifted students. This increase in the content is an indication of increase in vertical grouping with in the same subject.

### **4. *Classification Decision***

These are decisions that involve the assignment of persons to one of several categories, jobs, or programs that are not necessarily thought of as levels of work instruction. Like placement, classification decisions assume that the individual has been selected. Unlike placement decisions, classification involves “horizontal” grouping in different curricula or jobs.

**Classification decisions** are decisions made to place individuals in an optimal program to increase the probability of success.

---

For example, legislation in the area of educating persons with disabilities has given a legal status to many labels for classifying children with disabilities (i.e., blind, deaf, hard of hearing, speech disorders, etc.) into one (or more) of a few designated categories. These categories are unordered (that is, blindness is not higher or lower than deafness).

### **5. *Counseling and Guidance Decisions***

**Classifications** are different from **selection** and **placement**: classification refers to cases where the **categories are essentially unordered**; placement refers to cases where the categories **represent levels of education**, and selection refers to the case where students are **accepted or rejected**.

Evaluation also serves as a means to make counseling and guidance decisions. Students need to make decisions about their careers. Tests are frequently used to assist in exploring and choosing careers and directing them to prepare for the careers they select. However, it should be noted that a single assessment result is not used for making guidance and counseling decisions. Rather, a series of assessments is administered, including an interview, interest inventory, various aptitude tests, a personality questionnaire, and an achievement battery. Information from these assessments, along with additional background information, is discussed with the student during a series of counseling sessions. This facilitates a student's decision making processes and is an entrée (i.e., initial part) to the exploration of different careers.

### **6. *Credentialing and Certification Decisions***

Credentialing and certification decisions are concerned with assuring that a student has attained certain standards of learning. Student certification decisions (decisions to give on not to give certificates) may focus on whether a student has attained minimum competence or whether a student has attained a high standard performance.

### **7. *Educational Diagnostic and Remedial Decisions***

These decisions are made to identify the strengths and weaknesses of students. Before teachers and counselors can recommend remedial help, they must know in which specific areas an individual is having difficulty. Sometimes the instruction a teacher or school pre-arranged is not effective for an individual student: The student may need special remedial help or a special prescription, relying on alternative methods or materials. That is, when a student has a problem in Mathematics, a teacher may administer a test to identify his weaknesses and to make remedial actions.

**Diagnostic and remedial decisions** are made to determine a person's strengths and weaknesses in order to improve performance or well-being.

## **1.3 TYPES OF EVALUATION**

### **1. *PRELIMINARY EVALUATION***

Preliminary evaluations occur during the first days of school and provide a base for expectation thought of the school year. They are obtained through a teacher's spontaneous informal observations and oral questions and are concerned with student's skills, attitudes, and physical characteristics. These evaluations happen naturally. They are essential to guiding our interactions with others and with students (Oosterhof, 1994). Their functions are related to formative evaluation. The purpose is to determine the entry behavior of students, i.e., to know the knowledge students have about the subjects they are going learn.

### **2. *FORMATIVE EVALUATION***

Formative evaluation occurs during instruction by letting the teacher or evaluator know if students are meeting instructional objectives, if the program is taking place according to the schedule, and if the program might be improved. They establish whether students have achieved sufficient mastery of skills. Formative evaluations are also concerned with students' attitudes. The purpose is to determine what adjustments to instruction should be made.

Formative evaluations are based primarily on continuous informal assessments such as listening to what students say, using oral questions to probe comprehension, and watching student's facial expressions and other behaviors. Formative evaluations also are based on formally developed assessment such as quizzes, seatwork, and homework. They help students learn more efficiently, and improve the teaching learning process by overcoming students and teachers' weaknesses.

### **3. *SUMMATIVE EVALUATIONS***

Summative evaluations occur at the end of instruction, such as at the end of a unit, chapter, or the end of the course. They are used to

- certify student achievement and assign end-of-term grades or marks,
- promote students from one grade level to the next,
- group students into different categories,
- determine whether teaching procedures should be changed before the next school year.

### **4. *DIAGNOSTIC EVALUATION***

Diagnostic evaluations occur before or, more typically, during instruction. Diagnostic evaluations are concerned with skills and other characteristics that are prerequisite to the current instruction or that enable the achievement of instructional objectives. During instruction, diagnostic evaluations are used to establish underlying causes for a student failing to learn a skill. When used before instruction diagnostic evaluations try to anticipate conditions that will negatively affect learning. Diagnostic evaluations are based mostly on

---

informal assessments, although formal measures including standardized tests sometimes are used. Strictly speaking, these evaluation types are formative evaluation.

#### **1.4 TYPES OF EVALUATION PROCEDURES**

Evaluation procedure is broadly classified as quantitative and qualitative. In addition to this, there are two major ways of classifying evaluation procedures. They are in terms of

1. an aspect of behavior to be evaluated, and
2. evaluative method

##### **1. Classification by Aspects of Behavior**

Evaluation in terms of aspects of behavior evaluated can be subdivided into two general categories. They are that are used to

1. determine a person's abilities, and
2. reflect a person's typical behavior.

Procedures of the first type are concerned with how well an individual performs when he/she is motivated to put forth his/her best effort. In short, the evaluation results indicate what an individual can do. Learning abilities are referred to as person's ability. Aptitude and achievement tests are included in this category. An aptitude test is primarily designed to predict success in some future learning activity. It measures what a student will learn in the future. Whereas achievement test is designed to indicate degree of success in some past learning activities. It measures students' ability to remember what they have learned in classrooms.

Procedures of the second type are concerned to answer the question, "How does the individual usually behave in normal or routine situations?" Results in this area tend to indicate what an individual will do rather than what he/she can do. Evaluation of typical behavior falls in the general area of personality appraisal. Methods designed to evaluate interest, attitudes, and various aspects of personal social adjustment are included in this category.

##### **2. Classification by evaluation method**

###### **1. Testing Procedures**

As stated earlier, a test is a series of tasks which is used to measure a sample of a person's behavior at a given time. The most common tests used in schools are achievement tests.

###### **2. Self-report technique**

Every individual has a wealth of information about him/her self. Some of the information may be how he/she feels about certain situations, which activities interest him/her most, or what personal problems are of greatest concern to him/her, and the like. This information can only be obtained directly from the individual through interview or different types of questionnaire.

###### **3. Observational technique**



Reliable information about an individual's typical or usual behavior is best obtained from persons who have observed him/her in a variety of situations. There are various observational techniques. They include anecdotal records, checklists, rating scales, socio-metric techniques.

- Anecdotal records are continuous, objective descriptions of behavior as it occurs at a given time, place, and circumstance. They are continuous in the sense that sequential records are kept of a student's behavior over relatively long periods of time, such as semester or school year. They are objective since they describe what the student has done or accomplished.

They provide the least structured method of recording behavioral observation. They are simply a brief description of some observed behavior which appeared significant for evaluation purpose.

- Checklists usually contain lists of behaviors, traits, or characteristics that are either present or absent. They are frequently used to evaluate aspects of student's interests, attitudes, activities skills, and personal characteristics. For example, if a teacher wishes to check whether a child can draw a table or not, he/she gives a task for the child and can observe when the child draws.
- Ratings are observations that have been categorized or organized to provide summary information about the behavior of individuals or groups.
- The socio-metric technique is a method for evaluating the social relationships existing in a group. Each group member is asked to indicate those individuals they would prefer as associate for some group situation or activities.

### **1.5 TYPES OF TESTS**

As stated in the previous sections, tests are used to measure different behaviors of students. Most individuals think of tests as being limited to true-false, multiple-choice, matching, or essay questions that measure knowledge in arithmetic, reading, or school subjects. However, tests can take many forms. In other words, tests can be classified in many ways. To classify tests we use the following criteria.

- a. kind of item
- b. the nature of item scoring
- c. degree of standardization
- d. administrative conditions of the tests
- e. language emphasis of the responses
- f. emphasis on time to respond to items
- g. score-referencing scheme
- h. the attribute (behavior) being measured

Based on these criteria, we have a number of test types. We will consider the most common types of classification in the following sect.

#### ***A. Kind of Items***

---

The questions, exercise, and task appearing on a test are called items. Some items require students to write the answers for the items and others require them to select the correct alternative from the given list of responses. Thus, we have two types of test

- Selection type items, and
- Supply types items

Selection type items include: True-false items, multiple-choice items, and matching items. Supply type items include short answer items, completion items, and essay items.

### ***B. The Nature of Scoring***

Based on the way we score test items, we can classify tests into two types:

- Objective tests, and
- Subjective tests

Objective tests are any type of tests having a clear and unambiguous scoring criteria. That is different scorers (people who mark the test papers of students) can give exactly the same marks or they can agree on the number of points students can receive. Selection type items, completion items, and short answer items are objective type tests. Essay tests, on the other hand, are scored mainly differently by different scorers, and differently by the same person at different times. This shows that essay items are subjective. The same item can have different possible answers.

### ***C. Degree of Standardization***

Standardization refers to the degree to which the observational procedures, administrative procedures, equipment and materials, and scoring rules have been fixed so that exactly the same testing procedure occurs at different times and places. In other words, the same test is administered at different places at the same time and in the same manner. For example, the Ethiopian Schools Leaving Certificate Examination (ESLCE) has been administered to all students throughout the country at the same time using the same administration procedure. In this regard, we call it standardized. Based on this criterion, we classify test into two:

- Standardized tests, and
- Nonstandardized test

Standardized test are constructed by test specialists working with curriculum experts and teachers. Nonstandardized tests are tests that are administered by teachers any time they wish to get information about their students' progress or when they want to assign grades or marks to their students. Nonstandardized tests are usually called teacher made tests. They are constructed by teachers for use within their own classrooms.

Standardized test can be used to compare the performance of one school students with other schools because they are the same for different school students. Teacher made tests cannot be used to compare students of different schools.

***D. Administrative Conditions of Tests***

Tests may be administered to one person at a time or simultaneously to a group of persons. Accordingly we can classify tests into two:

- Individual test, and
- Group test

Individual tests are tests that are designed to be administered to one person at a time. These individual tests allow the maximum amount of interaction between the examiner and examinee and are rich in opportunities for getting information by observing the student when taking the examinations. The examiner can observe not only an examinee's approach to and performance on the examination tasks but sometimes also can ask questions that follow up on an examinee's response to clarify it and to understand it more completely.

On the other hand, **group tests** are administered to a group of persons at the same time and place. All students are given the same amount of time and to complete each task. Tests that are administered to students in classrooms are examples of group tests. Administering individual tests generally requires much training, time, money, and experience, whereas the administration of group tests is much less complicated.

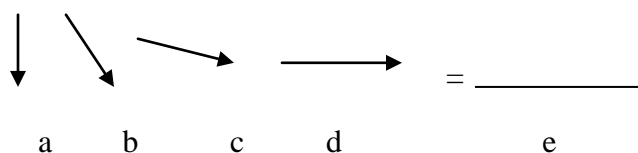
***E. Language Emphasis of the Response***

Based on the nature of the responses required by the items, we can classify tests into three types:

- Verbal tests,
- Nonverbal tests, and
- Performance test

**Verbal tests** emphasize the use of language as the primary means of responding to test items or questions. For example, tests you have taken in high schools are examples of verbal tests. They use written language. **Nonverbal tests**, on the other hand, deemphasize the role of reading or written language.

Example,



What will be in “e”?

This item requires the student to use the figure and not language. Nonverbal tests usually involve paper-and-pencil tasks that make extensive use of pictures to measure vocabulary, reasoning, and other skills. For example, the teacher might ask questions orally (“Which

---

animal gives us milk?") while students look at a picture of a horse, dog, cow, and pig. Students are instructed to select orally to indicate the correct. This type of test is important and useful for children and illiterates.

**Performance tests** require the student to perform a task other than answering questions. For example, measure writing skills by having students write, and measure science knowledge by having students perform or plan an experiment. Performance tests are often administered individually so that the examiner can observe the behavior of the examinee while performing the task.

#### ***F. Emphasis on Time to Respond to Items***

Another way to classify tests is by the speed students are required to complete the test items or to respond. Based on this criterion we classify tests items into:

- Speed tests, and
- Power tests

Speed tests are types of tests that require students to complete the items as fast and quickly as possible within a very short time. In other words, speed tests have severe time limits, but the items are so easy that few students are expected to make errors. In these types of tests, only the most exceptional students will complete the examination within the limited time given. For example, typing test is an example of this type of test. The examinees are required to type as many words as possible.

Quite frequently the main focus in testing students is to assess the amount of knowledge, comprehension, or understanding they possess. Often there is less concern about the rapidity of a student's responses to questions than about the content of those responses. This type of test is power test. A power test has generous time limits so that most students will be complete or attempt every item. In comparison to speed tests, power tests are difficult.

#### ***G. Score-referencing Scheme***

Though not exhaustive, there two types of score referencing schemes in measurement and evaluation. They are norm-referencing and criterion-referencing. Accordingly, there are two types of tests:

- Norm-referenced tests, and
- Criterion-referenced tests

**Norm-referenced tests** are designed to measure individual differences in achievement, intelligence, interests, attitudes, or personality. That is they involve **comparing** a student's performance to the performances of other group students where the student belongs. We often make comparisons to other individuals and events in order to help interpret what we see. How well a student did on a test is often described in terms of how others in the class

did. Describing the height of a student often involves comparing the student to other students. Similarly, measures such as how much it rained, how fast you were driving, and how you feel today are often interpreted through comparison to other similar events. It rained less than it usually does; I have never driven this fast, and so on. These are all norm-referenced interpretations.

This type of interpretation of test scores would be important if it were necessary to limit the number of persons who could be admitted into college or hired for a job or if the testing is to assign grades based on individual differences in knowledge or understanding.

**Criterion-referenced tests**, on the other hand, relate a student's score on an achievement test to a certain criterion or domain of knowledge rather than to another student's scores. That is, these types of tests involve comparing a student's performance to a well-defined content domain. Examples of content domains include being able to locate a word in a dictionary; or being able to solve an algebraic equation involving whole numbers and one unknown. Tests used in elementary and secondary schools are examples of criterion-referenced tests. Students who scored 50% and above in any subject are said to have passed the test.

#### ***H. The Attributes Being Measured***

Tests are used to get information about a variety of human attributes or characteristics. Among them are

- Cognitive tests (academic achievement tests, and variety of general scholastic and specific aptitudes tests), and
- Noncognitive tests (personality, interest, and attitudes)

Cognitive tests are tests that are designed to measure the intelligence, reasoning ability and academic achievements of students. These tests have correct or best answers. Academic achievement tests attempt to measure the knowledge, abilities, and skills that are the focus of direct instruction in schools. For instance, after you learn this course, you will take tests to check you knowledge of the subject. That is an academic achievement test. These tests measure what students have learned in the classrooms.

Scholastic aptitude tests and specific aptitude tests reflect learned behavior also. They are used mainly to predict future academic accomplishments of student in a particular instructional setting. They focus mainly on general knowledge rather than directly learned materials.

Still other human attributes tested are the noncognitive or nonacademic or affective (i.e., the psychological aspects). These are tests that do not have right or wrong, correct or incorrect answers. In psychological testing, the procedures for testing such attributes as emotional adjustment, interpersonal relationship, motivation, interest, and attitudes are called personality. These attributes are often measured by questionnaires asking the examinee's

degree of agreement with various statements. Box 1.1 summarizes the types of tests discussed so far.

BOX 1.1 Classification of tests using different criteria

<p>1. By kind of Item</p> <ul style="list-style-type: none"> <li>a. Selection Type <ul style="list-style-type: none"> <li>- True-False items</li> <li>- Multiple choice items</li> <li>- Matching items</li> </ul> </li> <li>b. Supply Type <ul style="list-style-type: none"> <li>- Short answer items</li> <li>- Completion items</li> <li>- Essay items</li> </ul> </li> </ul> <p>2. By the Nature of Scoring</p> <ul style="list-style-type: none"> <li>a. Objective tests</li> <li>b. Subjective tests</li> </ul> <p>2. By Degree of Standardization</p> <ul style="list-style-type: none"> <li>a. Standardized tests</li> <li>b. Nonstandardized tests</li> </ul> <p>4. By Administrative Conditions</p> <ul style="list-style-type: none"> <li>a. Individual tests</li> <li>b. Group tests</li> </ul>	<p>5. By the Language Emphasis of Response</p> <ul style="list-style-type: none"> <li>a. Verbal tests</li> <li>b. Nonverbal tests</li> <li>c. Performance tests</li> </ul> <p>6. By Emphasis on Time</p> <ul style="list-style-type: none"> <li>a. Power tests</li> <li>b. Speed tests</li> </ul> <p>7. By Score-Referencing Scheme</p> <ul style="list-style-type: none"> <li>a. Criterion-referenced tests</li> <li>b. Norm-referenced test</li> </ul> <p>8. By the Attributes being Measured</p> <ul style="list-style-type: none"> <li>i. Cognitive tests <ul style="list-style-type: none"> <li>a. Achievement tests</li> <li>b. Aptitude tests <ul style="list-style-type: none"> <li>- general aptitude tests</li> <li>- specific aptitude tests</li> </ul> </li> </ul> </li> <li>ii. Noncognitive tests <ul style="list-style-type: none"> <li>c. Personality, interest, and attitude tests</li> </ul> </li> </ul>
--	--

## 1.6 GENERAL PRINCIPLES OF EVALUATION

Evaluation process has the following general principles.

### 1. *Determining and clarifying what is to be evaluated*

No evaluation device should be selected or developed until the purposes of evaluation have been carefully defined or determined. This is what the teacher should do at the beginning of evaluation process. Because the effectiveness of the evaluation process depends as much upon a careful description of what to evaluate as it does upon the technical qualities of evaluation instruments used.

### 2. *Selecting evaluation techniques in terms of the purpose to be serve*

While selecting evaluation technique you should answer the question: Is this evaluation technique the most effective method for determining what I want to know about the students? Each evaluation technique is appropriate for some purposes and inappropriate for others. For instance, objective tests are most effective for measuring some educational objectives and

essay test are most effective for others. The question is not: Should this technique be used? But, rather when should this technique be used?

### ***3. Using a variety of evaluation techniques***

Comprehensive evaluation requires a variety of evaluation technique. No single evaluation technique is adequate for appraising students' progress toward the entire important outcomes of instruction. In fact, most evaluation techniques are rather limited in scope. An objective test of factual knowledge provides important evidence concerning a student's achievement but the result tells us little or nothing about how well he/she understood the material, the extent to which he/she is developing thinking skills, how attitudes are changing, how he/she would perform in an actual situation, and the like. Such outcomes require evidence beyond that can be obtained by an objective tests. Essay test, self report techniques, and various observational methods would all need to be used to evaluate such a diverse array of instructional outcomes.

One reason we have so many different types of evaluation procedure is that each provides unique but limited evidence on some aspect of behavior. To get a more complete picture of a student's achievement, we need to combine the results from a variety of techniques.

### ***4. Being aware of the limitations (weaknesses) of the evaluation techniques used***

Proper use of evaluation technique requires awareness of their limitations as well as of their strengths.

Evaluation techniques vary from fairly well developed measuring instrument (e.g. scholastic aptitude test) to rather crude observational methods. All are subject to one or more types of error.

1. Sampling error: When we are preparing tests, for example, we may focus on some parts of the course content but give no or little attention to others. This is a sampling error. Generally, however, since we can only measure a small sample of a student's behavior at any one time, there is always a question of the adequacy of the sample.
2. The other error is found in the evaluation instrument it self or in the process of using the instrument. For example, scores on objective tests are influenced by chance factor such as guessing; scores on essay tests are affected by the subjective judgment of the person doing the scoring, the result of self report techniques are distorted by the individual's desire to present himself in a favorable light; and observations of behavior are subject to all of the biases of human judgment. These and other errors inherent in the use of evaluation techniques must be recognized if the techniques are to be used wisely.
3. A major source of error arises from improper interpretation of evaluation results.

---

In general, a healthy awareness of the limitations of evaluation instruments makes it possible to use them most effectively.

## UNIT 2

### Classroom (Teacher-Made) Tests

#### 2.1. Planning Stage

Effective teaching is not a matter of what a teachers wishes to do; rather, good teachers know what changes in student behavior they want to produce and hold themselves accountable by modifying teaching strategies until objectives have been met. In practice, teaching involves setting goals that consider student need and backgrounds and selecting the most effective instructional strategies. Objectives stem (originate) from knowledge of student needs, societal demands, legal requirements and various philosophic assumptions concerning human nature.

Writing objectives in itself is not an end itself, it is a means to an end, a way of improving student learning. Teachers make sure whether students attained the stated objectives through tests. Therefore, teachers should prepare tests that can effectively and efficiently measure the intended learning outcomes, i.e., the objectives.

When planning a test, a teacher needs to take into account the following elements or guidelines.

#### 1. Defining the purpose of the test

This includes answering the following questions.

Why is the teacher testing? Is the teacher interested in making decisions about students' placement, selection, classification, or other type of decisions?

What does the teacher intend to measure? Is the teacher interested in measuring students' knowledge of facts, writing ability, understanding or other type of objectives?

How will the test scores be used or what type of score interpretation does the teacher want to make?

#### 2. Preparing table of specification

#### 3. Selecting appropriate test format

#### 4. Developing the initial draft of the test

#### PREPARING TABLE OF SPECIFICATIONS

Test items should represent important and clearly stated objectives. Some teachers unintentionally construct most of their test items from very few parts of the course. Because most tests are samples of behavior, teachers need to construct items that adequately sample subject matter from all major topics that are to appear on the examination. The most effective way to ensure adequate representation of items is to develop a two-way grid called a table of



specifications or test blueprint. This table relates the two major components of educational process: the content element and the objectives.

The numbers within the table of specifications refer to the number of items that are to be prepared. The number of items should reflect the emphasis the teacher wishes to give to each topic. The following criteria should determine the approximate number of items for each objective:

1. the relative importance or weight assigned content areas,
2. amount of instructional time devoted
3. roles as a future prerequisite
4. other opportunities to evaluate

Let us say the following are the instructional objectives and course contents. The teacher wishes to construct items for each of the instructional objectives and the course contents. The table of specifications combines the two elements together as indicated in Table 3.1.

### **Instructional Objectives**

At the end of the course, the student will be able to

1. know the concept of measurement and evaluation (implicit objective). Since this objective is difficult to measure. Hence we translate this objective into specific objectives as follows:
  - define the word measurement
  - define the word evaluation list types of evaluation
2. understand the role of evaluation (implicit objective)
  - differentiate the role of summative evaluation from formative
  - differentiate the role of diagnostic evaluation from summative
  - list Bloom's taxonomy of objectives
  - identify the six levels of cognitive domain
3. know the concept of table of specification (implicit objective)
  - define the phrase table of specification
4. apply the knowledge of preparation table of specifications
  - prepare table of specifications
5. compare and contrast Bloom's Taxonomy of objectives with Gagne's classifications of objectives.
6. produce different types of objectives.
7. evaluate the importance and use of objectives in the instructional process.

### **Content Outlines**

Definition of concepts  
Measurement  
Evaluation  
The role of evaluation

Preliminary evaluation  
 Summative evaluation  
 Formative evaluation  
 Diagnostic evaluation  
 Classification of objectives (Bloom's taxonomy)  
 Table of specification

Table 3.1 Test blueprint (table of specifications) for mid term examination of measurement and evaluation

Contents	Objectives						
	Knowledge	Understanding	Application	Analysis	Syntheses	Evaluation	Total
Definition of terms	3	2	1				6
The role of evaluation in education	2	6	3	2		1	14
Classification of objectives	2	5	3	1	1		12
Table of specification	1	1	1				3
Total	8	14	8	3	1	1	35

The row heading along the left margin lists the major topics to be covered in the test. The teacher can use a more detailed outline if he/she wishes, that is, it is possible to put main topics and subtopics in the table. The column heading across the top lists the major classifications of Bloom's taxonomy of cognitive educational objectives. Other taxonomies can also be used. Notice that there is an increasing complexity from left to right in the types of objectives to be measured, i.e., test items include simple knowledge of facts and higher learning objectives such as synthesis and evaluation. Performance that demonstrates knowledge or comprehension, for example, is lower level of cognitive domain. Those reflecting the ability to synthesize or evaluate are higher level cognitive performances.

#### ***Uses of table of specifications***

Generally, the use of table of specification or test blueprint in test development will help ensure that

1. only those objectives actually pursued in instruction will be measured
2. each objective will receive the appropriate relative emphasis in the test.
3. by using subdivisions based on content and behaviors, no important objectives will be overlooked or misrepresented.

In general, the behavior categories selected for a table of specifications are heavily influenced by the student's level and the nature of the subject matter.

### **SELECTION OF APPROPRIATE TEST FORMAT**

The two broad categories of test formats are subjective (essay) and objective item formats. Objective item format includes true-false, completion (fill in the blank), short answer, multiple-choice and matching. Subjective type of item format includes essay items both restricted and extended responses. Essays present a specific problem which requires the student to recall information, organize it in a suitable manner, derive a defensible conclusion, and express it within specific guidelines. The following are similarities and differences between subjective and objective item formats.

#### ***Similarities***

- 1) Both formats can be used to measure almost any important educational achievement that any paper-and- pencil (i.e., written type) tests can measure.
- 2) Both can be use to encourage students to study for understanding of principles, organization and integration of ideas, and application of knowledge of the solution of problems.
- 3) The use of either type necessarily involves the exercise of subjective judgment.

#### ***Differences***

1. Subjective test question requires students to plan their own answers and to express them in their own words. An objective test item requires examinees to choose among several given alternatives.
2. Subjective test consists of relatively few, more general questions that require extended answers. As objective test ordinarily consists of many specific questions, they require only brief answers.
3. Students spend most of their time in thinking and writing when taking subjective tests. They spend most of their time reading and thinking when taking an objective test.
4. Subjective tests are scored subjectively whereas objective tests are scored objectively.
5. Subjective test is relatively easy to prepare but rather tedious and difficult to score accurately. A good objective examination is relatively tedious and difficult to prepare but comparatively easy to score.
6. Subjective examination affords students much freedom to express their individuality in the answers they give and much freedom for the examiner to be guided by his or her individual preference in scoring the answer. An objective examination affords much freedom for the test constructor to express personal knowledge and values but allows students only the freedom to show, by the preparation of correct answers they give, how much or how little they know or can do.

- 
7. Objective tests cover a wide range of course content and instructional objectives which ensure the representativeness of test to measure the intended learning outcomes, whereas essay items include only a few areas of course content.
  8. Subjective test permits, and occasionally encourages, bluffing. An objective test permits, and occasionally encourages, guessing.

In view of these similarities and differences, when might it be most appropriate and beneficial to use subjective test?

Essay or subjective tests are favored for measuring educational achievement when

- the group to be tested is small and the test will not be used again.
- the teacher wishes to provide for the development of students skill in written expression.
- the teacher is more interested in exploring student attitudes than in measuring achievements.
- the teacher is more confident of his or her proficiency as a critical essay reader than as an imaginative writer of good objective test items.
- time available for test preparation is shorter than time available for test scoring
- relatively few areas of the content need to be tested

The teacher should know which type of test format does help him achieve the purpose he/she intends to measure. As stated earlier one type is more appropriate to measure certain types of behaviors than others and not appropriate for others types of behavior. So the teacher should know which type of test format to select.

### **Factors to be Consider When Selecting An Item**

There are a number of common problems in developing or selecting tests to assess student achievement. Among them

- Failure to consider objectives and instructional emphases when planning a test.
- Failure to cover all level of cognitive domain.
- Failure to assess all of the important objectives and instructional topics covered during instruction.
- Failure to select item types that permit students to demonstrate the desired behavior.
- Adopting a test without reviewing it for its relevance to the instruction provided
- Including topics or objectives not taught to students
- Including too few items to assess the consistency of student performance
- Using tests to punish students for inattentiveness or acting out.

Thus when you select test format consider the following factors.

1. The purpose of the test
2. The time available to prepare and score the test
3. The number of students to be tested
4. The physical facilities available for reproducing the test

5. Age of students
6. Skills in constructing different types of items.

### ***1. The purpose of the test***

The most important factor to be considered in the selection of item formats is what you want the test to measure. To measure expression of ideas, you would use the essay; for spoken self expression, the oral. To measure the extent of the student factual knowledge, his understanding of principles or ability to interpret, the objective test is preferable because it is more economical and tends to possess higher score reliability and content validity. If the purpose is to use the test results to make decisions for grading purposes or admission to college, the objective tests are recommended because of greater sampling of content and more objective scoring

### ***2. The time available to prepare and score the test***

It will take less time to prepare five extended response essay questions for a two-hour twelfth – grade test than it would be prepare 75 multiple choice items for that the same test. But the time saved in preparing the essay test may be used up in reading and grading the responses. The time element becomes of concern in relation to “When” the teacher has the time. If she/he is rushed before the test is to be administered, but will have sufficient time after it has been administered, she/he might choose to use an essay examination. But, if she/he must process the results within two or three days she/he should use the objective test, provided she has sufficient time to write good objective items.

### ***3. Number of examinees***

If there are only a few students to be tested and if the test is not to be reused, then the essay or oral test is practical. However, if a large number of students are to be tested and/or if the test is to be reused at a later time with another group, the objective test is appropriate. Because it is much harder to remember 75 objective items than to remember 5 or 6 essay items.

### ***4. Physical facilities***

If duplication or reproduction facilities are limited or unavailable, the teacher is forced to use either the essay test, with the questions written on the board, or the oral test, or he/she can use the true-false or short answer item by reading the questions aloud. However, multiple-choice items must be reproduced mechanically, i.e., they should be duplicated using duplication machines. Because they cover large amount of material that is difficult to write all items on blackboard.

### ***5. Age of students***

When constructing test items consider the age of students. Objective type items are more appropriate for young children than essay tests. Using a variety of test format creates

---

confusion for young children than the older one. For older students it is recommended to use a combination of different types of items.

#### ***6. The skills teachers have to construct test item***

Teachers develop test items that they have the skills to construct the items. Generally, constructing objective test needs more skill than essay test.

### **3.5 PREPARING TEST ITEMS**

The primary aim of assessing student achievement is to provide students a fair opportunity to demonstrate what they have learned from the instruction provided. It is not to trick students into doing poorly, entertain them, or ensure that most of them get “A” grades. It is not to determine how much total knowledge students have accumulated as a result of all their learning experiences, both in and out of school. It is simply to let students show what they have and have not learned from the things they have been taught in a particular classroom (Airasian, 1996). To do so the teacher should consider the following steps in preparing test items.

#### ***1. Determine how results of the test will be used***

In the classroom, written tests are most commonly used for formative or summative evaluations. Formative evaluations occur during a unit or chapter of instruction and are used to determine how instruction should proceed, and to identify the weaknesses and strengths of students and the instruction.

Summative evaluations occur at the end of the unit or chapter of instruction. These evaluations are used to certify achievement and assign grades. Thus, in the preparation of test items, the teachers first need to determine the purpose of the test.

#### ***2. Determining the type and number of skills to be measured by the test***

The type of skills limits which item format may be used. Any of the formats can measure recall of information. Students’ ability to recall information represents an important set of skills and deserves considerable emphasis. At the same time, classroom tests are often criticized for measuring only these types of skills. The essay, multiple-choice, and true-false formats can more easily measure intellectual skills than can the short answer format. The essay format can also directly measure students’ ability to organize and express ideas in writing. Problem solving skills are more easily assessed using direct observation and other performance measurement techniques.

The number of skills also influences the item format to be used. For example, essay items are appropriate only if the number of skills to be measured is quite limited.

#### ***3. Determining the type and number of items to be used in the test***

A test may contain a mixture of item formats, that is in a single test, it is possible to include different types of test items. If the class is small, short answer and essay items may be preferred, if there is enough time to measure adequately the skills to be assessed and to score students' responses. Short answer items are restricted in the type of skills they can assess and essays are more limited in the amount of content that can be sampled within a given period of time. Multiple choice and true-false items must be used when answers are to be quickly scored. The number of items is also limited by the amount of time available to administer the test.

**2. *Determine the number of items to be associated with each objectives or devise a table of specification***

At this point, the skills to be tested and total number and type of items to be used on the test have been established. Now, the number of items used to assess each skill is to be determined. This is accomplished by developing a table of specifications or by deciding how many items should be used to measure each performance objective. The number of points associated with each skill should be proportional to the emphasis it should receive. The number of test items to be associated with each skill are also limited by the total number of items that can be include in the test and the number of skills that the test will measure.

**3. *Preparing the required test items***

Preparing items is usually very time consuming. There are five primary criteria that should be used to guide item writing. These are

***i. Matching items with the stated objectives***

Congruence between items and objectives helps ensure that test scores are valid. When writing items, teachers should pay careful attention to the behavior, content and the conditions prescribed in the objectives. For example, if our objective is stated in as follows

At the end of this lesson student will be able to define the term measurement,

then, our item should be: Define the term measurement

***Behavior***

The response format selected for test items should match the behavior prescribed in the objective. It is important to note whether the student needs to define, list, select, solve construct, perform, or evaluate. As teachers write each item, they should consider alternative ways the student might respond to demonstrate the behavior and then write the items that represent the prescribed behavior.

***Content***

---

The content included in each objective may be simple or quite complex. Instances of content selected for test items should clearly reflect all facets of the content students have learned. For example, if a teacher has taught his/her student about proper nouns, he/she should develop items that are related to persons, places, and things. Writing items related to only one or two areas of content would be insufficient.

### ***Condition***

The conditions of behavioral objectives which describe what students are to be given can specify a variety of resource materials. They may require that students be provided with such materials as short stories picture, charts, diagrams, data tables, models, slides, specimens or particular products. They may also specify such equipment as computers, saws, dissecting tools, sewing machines, weighting instruments, or automobiles. In addition, conditions sometimes include facilities, such as the school library or a learning resource center. The resource materials provided to students during testing must be congruent with the conditions specified in the objective. These materials are just as critical to valid measurement as is the nature of the items.

Of the five primary criteria for designing test items, the congruence between the elements of objectives and the test items is the most critical. When objectives and test items do not match or are not congruent, then something different from the intended objective is most likely measured by the items.

### ***ii. Congruence of Items with target students' characteristics***

In order to create tests that result in valid and reliable score, the item must be congruent with the characteristics of target students. Factors that can influence the validity of a test are students' reading levels, vocabulary, and experience. Students of any age who are poor readers will have difficulty responding to items that require a lot of reading, their answers may reflect their reading skills more than their achievement of the objective. Therefore, vocabulary used in either written or oral items should be familiar to students to ensure that their response reflects their achievement of the objective and not their vocabulary. For example, if teachers are selecting words to use in measuring students' ability to classify nouns, each word used to measure the skill should be one that all students in the group can define. The words and sentences used in the test items should be familiar to the students. They should not be difficult and new to the students.

### ***iii. Clarity of items***

The clarity of items also affects the validity and reliability of a test.

To assure valid measures:

- items should be carefully constructed to avoid ambiguity and unintended complexity. Clearly written items allow students to focus their attention on the actual skill being measured.



- the language should be clear, simple, and precise
- grammar, sentence structure, and punctuation should be correct.
- each item should pose only one question. Questions that contain multiple ideas or that address several issues will confuse students of any age.
- irrelevant, unnecessary or extraneous materials should be avoided.
- instructions for responding should be clear, explaining how students are to respond as well as how resource materials are to be used.

To ensure reliability items should have only one correct or clearly best answer. Including more than one correct answer in one item confuses students and can result in inconsistent answers, scoring analysis and interpretation.

#### ***4. Assemble the Items into a test***

This includes ordering the items, determining the layout of items within the test paper and establishing instructions to be placed at the beginning of the test.

### **3.6. WHAT A TEACHER NEEDS TO HAVE TO BE A GOOD ITEM WRITER?**

The process of writing good test items is not simple - it requires time and effort. It also requires certain skills and proficiency on the part of the item writer. Some of which can be improved by formal course work, others require considerable practices. Rules, suggestions, guidelines, and textbooks may be useful, but they are not the panacea for writing valid test items. To be a good item writer one should be proficient in the following six areas:

#### ***1. Know the subject matter thoroughly***

The greater the item writer's knowledge the subject matter, the greater the likelihood that she/he will know and understand facts and principles as well as some of popular misconceptions. This latter point is of considerable importance when writing the selection type of item in general, and the multiple choice item in particular (because the item writer must supply plausible although incorrect answers).

#### ***2. Know and understand the students being tested***

The kinds of students the teacher deals with will determine in part the kinds of item format, vocabulary level, and level of difficulty of the test items. For example, primary school teacher seldom use multiple choice item because young children better able to response to the short answer type. The vocabulary level used for class of gifted children may be very different from that used with a class of educable but mentally retarded children. The classroom teacher who knows and understands his/her students will generally establish more realistic objectives and develop a more valid measurement device than will the teacher who fairly to consider the characteristics of her/his students.

#### ***3. Be skilled in verbal expression***

---

It is essential that the item writer clearly conveys (presents) to the examinees the intent of the question. In an oral examination, the student may have the opportunity to ask for and receive clarification of the question when he/she does not understand what the teacher is asking. But in paper-pencil test this is less possible. Hence, the items should be clearly written and to do that the teachers should have the skill to use the language instruction.

**4. *Be thoroughly familiar with various item format***

The item writer must be knowledgeable of the various item formats – their strengths and weaknesses, the error commonly made in this and that type of item – and guidelines that can assist her in preparing better test item.

**5. *Be preserving (try hard to write and improve)***

Writing good test items, regardless of their format, is both an art and a skill that generally improves with practice. There are very few professional item writers who are so gifted, able, and blessed to write items that requires absolutely no editing or rewriting. Depending upon the skill of the item writer, the number of the items that need rewording or will be rejected will vary. The important thing is that classroom teachers who are trained as teachers rather than as item writers should be preserving and not give up, even though the task seems demanding.

**6. *Be creative***

Item writing needs creativity. The teachers' ability of writing items in a novel way is very crucial.

# Unit 3

## Writing Objective Test Items

### 1. Writing Supply Type Items

In this part of the unite, we will see two types of supply type items that are commonly used to measure students' academic performances. They are short answer type and completion type. Their advantages and limitations and the guidelines to construct them will be presented.

The short answer item and the completion item both are *supply type test items* that can be answered by a word, phrase, number or symbol. They *are essentially the same*, differing only in the *method of presenting the problem*. The short answer item uses a direct question, whereas the completion item consists of an incomplete statement.

The **completion test item** asks the student to complete a sentence with a word or short phrase.

Example: What is the square root of 25?

Who invented radio?

The **short answer test item** poses a question that can be answered with a word or a phrase.

Example: The square root of 25 is \_\_\_\_\_.

The inventor of radio was called \_\_\_\_\_

### USES OF SUPPLY TYPE ITEMS

Both the short answer test item and completion item are suitable for measuring a wide variety of relatively simple learning outcomes. Some of its common uses are for measuring:

- Knowledge of terminology

#### *Example*

1. The family of those of animals that feed on the flesh of other animals is classified as \_\_\_\_\_. (carnivorous).

2. What is the common name of each of the following chemical substances?

- |  |       |              |
|--|-------|--------------|
| 1. $\text{CaCO}_3$                           | _____ | (Lime Stone) |
| 2. $\text{NaCl}$                             | _____ | (Salt)       |
| 3. $\text{C}_{12}\text{H}_{22}\text{O}_{11}$ | _____ | (Sugar)      |
| 4. $\text{NH}_3$                             | _____ | (Ammonia)    |

- Knowledge of specific facts.

---

**Example**

1. The battle of Awa took place in the year \_\_\_\_\_. (1888)
2. Water boils at \_\_\_\_\_ °C. (100)

- Knowledge of principles

**Example.** If the temperature of a gas is held constant while the pressure applied to it is increased, what will happen to its volume? (It will decrease)

- Knowledge of method or procedure

**Example.** What device is used to detect whether an electric charge is positive or negative? (Electroscope)

- Simple interpretations of data

**Example.** If an airplane flying northeast made a 180-degree turn, what direction would it be heading? (Southwest).

## **ADVANTAGES AND LIMITATIONS OF SHORT ANSWER AND COMPLETION ITEMS**

Dear students, in this section you will be introduced with the major advantages and weakness of supply type items. First, we present their advantages.

### **Advantages**

#### ***1. Construction is relatively easy***

The first advantage is they are the easiest to construct compared to other objective type items such as multiple choice items. As a result, they are popular partly because of the relatively simple learning outcomes they usually measure. But in the areas of mathematics and science it is possible to use them to measure somehow more complex learning where the solutions to problems can be indicated by numbers or symbols.

#### ***2. Guessing is eliminated or minimized***

A more important advantage of short answer and completion items is that the students must supply or give the answer. This reduces the possibility that the students will obtain the correct answer by guessing. They must either recall the information requested or make the necessary computations to solve the problem presented to them. In contrast, in other objective test items like multiple choice and true-false students who have partial knowledge may get the correct answer by guessing benefit. Partial knowledge is insufficient for answering a short answer test item correctly.

#### ***3. Item sampling is relatively high***

They allow for the construction of large number of items. Because completion and short answer items take relatively less time to read and respond than multiple choice items do,

---

classroom teachers can construct large number of items. However, this advantage will not hold if the items deal with computations.

## **Limitations**

### ***1. They typically measure rote memorization or simple learning outcomes***

They are unsuitable for measuring complex learning outcomes. Except for the problem solving outcomes measured in mathematics and science, the supply type items are used almost exclusively to measure the recall of memorized information. They cannot measure writing, organization, and synthesis ability of the learner. Since short answer and completion items are answered by a word, a phrase or number, they tend to focus on specifics.

### ***2. They are difficult to score***

In these types of items there is difficulty of scoring because students can provide many correct answers for the same item. That is, unless the question is very carefully phrased or prepared, many answers of varying degrees of correctness must be considered for total or partial credit.

***Example*** Atse Tewodros was born in \_\_\_\_\_.

For this question students can give different correct answers. It could be answered by the name of the village, the area, the region, the country, or the month, the year. Although the teacher may have had the year of his birth in mind when he wrote the question, the other answers could not be dismissed as incorrect. But even when this problem is avoided, the scoring is contaminated by the student's spelling ability. With misspellings it is difficult to determine what the student had in mind. The complications make scoring more time consuming and less objective than that obtained with selection type items.

### ***3. They are susceptible to bluffing***

Although supply type items eliminate guessing, this advantage may be offset by an increase in bluffing. If students do not know the required answer to an item, writing the answer for another item might still earn more credit than would they leave the item unanswered. This is what is called bluffing.

## **SUGGESTIONS FOR CONSTRUCTING SHORT ANSWER AND COMPLETION ITEMS**

Dear student, in this part you will learn about the suggestion that supply type test items writer should follow to produce valid, or good items that measure the intended learning outcomes.

### ***1. Word the item so that the required answer is both brief and specific.***

First of all the answer to an item should be a word, phrase, number or symbol. This can be easily conveyed to the students through the directions at the beginning of the test and by

---

proper phrasing of the question. The question should be clear so that students can provide a specific answer. Consider the following examples.

**Example**

An animal that eats the flesh of other animals is \_\_\_\_\_. (This is a poor item)

We can rewrite this item in the following way, and see the difference.

An animal that eats the flesh of other animals is classified as \_\_\_\_\_. (better)

In the first question there are many possible answers like dog, cat, wolf, lion, hyena, etc., can be written but the way the second question is phrased allows students to give only one answer, which is *carnivorous*.

**Example**

What is coal? (Poor)

This item can be answered by a number of correct answers such as fuel, a burning ember, impure carbon, etc.

From what substance is coal formed? (Better)

Can you see the difference? So when we are writing supply type items we should be careful in determining the answer more specific.

**2. Do not take statements directly from textbooks to use as a basis for short answer items.**

Taking statements directly from the context of text book statements are frequently too general and ambiguous (unclear) to serve as good short answer items.

**Example**

1. Chlorine is \_\_\_\_\_. (Poor)

2. Chlorine belongs to a group of elements that combine with metals to form salts. It is therefore called \_\_\_\_\_. (Better)

*Ans. Halogen*

In the first statement there is nothing to imply that the word halogen is wanted. The only students who are can supply the intended answer would be those who memorized the text book statements. Otherwise, it can be filled with variety of answers that include molecule, group seven element, and gas.

**3. A direct question is generally more desirable than an incomplete statement.**

It is better to use short answer type items than completion type items. There are two advantages to the direct question form. First, it is more natural to the students, as this is the usual method of phrasing questions in daily classroom discussions. Second, the direct

---

question is usually better structured and free of much of the ambiguity that creeps into items based on incomplete statements. The phrasing of a question requires us to decide what answer we want to know.

Poor: Emperor Menilik became a king of Ethiopia in \_\_\_\_\_ (Year)

Better: When did Emperor Menilik become a king of Ethiopia? (Year)

Best: In what year did Emperor Menilik become a king of Ethiopia? (Year)

In the first item it is possible to put answers like: Shewa, Ankober, etc., but this ambiguity does not exist in the later two questions.

**4. *If the answer is to be expressed in numerical units, indicate the type of answer wanted.***

For computational problems, it is usually preferable to indicate the units in which the answer is to be expressed.

Indicating the units will clarify the problem and will simplify the scoring. Consider the following examples.

Poor If oranges weigh 100 grams each, how much will a dozen oranges weigh?

Here the answer may be provided as 1200 grams or 1.2 kg. or 1 kg and 200 gm. All are correct.

Better If oranges weigh 100 gm. each, how much will a dozen oranges weigh?  
\_\_\_\_\_ kg. \_\_\_\_\_gm.

It is also usually helpful to indicate the degree of precision expected in the answers. For example, specifying that the answers should be “carried out to two decimal places” or “rounded off to the nearest tenth of a percent” makes clear to the students how far to carry their calculations. According to Linn & Gronlund (2000:177), specifying degree of precision apart from relieving problem of scoring saves students time that could be wasted in calculating to a degree of precision that is unwanted.

**5. *Blanks for answers should be equal in length and in a column to the right of the question.***

If blanks for answers are kept equal in length, they will not give a hint or a clue to the correct answers. When there is a difference in the length of the blanks, students will expect a long answer for the long blank, and a short answer for a short blank. And placing the blanks in a column to the right of the question makes scoring quicker and more accurate.

---

**6. When completion items are used, do not include too many blanks.**

If a statement is too mutilated by blanks, the meaning will be lost and the student will have to guess what the teacher had in mind.

Example

Poor \_\_\_\_\_ was established in \_\_\_\_\_.

Better Addis Ababa was established in the year \_\_\_\_\_.

In the poor item above a variety of options are correct. The blanks could be filled with any possible answer.

**7. Put the blank space at the end of the item, if possible.**

When blank spaces are put at the beginning of the sentences, students may need to read the item many times to understand what they are required to do. In contrast, if the blank appears towards the end, students can more easily understand what is being asked for.

Example

Poor \_\_\_\_\_ is a measure of the extent to which a test measures what it is intended to measure.

Better A measure of the extent to which a test measures what it is intended to measure is called\_\_\_\_\_.

**8. Avoid providing irrelevant clues**

Poor If a triangle has two equal sides then the triangle is an\_\_\_\_\_.

Better If a triangle has two equal sides then the triangle is a/ an\_\_\_\_\_.  
(isosceles triangle)

In the poor item, students are provided with irrelevant clue that is the presence of “an”. That is to mean students with no knowledge about the content but who know what will follow after “an” know that the missing word is something that starts with a vowel. In the better item, in contrast, students will not have any clue as to whether the missing word should start with a vowel or a consonant.

Similarly, in an item in which the answer called for is the name of two or more people the existence of are/were right before the blank space might suggest that the people are at least two. Therefore, it is suggested that you end the item with were/was or are/is.

Example

Poor: Gestalt psychology was formulated by psychologists\_\_\_\_\_, \_\_\_\_\_ and \_\_\_\_\_.

Better: Gestalt psychology was formulated by psychologist(s) \_\_\_\_\_.



---

In the poor item, the student is given a hint about the number of psychologists who have formulated the field of Gestalt psychology. But in the second case the clue is eliminated, so a student who has no knowledge of the subject matter cannot get the answer correct.

### **9. *Omit important words only***

Look the following completion item.

Poor    The first man to \_\_\_\_\_ the moon is Uri Gagarin.

In the above test item the blank space could be filled with a variety of words or phrases which include “land,” “go to,” and “visit.” Instead the words that could be worth omitting include the year in which the man landed at the moon and the name of the person.

## **2. TRUE – FALSE ITEMS (ALTERNATIVE RESPONSE ITEMS)**

### **I) True or False**

**Example.** Photosynthesis is the process by which plants make their own food.

True                      False

### **II) Yes or No**

**Example.** Is 51% of 38 more than 19?                      Yes                      No

### **III) Fact or Opinion**

**Example.** Other countries should adopt a constitution like that of the United States.

Fact                      Opinion

## **USES OF TRUE – FALSE ITEMS**

The most common use of the true-false item is in *measuring the ability to identify the correctness of statements of facts, definition of terms, and statements of principles*. However, they can also become measures of understanding. For true-false items to be measures of understanding there should be some sort of novelty in them.

## **ADVANTAGES AND LIMITATIONS OF TRUE FALSE ITEMS**

### **Advantages**

#### **1. *They are Efficient***

True-false items save time to respond. That is, in a given period of time, students can typically respond to roughly three true-false items in the time it takes to respond one or two multiple-choice items.

#### **2. *Item Sampling***

Because true-false items and answers tend to be short, teachers can examine students on more material (content area) in a time period than any other kind of item. The true-false



---

### ***3. Emphasis on rote memorization***

Although there is a tendency, at least theoretically, to deemphasize rote memorization, there is still a high practice of measuring lower levels of the cognitive domain, i.e., knowledge of facts, terms, and definition of concepts. True-false items are used to measure mainly these levels of the cognitive domain. If they are carefully constructed, true-false items are used to measure comprehension, and application.

### **SUGGESTIONS FOR CONSTRUCTING TRUE FALSE ITEMS**

Dear student we now see the suggestions by which teachers follow to develop quality true-false items. Item construction requires technical and creative skills on the part of teachers. Learning to write good items is important if the test is to be of value to students and teachers to measure the required learning outcomes, i.e., students' academic performance.

The main task in constructing true-false items is formulating statements free from ambiguity and irrelevant clues. This is extremely difficult and the only guidance that can be given is a list of things to avoid when phrasing the statements.

#### ***1. Avoid trivial statements.***

The item should test an important idea. In an attempt to obtain statements that are unequivocally true or false, we sometimes turn to specific statements of facts that fit this criterion beautifully but have little significance from a learning stand point.

#### ***Example 1***

Emperor Menilik was 10 years younger than his wife. (T)

This item has little to do with the important events of the time of Menilik.

#### ***Example 2***

EPRDF controlled Addis Ababa on the 19<sup>th</sup> of Ginbot 1983 E.C. (F)

This item expects the student to remember that the EPRDF did not control Addis Ababa until Ginbot 20. Such items cause students to direct their attention toward memorizing details at the expense of more general knowledge and understanding.

#### ***2. Avoid tricky statements.***

Some classroom teachers try to make a true statement false by making some insignificant changes like misspelling names of persons. This should be avoided because the purpose of testing is not measuring to what extent students can be fooled. The purpose rather is to see to what extent students are achieving important outcomes of a given course or subject.

#### ***3. Avoid the use of negative statements, especially double negatives.***

---

Students tend to overlook negative words such as no or not, and double negatives contribute to the statement's ambiguity.

**Example**

None of the steps in the experiment was unnecessary (poor item)

All of the steps in the experiment were necessary. (better item)

When a negative word must be used, it should be CAPITALIZED, underlined or put in *italics* or in **bold** so that students do not overlook it.

**Example**

Poor Sigmund Freud was not the first person to identify the subject matter of psychology.

Better Sigmund Freud was NOT the first person to identify the subject matter of psychology.

Best Sigmund Freud was the first person to identify the subject matter of psychology.

**4. Avoid long complex sentences.**

It is not good to use long sentences because they are difficult to be understood. A test item should indicate whether a student has achieved the knowledge and understanding being measured. Long complex sentences tend also to measure the extraneous factor of reading comprehension. It, therefore, should be avoided in tests designed to measure achievement.

**Example**

Despite the theoretical and experimental difficulties of determining the exact PH value of a solution, it is possible to determine whether a solution is acid by the red color formed on litmus paper when it is inserted in to the solution. (poor)

Litmus paper turns red in an acid solution. (better)

It is frequently possible to shorten and simplify a statement by eliminating nonfunctional material and restating the main idea.

**5. Avoid including two ideas in one statement, unless cause-effect relationships are being measured.**

An item based on a simple idea is usually easier to understand than one based on two or more ideas. It is not advisable to use items that contain two ideas. The problem here is one idea may be false and the other may be true. The student will be in problem to decide the answer for the item. To see this look at the following true false item.

Poor      T      F      A mule does not give birth because it does not have sexual apparatus.

The above statement has to be keyed false. A student may answer it false. But when asked why he /she may say a mule gives birth which is erroneous. The teacher, however, may be of

---

the opinion that those who know the reason why mule does not give birth will get the item right. To avoid such dilemma, Linn & Gronlund (2000:184) suggest making both the cause and the result true and the relationship either true or false.

**6. Avoid use of ambiguous words.**

For example look at the following true false items.

- |   |   |  |
|---|---|--|
| T | F | 1) Large numbers of endemic animals are found in Awash National Park |
| T | F | 2) Blood clotting takes place in a few minutes.                      |

The above two true-false items cannot be unequivocally answered either true or false. This is because the terms *large* and *a few* are not definite. To what extent should the number of the animals be to be qualified as large and how fast the blood should be to be qualified as taking a few minutes are not well defined. The same is true with the use of words like some, few, a lot of, etc. Thus, it is suggested that instead of using these words it is preferable to indicate numbers in items.

**7. Avoid the use of specific determiners.**

A specific determiner is a word or phrase that provides unintended clue to the correct answer. A specific determiner helps the unprepared student to respond correctly. To see what specific determiners are like, look the following true-false items.

- |   |   |   |
|---|---|---|
| T | F | 1. All large cities are connected by railways.              |
| T | F | 2. No school system is supported entirely by local funds.   |
| T | F | 3. It is possible to bisect any angle.                      |
| T | F | 4. It is impossible to run a mile in less than 3:30         |
| T | F | 5. Revolutions have always led to socially desirable goals. |
| T | F | 6. Wars are never justified in democracy.                   |
| T | F | 7. Some wars could have been prevented.                     |

In the above list of true false items generalizations indicated by absolute terms like *always*, *all*, and *never* (items 1, 2, 4, 5, and 6) are likely to be keyed *false*. Item 3 should most probably be keyed *true* because few things or activities are impossible. The last item is also likely to be keyed *true* because it involves a reasonable qualification.

**8. True statements and false statements should be approximately equal in length.**

There is a natural tendency for true statements to be longer because such statements must be precisely phrased in order to be absolutely true. This can be overcome according to Linn & Gronlund (2000:184) by lengthening the false statements through the use of qualifying phrases similar to those found in true statements. Thus the length of the statement will be eliminated as a possible clue to the correct answer.

---

9. *The number of true statements and false statements should be approximately equal.* Constructing a test with an approximately equal number of true statements and false statements will prevent response sets from unduly inflating or deflating the student's scores. Some students consistently mark statements "true" when in doubt about the answer, where as others consistently mark them "false". But the teacher should not consistently use exactly the same number true and false items; this will provide a clue to the student who is unable to answer some of the test items. The best procedure is to vary the percentage of true-false statements somewhere between 40 and 60 percent. Under no circumstances should the statements be all true or all false.

### **3. MATCHING EXERCISES**

Matching exercise is used to measure factual information, i.e., knowledge of facts, based on simple associations. Examples of relationships considered important (by teachers) include between persons and achievements; dates and historical events; terms and definitions; authors and titles of books; machines and uses; plants/animals and classifications and parts and functions.

#### **ADVANTAGES AND LIMITATIONS OF MATCHING EXERCISE**

As stated above matching exercises have their own advantages and disadvantages. What advantages and limitations do they have? Can state some of their weaknesses and strengths? In the following section we will look at the advantages and disadvantages.

#### **Advantages**

##### **1. *It is efficient***

The major advantage of the matching exercise is its compact form, which makes it possible to measure a large amount of related factual material in a relatively short time.

##### **2. *It reduces the effect of guessing***

Compared to multiple choice items it reduces the effect of guessing. However, it should be noted that as one goes from item to item unless using a response once, more than once or not at all is the rule the effect of guessing increases. In Example 2 below, the student has only a 1 in 10 chance of guessing correctly on the first item, a 1 in 9 chance for the second item, a 1 in 8 for the third, and so on. By the time the gets to the last item, there are still five alternatives from which to select. Obviously, the more options are provided, the less chance there will be to guess correctly.

##### **3. *Ease of construction and scoring***

Another advantage often cited by people for the matching exercise is its ease of construction and scoring. But this is not very true. Poor matching items can be rapidly constructed, but good matching items require a high degree of skill.

---

## Limitations

The following are the weaknesses of matching exercises.

### ***1. It is limited to measuring simple learning outcomes***

A major weakness of the matching exercise is that tends to ask students to associate trivial information. Unfortunately, most matching tests do emphasize memorization, although it is possible to construct items that measure more complex cognitive skills.

### ***2. It is highly susceptible to the presence of irrelevant clues***

If teachers are not careful in the construction of matching tests, there is high degree of including irrelevant clues to the responses that would lead students to the correct answers.

### ***3. Difficulty of obtaining homogenous materials***

Matching exercise is constructed based on a homogenous material to increase the difficulty level of the test. But it is usually difficult to find homogenous material that is significant from the view point of our objectives and learning outcomes.

## SUGGESTIONS FOR CONSTRUCTING MATCHING EXERCISES

### ***1. Use only homogeneous material in a single matching exercise***

This is the most important rule of constructing matching exercise and yet the one most commonly violated. One reason for this is that homogeneity is a matter of degree and what is homogeneous to one group may be heterogeneous to another. For example let's see the following exercise.

Consider the following matching exercise items.

Example 1.

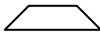

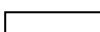


	Column A	Column B
_____	1. The king who introduced railroad service to Ethiopia.	A. 1974
_____	2. The Commander-in-chief of Army during Atse Tewodros's time.	B. Adwa
_____	3. Year Ethiopian revolution took place.	C. Atse Menilik
_____	4. The place Ethio-Italian war took place.	D. Gebrye
_____	5. The first country that made diplomatic relationship with Ethiopia.	E. Italy

Although this is a matching exercise, it does not contain a homogeneous material. Some items deal with names of people, and other items deal with historical events. To obtain homogenous material, it is necessary to have only inventors and their inventions in one

matching exercise, explorers and their discoveries in another and kings or presidents and their achievements in another. If matching exercises are not kept homogeneous, the items are likely to test only the simplest associations and to provide many commonsensical clues to the correct answer. Look the following matching exercise.

### Example 2

**Direction:** In column A are five diagrams. In column B are their names of different figures. Match the name of the figure with the diagram that best shows it by placing the letter in Column B on the appropriate space under Column A. Each option under Column B may be used once, more than one or not at all.

COLUMN A	COLUMN B
_____ 1. 	A. Circle
_____ 2. 	B. Cone
_____ 3. 	C. Cube
_____ 4. 	D. Cylinder
_____ 5. 	E. Parallelogram
	F. Pyramid
	G. Rectangle
	H. Sphere
	I. Square
	J. Trapezoid

This matching exercise is more homogenous than the above one. This is because it deals with figure and their names. The chances are very low to get the correct by guessing.

### 2. *Include an unequal number of responses and premises, and instruct the student that responses may be used once, more than once, or not at all.*

When an equal number of responses and premises are used and each response is to be used only once, the probability for guessing the remaining responses correctly is increased each time a correct answer is selected and the final response can be selected entirely on the basis of this process of elimination. On the other hand, in this type of matching if an error is made in one match, there is certain to be an error in another. Using more responses than premises eliminates these potential dangers of perfect matching. Or we can make use of more premises than responses. In either case the directions should instruct the student that each response might be used once, more than once or not at all. So avoid equal number of responses and premises in matching exercises. Instead use unequal number of responses and premises in matching exercises. This will make all the responses eligible for selection for each premise and will decrease the likelihood of successful guessing.



---

**3. *Place the shorter responses in Column B, i.e., on the right***

Putting the shorter responses in column B is time saving. This timesaving practice allows the student to read the longer item first in column A and then the search quickly through the shorter options to locate the correct alternative.

**4. *Use limited number of the items within each set***

A brief list of items is advantageous to both the teacher and the student. From the teacher's standpoint, it is easier to maintain homogeneity in a brief list. From the student's view point a brief list enables them to read the responses rapidly and without confusion. Approximately four to seven items in each column seems best. There certainly should be no more than ten items in either column.

**5. *Arrange the list or responses in logical order: place words in alphabetical order and numbers in sequence.***

If options are organized alphabetically or numerically, students do not waste time searching for the correct response. This is especially important if there are many options.

**6. *Provide complete directions***

Although the students basically know what to do in matching exercises and basis for matching is rather obvious in most matching exercises, there are advantages in clearly stating it. First ambiguity and confusion will be avoided. Second, testing time will be saved because the student will not need to read through the entire list of premises and responses and then "reason out" the basis for matching.

**7. *Place all the items in one matching exercise on the same page.***

This will prevent the disturbance created by the students switching the pages of the test back and forth. It also will prevent them from missing the responses appearing on another page and generally adds to the speed and efficiency of test administration.

**4. WRITING MULTIPLE CHOICE ITEMS**

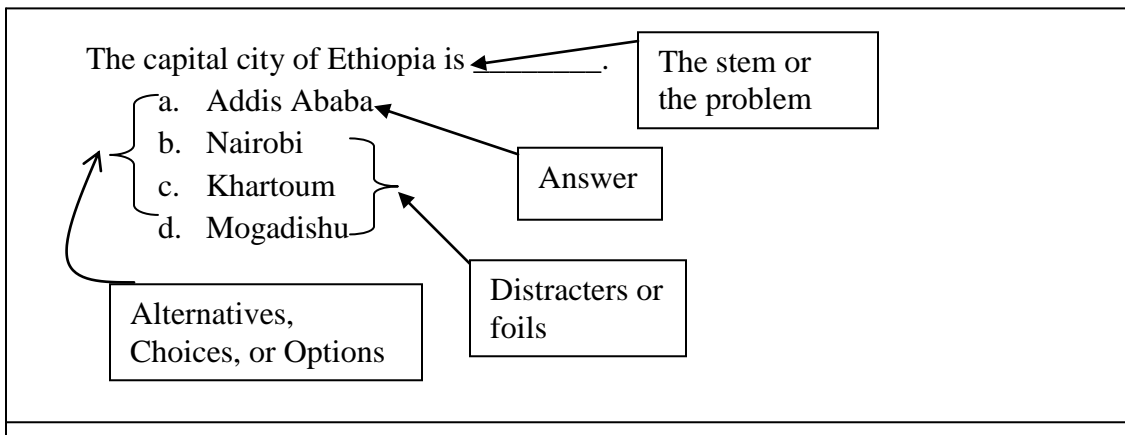
The multiple choice item is the most flexible objective item types. It has great versatility in measuring objectives from the rote knowledge level to the most complex level except for synthesis.

According to Ebel & Frisbie (1991), multiple choice items have long been the most highly regarded and widely used form of objective test items. They are adaptable to the measurement of most important educational outcomes of knowledge, understanding and judgment, ability to solve problems, and ability to make predictions. Almost any understanding or ability that can be tested by means of any other item form, short answer, completion, true-false, matching or even essay, can also be tested by means of multiple

choice items. This flexibility added with higher quality items usually found in the multiple choice form, has led to its extensive use in achievement testing

A multiple-choice item consists of two parts: the stem, which presents the problem, and a list of possible or suggested answers or options. The problem can be stated as a direct question, i.e., in a question form, or an incomplete statement. The list of suggested answers or solutions may include words, numbers, symbols, phrases and are called alternatives (also called choices or options). The correct alternative in each item is called the *answer*, and the remaining alternatives are called *distracters* (also called *decoys* or *foils*). These incorrect alternatives receive their name from their intended function - to distract those students who are in doubt about or who do not know the correct answer.

### Parts of multiple choice items



In writing the stem of the item, a direct question or an incomplete statement can be used. A direct question is easier to write, is more natural for students, and is more likely to present a clearly formulated problem. On the other hand, the incomplete statement is more concise and if skillfully phrased, it too can present a well-defined problem. A common procedure is to start each stem as direct question and shift to the incomplete statement form only when the clarity of the problem can be retained and greater conciseness achieved.

Multiple choice items can be of a correct answer type or a best answer type. In case of the correct answer type there will be one absolutely correct response and it mainly measures simple aspects of knowledge represented by questions of 'who', 'what', 'when' and 'where' varieties.

### Example

What is the largest planet in the universe?

- A. Earth
- B. Mars

- 
- C. Jupiter
  - D. Neptune

In the above multiple choice item, it is only Jupiter that can be chosen as an answer. Thus the item is correct answer type.

In other cases answers of varying degrees of acceptability may be the rule. For example, questions of the ‘why’ variety tend to reveal a number of possible reasons, some of which are better than the others. Likewise question of the ‘how’ variety usually reveals possible procedures, some of which are more desirable than the others. Therefore, measures of achievement in these areas become a matter of selecting the best answer. Look at the following multiple choice item.

Why does the planet Mercury have a year of 88 Earth days?

- a. Mercury’s year is shorter than Earth’s.
- b. Mercury’s orbit is closer to the sun than the Earth’s. *Answer*
- c. Mercury’s small size and elliptical orbit make it travel faster than Earth.

In this example, option “a” is true, but it is not the best alternative because it fails to answer the question posed in the stem; option “c” is true *in part* (Mercury is small and has a elliptical orbit) but does not explain its orbital speed as well as does option “b”.

## USES OF MULTIPLE CHOICE ITEMS

The multiple choice item is the most versatile type of test item available. It can measure a variety of learning outcomes from simple to complex and it is adaptable to most types of subject matter content. Multiple Choice items can be employed for measuring common learning outcomes in the areas of knowledge, understanding, and application.

### A. Measuring Knowledge Outcomes

Learning outcomes in the knowledge area are so prominent in all school subjects, and multiple choice items can measure such a variety of these outcomes. Let us look at some of the more typical uses with this respect to measuring knowledge objectives.

#### i. Knowledge of terminology

Here students can be requested to show their knowledge of a particular term by selecting a word that has the same meaning as the given term or by choosing a definition of the term. Special uses of the term can also be measured by having students identify the meaning of the term when used in context.

#### *Examples*

1. Which one of the following words has the same meaning as the word *egress*?

- 
- A. Depress
  - B. Enter
  - C. *Exit*
  - D. Regress
2. Which one of the following statements best defines the word ***egress***?
- A. An expression of disapproval
  - B. *An act of leaving an enclosed place*
  - C. Proceeding to a higher level
  - D. Proceeding to a lower level
3. What is meant by the word ***egress*** in the following sentence: “The astronauts hope they can now make a safe ***egress***”
- A. *Separation from rocket*
  - B. Re-entry to the earth's atmosphere
  - C. Landing on the water
  - D. Escape from the space capsule.

## **ii. Knowledge of specific facts**

Multiple choice items designed to measure specific facts can take many different forms, but questions of the ‘who’, ‘what’, ‘when’ and ‘where’ varieties are most common.

### ***Examples***

1. Who was the king of Ethiopia that people commonly say was ahead of his time?
- A. Tewodros
  - B. Menilik
  - C. Yohannes
  - D. Hailelassie
2. When did the battle of Adwa take place?
- A. 1896 E.C.
  - B. 1888 E.C
  - C. 1890 E.C
  - D. 1892 E.C

## **iii. Knowledge of principles**

The items can be constructed to measure knowledge of principles as easily as those designed to measure facts.

### ***Example***

- The principle of capillary action helps explain how fluids
- A. enter solutions of lower concentration

- 
- B. escape through small openings
  - C. pass through semi-permeable membranes
  - D. rise in fine tables

**iv. Knowledge of methods and procedures.**

This includes such diverse areas as: knowledge of laboratory procedures; knowledge of methods underlying communication, computational and performance skills, knowledge of methods used in problem solving, knowledge of governmental procedures, etc.

***Example***

Alternating electric current is changed to direct current by means of a

- A. condenser
- B. rectifier
- C. generator
- D. transformer

**B. Measuring outcomes at the understanding and application levels**

Although it is difficult to go beyond the knowledge level with most of the other types of objective items, the multiple choice item is especially adaptable to the measurement of more complex learning outcomes. But here we have to know that such items will measure learning outcomes beyond factual knowledge only if the applications and interpretations are new to the students. To measure understanding and application, an element of novelty must be included in the test item.

**i. Ability to identify application of facts and principles**

A common method of determining whether students' learning has gone beyond the mere memorization of a fact or principle is to ask them to identify its correct application in a situation that is new to the student. Application items measure understanding but they also include the ability to transfer learning to situations that have not been previously studied.

***Example***

Which of the following would result in the greatest reduction of calories if it were eliminated from the daily diet?

- a) 1 tablespoon of butter
- b) 1 tablespoon of granulated sugar
- c) 1 slice of white, enriched bread
- d) 1 boiled egg

**ii. Ability to interpret cause and effect relationships**

Understanding can frequently be measured by asking students to interpret various relationships among facts.

---

*Example*

An increased quantity of carbon dioxide is produced when fuel is burned in a limited supply of oxygen because

- a) carbon reacts with carbon monoxide
- b) carbon reacts with carbon dioxide
- c) carbon monoxide is an effective reducing agent
- d) greater oxidation takes place

**iii. Ability to justify methods and procedures**

Here the student is asked to select the best of several possible explanations of a method or procedure.

*Example*

Why do farmers rotate their crops?

- a) To conserve the soil
- b) To make marketing easier
- c) To provide far strip cropping
- d) It removes the brownish yellow

**ADVANTAGES AND LIMITATIONS OF MULTIPLE CHOICE ITEMS**

**Advantages**

Multiple choice item has a number of advantages. Some of them are the following.

**1. It is flexible**

It is one of the most widely applicable test items for measuring achievement. It can effectively measure various types of knowledge and complex learning outcomes.

**2. It is free from the common weaknesses of the other type items**

For example the ambiguity and vagueness that frequently are present in the short answer item are avoided because the alternatives better structure the situation

*Examples*

Short answer type item

African union was established in \_\_\_\_\_. ( poor)

If we change this item into multiple choice item form, it will be as follows.

African union was established in

- A. South Africa

- 
- B. Ethiopia
  - C. Libya
  - D. Algeria

Can you see how the problem of the short answer type is avoided by the alternatives of the multiple choice item? This is to mean that multiple choice items clarify the specific type of response called for.

**3. *It avoids the problem of spelling errors by students***

In multiple choice items students are required to select one answer from the given list of alternatives. This avoids the problem of deciding how to score misspelled answers because students are not supposed to supply responses. In addition, Multiple Choice item relieves the problem.

**4. *Students should know the answer***

One advantage of multiple choice item over True-False item is that students cannot receive credit for simply knowing that a statement is incorrect: they must also know what is correct.

*Examples*

T      F      African union was established in Algeria.

African union was founded in\_\_\_\_\_.

- A. South Africa
- B. Ethiopia
- C. Libya
- D. Algeria

In the true-false item above a student may get it right if he responds *false*. But the problem is we are not sure whether the student knows that the union was founded in South Africa. In the multiple choice format, in contrast, unless the student knows that the union was founded in South Africa he/she will not get it right.

**5. *Multiple choice items have a greater reliability per item***

When compared to true-false items, multiple choice items have a greater reliability per item because the number of alternatives is increased from two to four or five. As a result of this the opportunity for guessing the correct answer is reduced, and the reliability is correspondingly increased. According to Linn & Gronlund (2000) the effect of increasing the number of alternatives for each item is similar to that of increasing length of the test.

**6. *The need for homogeneous material is minimized or avoided***

An advantage of the multiple choice item over the matching exercise is that the need for homogenous material is minimized or avoided. In many content areas, it is difficult to obtain

---

enough homogeneous material to prepare effective matching exercises. But this problem is avoided with multiple choice items because each item measures a single idea.

***7. It is relatively free from response sets***

That is, students generally do not favor a particular alternative when they do not know the answer. In True-False, item response set is more because the mere alternatives are always true or false.

***8. Using a number of plausible alternatives makes the results useful in diagnosing students' learning errors***

This means the kind of incorrect alternatives students select provides information on students' misunderstandings.

**Limitations**

***1. It is limited to the measurement of verbal material***

Multiple choice items do not measure students' ability to do something or to perform a task. They do not measure their skills. As with all other paper and pencil tests, it is limited to learning outcomes at the verbal level. The problems presented to students are verbal problems, i.e., they are given written test papers and answer by writing their responses.

***2. It is unsuitable to measure synthesis and evaluation levels of the cognitive domain***

Like other types of selection items, it is not well adapted to measuring some problem solving skills and the ability to organize and present ideas

***3. Difficulty of getting plausible distracters***

There is difficulty of getting sufficient number of incorrect but plausible distracters. The purpose of distracters is to distract the unformed student from getting the correct answer. To do that the distracters should be equally attractive. But finding distracters that serve this purpose is very difficult. Specially obtaining plausible distracters is acute at the early primary levels because of their limited vocabulary and the problem becomes less as students go up grade levels when their vocabulary improves.

**4.4.3. SUGGESTIONS FOR CONSTRUCTING MULTIPLE CHOICE ITEMS**

The general applicability and the superior quality of multiple choice test items are realized most fully when care is taken in their construction. This involves formulating a:

- clearly stated problem,
- identifying plausible alternatives, and
- removing irrelevant clues to the answer.

The following suggestions provide more specific guidelines for this purpose.



---

**1. *The stem of the item should be meaningful by itself and should present a definite problem.***

In other words, students should have tentative answers after reading the stem only. Often the stems of test items placed in multiple-choice form are incomplete statements that make little sense until all of the alternatives have been read. A properly constructed multiple choice item presents a definite problem in the stem that is meaningful without the alternatives.

*Example*

- |        |   |
|--------|---|
| Poor   | Ethiopia is                                 |
|        | A. never colonized.                         |
|        | B. found in eastern Africa.                 |
|        | C. the seat of African Union.               |
|        | D. rich in natural resources.               |
| Better | Which country is the seat of African Union? |
|        | A. Sudan                                    |
|        | B. Kenya                                    |
|        | C. Ethiopia                                 |
|        | D. Mozambique                               |

In the first item, the alternatives are concerned with widely dissimilar ideas. This heterogeneity is possible because of the stem's lack of structure. In the second item, the clearly formulated problem in the stem forces the alternatives to be more homogeneous. A good check on the adequacy of the problem statement is to cover the alternatives and read the stem by itself. It should be complete enough to serve as a short answer item.

**2. *The item stem should include as much of the item as possible and should be free of irrelevant material***

This will increase the probability of a clearly stated problem in the stem and will reduce the reading time required. It is possible to increase the conciseness of an item by removing irrelevant materials and those words repeated in the alternatives.

*Example*

- |         |  |
|---------|--|
| Poor    | The period and the group to which chemical elements belong can be easily determined by its electronic configuration. If an element has an atomic number of 20 then |
|         | A. It is in the second period.   |
|         | B. It is in the third period.  |
|         | C. It is in the fourth period.   |
|         | D. It is in the fifth period.  |
| Better: | If an element has an atomic number of 20 then it is in the _____ period.   |
|         | A. second  |

- 
- B. third
  - C. fourth
  - D. fifth

Best If an element has an atomic number of 20, then in what period is it?

- A. second
- B. third
- C. fourth
- D. fifth

In the poor multiple choices above the stem has been made unnecessarily long due to the inclusion of irrelevant material. In addition, in all of the alternatives there is unnecessary repetition of terms. That is, “It is in the” has been repeated in all of the alternatives.

There are few exceptions to this rule. In testing problem solving ability, irrelevant material might be included in the stem of an item to determine whether students can identify and select the material that is relevant to the problem’s solution. Similarly, repeating common words in the alternatives is sometimes necessary for grammatical consistency or greater clarity.

### ***3. Construct stems and options that are stated positively***

Stems and options should be stated positively whenever possible. Most problems can and should be stated in positive terms. This avoids the possibility of students’ overlooking ***no***, ***not***, ***least***, and similar words used in negative statements. Stating items in positive terms also avoids measuring relatively insignificant learning outcomes. Knowing the least important method, the principle that does not apply, or the poorest reasons are seldom important learning outcomes. We are usually interested in students’ learning of the most important method, the principle that does apply, and the best reason.

#### ***Example 1***

Poor Which one of the following cities is not found South of Addis Ababa?

- a. Shashemene
- b. Awasa
- c. Bahir Dar
- d. Nazareth

Better Which one of the following cities is found north of Addis Ababa?

- a. Shashemene
- b. Awasa
- c. Bahir Dar
- d. Nazareth

#### ***Example 2***

Poor Which of the following is not a characteristic of gifted children?

- 
- a. They are emotionally stable.
  - b. They are not awkward.
  - c. They are not as old as their classmates.
  - d. They are friendly.

Better      Which of the following is a characteristic of gifted children?

- a. They tend to have many emotional problems.
- b. They are awkward.
- c. They are younger than their classmates.
- d. They are unfriendly.

In the above examples, the poor and better versions of the items measure the same knowledge. But the students may overlook the negative marker, i.e., the word “not”, in the stem and the options of the poor items.

Although negatively stated items are generally to be avoided, there are occasions when they are useful mainly in areas in which the wrong information or wrong procedure can have serious consequences. In the health area, for example there are practices to be avoided because of their harmful nature. In shop and laboratory work, there are procedures that can damage equipment and result in bodily injury. And in driver training there are safe practices to be emphasized. When the avoidance of such potentially harmful practices is emphasized in teaching, it might well receive a corresponding emphasis in testing through the use of negatively stated items. When used, the negative aspects of the item should be made obvious by making it bold, capitalized, underlined, italicized, etc.

*Example*

Poor      Which one of the following is not a safe driving practice on icy roads?

- A. Accelerating slowly
- B. Jamming on the brakes
- C. Holding the wheel firmly
- D. Slowing down gradually

Better      All of the following are safe driving practices on icy roads EXCEPT

- A. Accelerating slowly
- B. Jamming on the Brakes
- C. Holding the wheel firmly
- D. Slowing down gradually

In the first version the ‘not’ can easily be overlooked. In the second example, no student would probably overlook the negative element because it is placed at the end of the statement and is capitalized.

---

4. ***All alternatives should be grammatically consistent with the stem of the item*** This rule is not presented merely to perpetuate proper grammar usage, however its main function is to prevent irrelevant clues from creeping in.

*Example*

- Poor    An electric transformer can be used
- A. for storing electricity.
  - B. to increase the voltage of alternating current.
  - C. it converts electrical energy into mechanical energy.
  - D. Alternating current is changed to direct current.
- Better    An electric transformer can be used to
- A. store electricity
  - B. increase the voltage of alternating current.
  - C. convert electrical energy in to mechanical energy
  - D. change alternating current

In the poor example above, alternatives C and D are not grammatically consistent with the stem. So, they will be easily eliminated by students resulting only in two plausible alternatives. Similar difficulties arise from a lack of attention to verb tense, to the proper use of the articles ‘a’ or ‘an’, and to other common sources of grammatical inconsistency.

5. ***An item should contain only one correct or clearly best answer***

Unless students are given directions, to the contrary, they will presume that only one answer is the correct or the best one for the multiple choice items. In best answer type items care must be taken to make certain that the answer is clearly the best one.

*Example*

- Poor:    Most fatalities are due to
- A. acts of God
  - B. automobile accidents
  - C. home accidents
  - D. jobs
  - E. old age
- Better:    Most accidental deaths occur in which of the following places?
- A. Automobiles
  - B. Homes
  - C. Jobs (excluding auto and home accidents)
  - D. Schools

In the poor example, can old age be considered a fatality? Are parents at a job when at home? Are home accidents acts of God? To avoid pitfalls (difficulties) each option should be

---

examined to make sure it is either the most defensible answer or clearly wrong. Being able to justify the reasons for incorrect options is as important as being able to defend the correct one.

**6. *Use novel materials in formulating problems that measure understanding or ability to apply principles. But beware of too much novelty***

The construction of multiple choice items that measure understanding requires a careful choice of situations and skilful phrasing. The situations must be new to the students but not too far removed from the examples used in class. If the test items contain problem situations that are identical with those used in class, the test is not going to be test of understanding or application. It rather becomes a test of memorized facts. On the other hand, if the problems are extremely new to students, some students may respond incorrectly merely because they lack necessary factual information about the situations used. This problem of too much novelty usually can be avoided by selecting situations from the students' everyday experiences.

*Example*

If you wish to determine the angle measure of a certain plot of land to be  $90^0$  where you have a rope but not a tri-square or other measuring devise, which one of the following principles you may use?

- A. Pythagoras theorem.
- B. Einstein's relativity theory.
- C. Archimedes principle
- D. Newton's third law

**7. *All distracters should be plausible***

As stated earlier the purpose of a distracter is to distract the uninformed from the correct answer. To the student who has not achieved the instructional objectives, the distracters are supposed to be at least as effective as the correct answer and preferably more so to attract him or her. Distracters are included with the assumption that they will be selected by some students. If a distracter is not selected by any one, it is useless thus need to be eliminated or revised. One factor contributing to the plausibility of distracters is their homogeneity. If all of the alternatives are homogeneous with regard to the knowledge being measured, the distracters are more likely to function as intended.

*Example*

- Poor      The concept of inferiority complex was contributed by
- A. Adler.
  - B. Freud.
  - C. Marx.
  - D. Lincoln.

- 
- |        |   |
|--------|---|
| Better | The concept of inferiority complex was contributed by |
|        | A. Adler.   |
|        | B. Freud.   |
|        | C. Jung.  |
|        | D. Horney.  |

In the example above, the first multiple choice item is poor because the alternatives are not plausible. To explain, while Adler and Freud are both personality psychologists and in fact, psychoanalysts, Marx and Lincoln were not. Marx and Lincoln do not belong to the field of psychology. Marx was a philosopher and Lincoln was rather politician. In contrast, the second item consists of list of names of psychodynamists (psychoanalysts).

### ***Example 2***

- |        |   |
|--------|---|
| Poor   | The number of photoreceptors in the retina of each human eye is about |
|        | A. 100,000  |
|        | B. 2 million  |
|        | C. 115 million  |
|        | D. 2.37 billion   |
| Better | The number of photoreceptors in the retina of each human eye is about |
|        | A. 5 million  |
|        | B. 35 million   |
|        | C. 65 million   |
|        | D. 115 million  |

In the second example the first multiple choice item is followed by implausible distracters. This is because the numbers vary in units: one in thousands another in billion which are not likely to be equally attractive. In the better version, units are made the same: all are indicated in millions only.

In selecting plausible distracters, the students' learning experiences must not be ignored. Obviously, distracters must be familiar to students before they can serve as reasonable alternatives.

### ***8. Avoid unintentional clues to the correct answer.***

Inexperienced test constructors frequently give away the correct answer to an item or give clues that permit the examinee to eliminate one or more of the incorrect answer choices from consideration. Items that contain

- irrelevant clues,
- specific determiners,
- correct answers that are consistently longer than incorrect answers, or
- grammatical inconsistencies between the stem and the wrong alternatives

---

tend to be easier than items without these faults.

Example items containing some types of unintentional clues are given below. Item 1 is an example of a clang association, that is, a repetition of a word, a phrase, or sound in the keyed answer and in the stem. Item 2 contains specific determiners that have the same effect in multiple choice options as in true-false statements. In item 3, the keyed answer is much longer than the other options. Item 4 is an example of a grammatical inconsistency; the word “a” in the stem implies a singular word, but Options A and C are both plural. The revised items show how each of these faults can be corrected to make the items more effective in measuring knowledge rather than test-wiseness.

### ***Examples***

#### Clang Association

Poor: 1. Which one of the following instruments is used to determine the direction of wind?

- A. Anemometer
- B. Barometer
- C. Hygrometer
- D. Wind Vane (*Answer*)

Better: Which one of the following instruments is used to determine the speed of wind?

- A. Anemometer (*Answer*)
- B. Barometer
- C. Hygrometer
- D. Wind Vane

Specific Determiners (i.e., never, always, etc.)

Poor: 2. Which of the following is characteristic of anaerobic bacteria?

- A. The **never** live in soil.
- B. They can live without molecular oxygen.
- C. They **always** cause disease.
- D. They can carry on photosynthesis.

Better: One characteristic that distinguishes all anaerobic bacteria is their ability to

- A. withstand extreme variation in air temperature.
- B. live without molecular oxygen. (*Answer*)
- C. live as either saprophytes or parasites.
- D. reproduce either in living cells or nonliving culture media.

#### Length Clues

Poor: 3. The term *side effect* of a drug refers to

- A. additional benefit from the drug.
- B. the chain effect of drug action.
- C. the influence of drugs on crime.

- 
- D. any action of the drug in the body other than the one the doctor wanted the drug to have. (*Answer*)
- Better: Which of the following, if it occurred, would be a side effect of aspirin for a man who had been taking two aspirin tablets every 3 hours for a heavy cold and slight fever?
- A. Normal body temperature
  - B. Reduction in the frequency of coughing
  - C. Easier breathing
  - D. Ringing in the ears. (*Answer*)
- Grammatical Inconsistency
- Poor 4. Penicillin is obtained from a
- A. Bacteria.
  - B. mold. (*Answer*)
  - C. coal tars.
  - D. tropical trees.
- Better Penicillin is obtained from
- A. bacteria.
  - B. molds. (*Answer*)
  - C. coal tars.
  - D. topical trees.

**9. *The correct answer should appear in each of the alternative positions an approximately equal number of times but in random order.***

We should have ‘a’ answers, ‘b’ answers, ‘c’ answers, and ‘d’ answers and if possible an approximately equal number of times but in random order.

If 40 multiple choice items with 4 alternatives are used, it is suggested to have approximately 10 a’s, 10 ‘b’s, 10 ‘c’s and 10 ‘d’s. But the order of the answers should be random to avoid getting the correct answer by some pattern. A way of placing correct answers in all the alternatives randomly is use of right side page numbers of books. The right sides of books are given page numbers that are odd. The following is the procedure of using them. You open any book randomly and look at the page number on the right side. If you see the last digit to be 1 the answer for the item should be placed at alternative A, if the last digit is 3, the answer for the item should be placed at alternative B, if the last digit is 5, the answer for the item should be placed at alternative C, if the last digit is 7, the answer for the item should be placed at alternative D, and if the last digit is 9 the answer for the item should be placed at alternative E.

**10. *Use sparingly “none of the above” and “all of the above” as alternatives***

The phrases “none of the above” or “all of the above” are sometimes added as the last alternatives in multiple choice items. This is done to force the student to consider all of the



---

alternatives carefully and to increase the difficulty of the items. All too frequently, however, those special alternatives are used inappropriately. In fact, there are certain situations in which their use is appropriate.

The use of “none of the above” is restricted to correct answer type multiple choice items and consequently to the measurement of factual knowledge to which absolute standards of correctness can be applied. It is inappropriate in best answer type items because the student is told to select the best of several alternatives of varying degrees of correctness. Use of “none of the above” is frequently recommended for items measuring computational skill in mathematics and spelling ability.

When “all of the above” is used, some students will note that the first alternative is correct and select it without reading further. Other students will note that at least two of the alternatives are correct and thereby know that “all of the above” must be the answer. In the second instance, students obtain the correct answer on the basis of partial knowledge and in the first instance, students mark the item incorrectly because they do not read all of the alternatives. In stead it is suggested that you use combined alternatives like “A and B are correct”, “B and D are correct”, etc.

## **5. THE ESSAY TESTS**

### **The Nature and Types of Essay Items**

According to Linn & Gronlund (2000) essay tests allow for freedom of response. Students are free to select, relate, and present ideas in their own words. But the freedom is a matter of degree. In some instances that freedom is delimited to specific size. In other cases, no restriction is put. So, based on the extent of freedom essay tests can be classified into restricted essay tests and extended response essay tests.

#### **1. Restricted Response Essay Questions**

These questions usually limit both the content and the form of the response. The content is usually restricted by the scope of the topic to be discussed. Limitations on the form of the response are generally indicated in the question.

#### ***Example***

- a. Why is multiple choice item considered the most versatile type? Answer in a brief paragraph.
- b. Describe two situations that demonstrate the application of the Newton's third law of motion. Do not use those examples discussed in class.

Although delimiting students' responses to essay questions makes it possible to measure more specific learning outcomes, these same restrictions make them less valuable as a

---

measure of those learning outcomes emphasizing integration, organization, and originality. This is because for higher order learning outcomes, greater freedom of response is needed.

### **E. Extended Response Essay Questions**

In this type of test no restriction in either form or content are placed. Students can provide answers by organizing their ideas the way they like. Moreover, it is the student himself/herself that determines the size of the answer. However, in spite of the fact that this freedom allows for the measurement of higher order skills, scoring difficulties come into play. The following are examples of extended response essay tests.

#### ***Example***

- a. Describe the influence of textbooks on sex stereotyping.
- b. Write your own evaluation of the value of the New Pre-service Teacher Education System Overhaul (TESO) in preparation of qualified or well trained secondary school teachers.

## **ADVANTAGES AND LIMITATIONS OF ESSAY TEST**

### **Advantages**

#### ***1. They can measure complex learning objectives***

Essay type items measure complex learning outcomes that cannot be measured by other objective type items. They also can measure divergent thinking. Divergent thinking is indicated by unconventional, creative, and relatively rare responses. Because essay tests allow great freedom in responding, the opportunity for obtaining unusual responses is increased. Extended response questions emphasize on the integration and application of thinking and problem solving skills.

#### ***2. They allow free response***

Essay tests give students the freedom to respond within broad limits. Although recognition of good writing skills can be measured by multiple-choice test, the teacher's objectives may call for the student to write. The essay test allows students to express their ideas. They are the most direct measurement of writing skills. According to Linn & Gronlund (2000) in some cases, the evaluation of specific writing skills may be combined with the assessment of subject matter knowledge and understandings (e.g. communication of mathematical or scientific principles, ideas, or concepts). In other cases, the assessment of writing skills may be the only assessment area (e.g. skills in developing characters in a narrative story or writing mechanics).

#### ***3. They eliminate guessing***

Because essay achievement tests provide no options for students to select from, guessing is eliminated. The student must supply rather than select the proper response.

Compared to some objective type items it is easy to construct.

---

#### **4. *They easy to construct***

Constructing essay items is relatively easy. The teacher can construct essay items within a limited time range.

#### **5. *They have a desirable effect on the learners study habits.***

Essay items encourage students to study hard because they know that they are required to write rather than select a single answer. To write an answer to a given item, a student has to be prepared well.

### **Limitations**

#### **1. *Scoring is Unreliable***

Essay responses are difficult to score objectively because the student has greater freedom of expression. Also, long, complex essays are more difficult to score than shorter, more limited ones.

Different scorers assign different scores for the same response to an essay test. Even the same teacher might score the same test paper at different times. It requires an extensive amount of time to read and grade.

#### **2. *They are time consuming***

Essay items are time consuming for both the teacher and student. Students often spend much time answering only one or two extended essay questions, which may severely limit sampling of their knowledge. Teachers also devote many hours to reading lengthy responses.

#### **3. *They cannot measure a large amount of content or objectives.***

It provides limited sampling of content. So few questions can be included in a given test that some areas are measured thoroughly, but many others are neglected. As a result of this they are inefficient measures of factual information.

#### **4. *They are subject to bluffing***

Although essay tests eliminate guessing, they do not prevent bluffing. Poorly prepared students often attempt to get a passing grade by answering *something*, even if the responses are unrelated to the questions asked.

### **SUGGESTIONS FOR CONSTRUCTION**

Linn & Gronlund (2000) said that use of essay questions as a measure of complex learning outcomes necessitates giving attention to two important points;

1. how to construct essay tests, and
2. how to score them reliably.

Accordingly, they suggest the following when constructing essay tests.

---

***1. Restrict the use of essay questions to those learning outcomes that cannot be measured satisfactorily with objective items.***

That is, if it is possible to test a certain learning outcome using objective test items essay test should not be used. This is so because, other things being equal, objective measures have the advantage of efficiency and reliability. You have to use essay test only when objective test items cannot measure the behavior of interest.

***2. Construct questions that will call forth the behavior specified in the learning outcomes***

Like objective items, essay questions should measure the achievement of clearly defined instructional objectives. Because they are easy to construct it may happen that essay tests might be constructed without giving attention to the specific learning outcomes they are intended to measure. Hence, great care must be taken when constructing essay tests. If the ability to apply principles is being measured, for example, the questions should be phrased in such a manner that they call forth that particular behavior.

***3. Phrase the question so that the student's task is clearly indicated.***

Like the case in objective test items, essay tests should be developed in such a way that they are not ambiguous (i.e. they should be clear). If they are ambiguously stated, they will be problems both to the students and to the teacher. Students will have to guess what was in the teacher's mind when he/she constructed the item. Because students will come with variety of responses for they may see the question in different ways, the teacher will have hard time of scoring. Look at the following pair of essay questions.

Example

Poor: How does human kind come to this world?

Better: Explain from the viewpoint of evolution how human kind came to this world.

In the first version whether the student should depend on some evolutionary viewpoint or provide his/her personal justification is not known. In contrast, the second item specifies how the student should attack the problem.

Even in extended response items, it is preferable to offer students a common basis for responding to the question. Otherwise, the teacher himself/herself will lack frame of reference for evaluating the response. Look at the following examples.

Poor: Compare Coalition for Unity and Democracy, and the Ethiopian Peoples Revolutionary Democratic Front.

---

Better: Compare the land policies of Coalition for Unity and Democracy, and the Ethiopian Peoples Revolutionary Democratic Front in view of resulting in sustainable development.

The first essay item is too broad and can be seen from any angles with no common frame of reference. The second one in contrast while still preserving freedom of response guides students how to approach it.

**4. *Indicate an approximate time limit for each question.***

It is suggested that time should not be over when students are one or more questions to go. Therefore, as each question is written we should estimate the approximate time needed for a satisfactory response. We have also to take into account slower students when we determine the time. It is better to use fewer questions and give more generous time limits than to put some students at a disadvantage.

Students often waste time unnecessarily on some questions and leave others unattempted because of shortage of time. Therefore, to minimize this problem the time limits allotted to each question should be indicated to the students. This will help students adjust their pace to each question of the test. The students should also be told approximately how much time to spend on each part of the test either orally or on the test paper it self.

**5. *Avoid the use of Optional Questions.***

Teachers commonly give students more essay test items that they are supposed to answer. For example, a teacher may construct five essay questions and direct students to pick any three of them and write their answer. Hence, students will be in a position to select those questions they know most. If student answer different questions, it is clear that they are taking different tests, and the common basis for evaluating their achievement is lost. Then ultimately, it will be possible to say each student is demonstrating achievement of different learning outcomes.

**SCORING ESSAY ITEMS**

One of the most serious problems with essay items is *unreliability in scoring*. There are some ways through which this unreliability could be minimized. The following are list of suggestions.

**1. *Prepare an outline of the expected answer in advance.***

The outline you prepare should contain the major points to be included, the characteristics of the answers (e.g., organization) to be evaluated and the amount of credit to be allotted. In case of restricted response items, you can have a list of acceptable responses. For example, if students are given an aim statement form Ethiopian Training and Education Policy and are required to derive three goal statements of their own, as your scoring guide you can prepare

---

three acceptable goal statements. In case of extended response essay tests you can have an outline of major points to be emphasized in students' responses. The outline may include

- accuracy of factual information,
- relevance of examples,
- coherence of paragraphs, etc.

In addition to this you have to determine and make them the weights given to each of the components known to students.

Preparing a scoring key provides a common yardstick for evaluating students' answers and increases the consistency of our standards for each question throughout the scoring. If prepared during the test construction, such scoring key also helps us phrase questions that clearly specify the types of answers expected.

## ***2. Use the scoring method that is most appropriate.***

There are two common methods of scoring essay questions: analytical method and holistic method.

### **Analytical scoring method**

In the analytic method, each answer is compared with the ideal answer in the scoring key, and a given number of points are assigned according to the adequacy of the answer. It enables the teacher to focus on one characteristic of a response at a time. Examples for analytic scoring rubrics may include organization, word choice, content, etc. In analytic scoring thus specific feedback can be given to the testee (student). The following Northwestern Region Educational Laboratory (NWREL) and Gearhart, Herman, Baker, & Whittaker analytic scoring rubrics are descriptions of analytic scoring rubrics taken from Linn & Gronlund (2000).

#### **Example of the NWREL Scoring Rubric for a Score of 5 on the Organization Dimension**

"The organization enhances and showcases the central idea or storyline. The order, structure, or presentation of information is compelling and moves the reader through the text."

- "Details seem to fit where they are placed: *sequencing is logical and effective.*"
- "*An inviting introduction* draws the reader in: *a satisfying conclusion* leaves the reader with a sense of resolution."
- "*Pacing is well controlled*; the writer knows when to slow down and elaborate, and when to pick up the pace and move on"
- "*Thoughtful transitions* clearly show how ideas connect."
- "Organization *flows so smoothly* the reader hardly thinks about it."

---

**Table 5.1 Analytic Scales for Expository Essays or Descriptive Summaries**

Score	<i>General Impression</i>	<i>Focus/ Organization</i>	<i>Language</i>	<i>Elaboration</i>	<i>Mechanics</i>
<b>6</b>	Exceptional Achievement	<ul style="list-style-type: none"> <li>Clearly stated main idea</li> <li>Unified focus and organization</li> <li>Effectively orients reader</li> </ul>	<ul style="list-style-type: none"> <li>Specific and concrete</li> <li>Details consistent with intent</li> <li>Details create and vivid image</li> </ul>	<ul style="list-style-type: none"> <li>Extended elaboration of one main point</li> </ul>	<ul style="list-style-type: none"> <li>One or two minor errors</li> <li>No major errors</li> </ul>
<b>5</b>	Commendable Achievement	<ul style="list-style-type: none"> <li>Stated or implied main idea</li> <li>Focused and organized effectively orients reader</li> </ul>	<ul style="list-style-type: none"> <li>Specific sensory details</li> <li>Most details consistent with intent</li> </ul>	<ul style="list-style-type: none"> <li>Full elaboration of one main point</li> </ul>	<ul style="list-style-type: none"> <li>A few minor errors</li> <li>No more than one major error</li> </ul>
<b>4</b>	Adequate Achievement	<ul style="list-style-type: none"> <li>Main idea present but may not maintain consistent focus</li> <li>Some orientation of reader</li> </ul>	<ul style="list-style-type: none"> <li>Some specific details</li> <li>Details usually clear</li> <li>Generally clear images</li> </ul>	<ul style="list-style-type: none"> <li>Moderate elaboration of one main point</li> </ul>	<ul style="list-style-type: none"> <li>Some minor errors</li> <li>One or two major errors</li> <li>Errors do not cause reader confusion</li> </ul>
<b>3</b>	Some Evidence of Achievement	<ul style="list-style-type: none"> <li>Main idea not clear</li> <li>Usually on topic but with some digressions Focused and organized effectively orients reader</li> </ul>	<ul style="list-style-type: none"> <li>Few or inconsistent details</li> <li>Some details but not all may be appropriate</li> </ul>	<ul style="list-style-type: none"> <li>Restricted elaboration of one main point</li> </ul>	<ul style="list-style-type: none"> <li>Some minor and some major errors</li> <li>Some cause reader confusion</li> </ul>
<b>2</b>	Limited Evidence of Achievement	<ul style="list-style-type: none"> <li>Vague identification of main idea or focus</li> <li>Significant digressions</li> <li>No sense of closure</li> </ul>	<ul style="list-style-type: none"> <li>Little concrete language</li> <li>Simple or generic naming</li> </ul>	<ul style="list-style-type: none"> <li>Limited elaboration of one main point</li> </ul>	<ul style="list-style-type: none"> <li>Many minor and major errors</li> <li>Errors interfere with reader understanding</li> </ul>
<b>1</b>	Minimal Evidence of Achievement	<ul style="list-style-type: none"> <li>No apparent main idea</li> <li>No apparent plan of coherence</li> </ul>	<ul style="list-style-type: none"> <li>No concrete language</li> </ul>	<ul style="list-style-type: none"> <li>No elaboration of any point or central statement</li> </ul>	<ul style="list-style-type: none"> <li>Many major errors causing reader confusion</li> </ul>



---

### **Holistic scoring method**

In this method, a single overall score is given taking into account the entire response. Because no detailed criteria are needed, scoring papers in this method is rapid. However, this has its own shortcomings. Unless supplemented with comments teachers write on test papers, holistic scores alone provide less specific guidance to the student. Linn & Gronlund (2000) commented that like analytical scoring holistic scoring needs to have the scores or labels elaborated by statements.

Restricted response essay questions can usually be satisfactorily scored by the analytic method. The restricted scope and the limited number of characteristics included in a single response make it possible to define degrees of quality precisely enough to assign point values. The extended response question, however, usually requires the holistic method.

### ***3. Decide how to handle factors that are irrelevant to the learning outcomes being measured.***

Several factors influence our evaluation of answers that are not directly pertinent to the purpose of the measurement. Prominent among these are:

- legibility of hand writing,
- spelling,
- sentence structure, and
- punctuation,

We should make an effort to keep such factors from influencing our judgment when evaluating the content of the answers.

### ***4. Evaluate the responses of all students to one question before going to the next one.***

One factor that contributes to unreliable scoring of essay questions is a shifting of standards from one student's answer to the next. A paper with average answers may appear to be of much higher quality when it follows a failing paper than when it follows a near perfect one. One way to minimize this is to score all answers to the first question, reorder the papers, and score all answers to the second question, and so on until all the questions have been scored. A more uniform standard can be maintained with this procedure because it is easier to remember the basis for judging each answer, and answers of various degrees of quality can be more easily compared. It also helps to counteract another type of error. When we evaluate all of the answers to a single student, the first few answers may create a general impression of the student's achievement that affects our judgment of the remaining answers. Thus if the first answers are of high quality, we tend to overrate the following answers; if they are of low quality, we tend to underrate them. We call this condition a *carry over effect*—where our impression of the answer for one item affects the answer for the next item.

---

**Carry over effect** is the tendency of the scorer to rate the following items based on the impression he/she formed from the previously rated item. If a student did well on the first item, the teacher will give high scores for the next items though the answers may be poor, or vice versa.

When possible evaluate the answers without looking at the students' names. The general impression we form about each student during our teaching is also a source of bias in evaluating essay questions. This is called halo effect. This is a tendency on part of the scorers to allow their general impressions of a person to influence their evaluation of specific behaviors. If a teacher expects that the student is clever, is not uncommon for a teacher to give a high score to a poorly written answer by rationalizing that "the student is really capable, even though he/she didn't express it clearly".

**Halo effect** is a tendency on part of the scorers to allow their general impressions of a person to influence their evaluation of specific behaviors.

## **Unit 4**

# **Assembling, Administering, Scoring and Analyzing Classroom Tests**

### **6.1 ASSEMBLING TEST ITEMS**

Assembling involves recording the test items, reviewing them, arranging the items and the formats, writing directions, and reproducing the test.

#### **6.1.1 RECORDING TEST ITEMS**

When writing items, it is recommended that each test item be recorded on a separate page in our notebook. It is also recommended that in addition to the item the content from which the item is extracted, and the specific learning outcomes the item is measuring should be recorded. This enables us to make cross checking with table of specification easily.

#### **6.1.2 REVIEWING TEST ITEMS**

Before tests are made ready for reproduction and administration they should be carefully reviewed. This is because there are many errors that might be committed when we construct tests. When we concentrate so closely on some aspect of item construction, we may overlook (forget) others. This results in a number of unwanted errors that may distort the function of the item. However, such problems can be detected and minimized by

- i) reviewing the items after they have been set aside for a few days, and
- ii) asking a fellow teacher to review and criticize them.

#### **6.1.3 ARRANGING ITEMS IN THE TEST**

Once the test items are written, edited, and revised for errors, the next task is to arrange them in some order. Linn and Gronlund (2000) suggest the following arrangement order of items by format

- True-False
- Matching exercises
- Supply type (Short answer and completion)
- Multiple choice
- Interpretive exercises
- Essay

The above arrangement is based on the assumption that as one goes down difficulty of items as a group increases.

Then it is logical to arrange items in order of increasing difficulty, i.e., beginning with the easiest items and proceeding gradually to the most difficult ones. Beginning with easy items

helps to increase motivation and confidence in students to work on the items that follow. On the other hand, if our test starts with an item that is difficult students may lose confidence right from the start and may likely be to miss even simple items that follow.

## ACTIVITY

In the former school leaving examination, ESLCE, to avoid cheating especially copying from one another, differently coded test papers were used. It happened that question number 1 for one paper could be question number 9 for the other. Does this have any problem from assembling perspective? Explain.

### 6.1.4 PREPARING DIRECTIONS FOR THE TEST

Sometimes due to problems with clarity of directions students get confused with regard to how they are supposed to respond to the questions. The problem of student confusion can be reduced if certain guidelines are followed in writing test directions. What are those guidelines? They are as follows.

Test directions generally should include:

- i) purpose of the test,
- ii) time allowed for completing the test,
- iii) directions for responding,
- iv) how to record the answers,
- v) basis for scoring open ended or extended response tests.

#### i. Purpose of

#### the Test

The direction should inform students about the purpose of the test. This includes the course or the subject the exam is for, whether it is diagnostic, mid semester exam, final examination, etc. Though this can be done orally when the examination program is set, it is preferable if the heading of the exam says something about purpose of the test.

For instance, when a final examination is prepared for the course Educational Measurement and Evaluation, the examination paper will have the following information on the top of the first page.

<b>Bahir Dar University</b> <b>Education Faculty</b> <b>Department of Pedagogical Sciences</b> <b><u>Educational Measurement and Evaluation (Epsy. 212) Final Examination (60%)</u></b> <b>Date: 24-04-98 E.C.</b> <b>Time allowed: 1:45 hrs.</b>			Purpose of the test
Name _____ ID No. _____ Dept. _____	How to record responses	Time allocated for the test	
<b>General Directions:</b> In this test paper there are four parts: True-false, matching, short answer, and multiple choice type items. Write your answers on the blank space given. For the selection type items use CAPITAL LETTERS only.		Direction for responding	

## ii. Time Allowed for Completing the Test

It is important to inform students how much time is given to the entire test. In addition, it is good if students are informed about length of time to be allocated to each section in the test. This enables students to effectively use their time. It also rescues less able students from spending unnecessarily long time on questions that appear difficult to them.

In relation to length of time that students should be allowed in a given test, there is *no any limit*. It is suggested that time should not constrain students from responding to items. This does not, however, mean there should not be any limit. The rule of thumb is 1 minute per two true false items or one multiple choice or one short answer item.

## iii. Directions for Responding

How students should respond to the questions should be well spelt. For instance, in multiple choice items students should know whether they have to choose the correct answer or the best answer. When the testees are young children to help them understand how to respond, it is important to include correctly marked sample items.

## iv. How to Record the Answers

This refers to how students indicate their answers and whether they should provide their answers on the test booklet itself or on a separate answer sheet. In selection type of test items

like multiple choice students may be instructed to circle, underline, or write the letter of their choices. But for young children instructing them to underline their answers is preferable.

- number of testees, and
- length of the test

determine where students should preferably record their answers.

If the number of testees is small and the test is too short, answers can be recorded on the test paper itself. On the other hand, with large number of testees and long test it is preferable to get students record their answers on a separate answer sheet.

#### **v. Basis for Scoring Open Ended or Extended Response Tests**

If there are many essay items it is suggested that we indicated in the direction the weight of each item. In addition, we have to also indicate the relative weight given for each of the components we need to appear in the students' responses to essay test items. For instance, we need to determine how much is given to factual accuracy, organization, comprehensiveness, originality etc.

### **1. ADMINISTERING TESTS**

The guiding principle of test administration is that there should be a fair chance for all students to demonstrate their achievement of the learning outcomes being measured. In addition to this, we have to consider physical and psychological conditions of testees as they may help or hamper students from demonstrating their full performances or achievements.

The physical environment should be as conducive as possible. Conducive physical environment includes that the testing room should be quiet, there should be adequate light, ventilation, adequate workspace, comfortable seats etc. The psychological conditions influence students' scores more seriously than the physical ones. The psychological conditions include mental preparedness of testees to take and pass exams. Any condition that may result in tension should be eliminated. Some sources of anxiety among students when taking tests include the following:

- i) threatening students with tests if they do not behave in a required way.
- ii) warning students to do their best because the test is important.
- iii) telling students to work fast in order to complete on time.
- iv) threatening students on consequences if they fail.

Apart from working for the conduciveness of physical and psychological factors, there are some practices we need to avoid when we administer our tests. Some of them are the following:

- i. Don't talk unnecessarily before letting students start working.

---

First of all since students are likely to think about the exam other instructions unrelated to the test may be overlooked by testees. For instance, many students may fail to grasp and remember when you have arranged for make up classes if you tell them during test administration. For some students this may cause frustration.

ii. Keep interruptions to a minimum.

If we have some corrections or related instructions, we have to make them at the beginning of the test. Otherwise, interrupting the testees now and then by giving corrections or other instructions is not suggested, because it may disturb them.

iii. Avoid giving hints to students about individual items.

While taking a test, students may ask you about many things related to the test. That may be about lack of clarity, definition of terms, ambiguity, and unreadability of items. In this case providing individual clarification is not suggested. But if you believe that the item needs clarification, you should do it for the entire testees by calling their attention to it. In contrast, if you feel that no correction is needed you have to be silent and don't provide individual clarification. While you are making clarifications on individual basis, you may unknowingly provide unintended clues.

iv. Discourage cheating.

It is not uncommon to see invigilators do other things when students take exams. For example they may score previous tests, work on research proposals or read books. This undoubtedly gives chance for students to cheat. Cheating may take different forms. Copying answers from each other, copying answers from scratches of papers, getting others not belonging to the class take exams for others are some of them. In order to get valid results on students' achievement or performance we have to discourage cheating. The best way to avoid cheating is careful proctoring of testees. In a condition in which there is a large number of testees, it is advisable to have another person assist you. Another and complementary way is being careful about seating arrangements.

## **ii. SCORING THE ANSWERS**

There are basically three types of scoring. They are

- hand scoring,
- machine scoring, and
- self scoring.

Which scoring method to use depends on availability of scoring equipment, the speed with which the test results are needed.

### **Hand Scoring**

In this type of scoring, the teacher or others can score the test papers of students. In this case, the teacher prepares the answer keys and provides it to the scorers.

**Self Scoring**

This type of scoring is done by the students themselves. The teacher will provide the answer keys to the students and the students score the papers. So in this type of scoring the students can determine their total score on a test on their own. The disadvantage is the students may cheat when scoring their own papers.

**Machine Scoring**

As the name itself implies in this type of scoring a scoring machine is used. This is quite useful with large number of testees like for example students taking national examination, EGSCE, all over Ethiopia.

**III. ITEM ANALYSIS**

Item analysis is the process of examining students' responses to each item to determine the quality of test items. In item analysis, the specific activities done are determining difficulty level and discrimination power of test items and judging how effectively distracters are functioning in case of multiple choice items.

The purpose of item analysis is to select the best items from the poor ones. It is preferable from educational measurement and evaluation viewpoint to use test items that are judged good even if they were used before. Thus, here comes the role of item analysis. In item analysis we will determine which items can be used directly which with revision and which should be discarded. Apart from the above purpose, item analysis has the following purposes in classrooms.

**1. Item analysis data provide a basis for efficient class discussion of test results**

One of the tasks in item analysis is counting the number of times an alternative is chosen by students as a correct answer. This gives chance for both teachers and students to discuss on misinformation and misunderstandings. Item analysis also helps teachers to identify technical defects. They also suggest needed change on scoring keys or scoring rubric in essays for instance in a case in which high achieving students most frequently chose an alternative you think is a distracter.

**2. Item Analysis data provide a basis for remedial work**

Although discussion of test results can provide chance to clarify specific problems, item analysis suggests general areas of students weaknesses that need more attention. If for example students' score is less than expected, this may suggest that you need to revisit critical concepts or topics.



---

### **3. Item analysis data provide a basis for the general improvement of classroom instruction.**

Item analysis data provide information that can assist in determining the appropriateness of learning outcomes and course contents defined for some group of learners. Students' scores may lead one to the extent of revising curricula.

### **4. Item analysis procedures provide a basis for increased skill in test construction**

In item analysis, you will identify existence of ambiguity, unintended clues, ineffective distracters, etc. All this information is useful in revising the items for future use. Mostly teacher who make item analysis are better than those who do not make in terms of constructing good test items. This is because the former type of teachers will get the chance to learn from their own errors.

## **I. PROCEDURES OF ITEM ANALYSIS FOR OBJECTIVE TEST ITEMS**

Dear student, you may ask how item analysis is done. Item analysis is carried out based on the following procedures.

1. First, arrange the scored test papers in order from the highest score to the lowest score.
2. Divide the ordered papers in two halves. Put those highest scores in one group and those with the lowest scores in another. Take the top 27% from high achieving students, and the bottom 27% from low achieving students. For example, if the number of your testees is 80 and you want to make an item analysis, first you will consider the top 27% and the bottom 27% of the students, i.e.

$$80 \times \frac{27}{100} = 21.6$$

So using this formula we will have to use 22 papers from the upper group and another 22 papers from the lower group, which is a total of 44 student papers.

3. If the number of students is small like 40 or 50, there is no need to take the upper and the lower 27%: you can simply divide it into two halves and take 20 of the upper papers and 20 of the lower papers. For each item count the number of examinees in the upper group and in the lower group that choose each response alternative (in the completion, short answer, and true-false questions count the number of students who answered the question correctly) and record the counts separately for the upper group and the lower group. Add the counts of the lower and the upper group for the correct answer, and divide the sum by the total number of upper and lower group students and multiply the value by 100%. This will provide index of item difficulty (P). The formula is

$$P = \frac{R_u + R_L}{T} \times 100$$

T= Total number of upper and lower group students

P= Item difficulty Index

$R_U$  = number of upper group students who got the item right

$R_L$  = number of lower group students who got the item right

- Subtract the counts of the lower group from the counts of the upper group and divide the result by half of the total number of upper and lower group students. This will provide index of item discrimination (D).

$$D = \frac{R_u - R_L}{\frac{1}{2}(T)}$$

- Evaluate how distracters are functioning. The purpose of distracters is distracting the unprepared student from getting the correct answer. Thus in good items there should be more students from lower than from the upper who choose them. If more students from the upper than from the lower group happen to mark distracters as correct answers, the item will have poor discrimination power and even negative one.

Example: Let us assume that a 10-multiple choice items test was administered to 40 students. The teacher wanted to conduct item analysis. The results of students for the first item where the correct answer is B are presented below.

Item Number 1	Alternatives				
	A	B*	C	D	Omit
Upper Group (20)	1	19	0	0	0
Lower Group (20)	5	9	0	6	0

For the above item, the difficulty level (P) is given as

$$P = \frac{20+8}{40} \times 100$$

$$= 70\%$$

and the discrimination power of the item (D) is given as

$$D = \frac{19-9}{\frac{1}{2}(40)}$$

$$= 0.5$$

As regards the distracters, alternatives A and D were functioning as intended because they attracted larger number of students from the lower group than from the upper one. In contrast, alternative C did not function as intended because it attracted no student. Therefore,

this alternative needs improvement for future use or the item should have had three alternatives only.

## **I. INTERPRETING ITEM DIFFICULTY AND ITEM DISCRIMINATION IN OBJECTIVE TEST ITEMS**

Though there is no clear cut guideline to interpret items based on their level of difficulty and discrimination there are rule of thumbs. The following guidelines are suggested to determine the difficulty levels of different formats of test.

Item Format	Ideal Difficulty Level
Completion and short answer	50
5 response multiple choice	70
4 response multiple choice	74
3 response multiple choice	77
True false	85

As regards item discrimination index, Ebel & Frisbie (1991) suggested the following rule of thumb.

Index of D	Interpretation
0.40 and up	Very good item
0.30 to 0.39	Reasonably good but possibly subject to improvement
0.20 to 0.29	Marginal that needs improvement
Below 0.20	Poor items

### **ACTIVITY**

Following are an item analysis data. The data indicate the number of students choosing each alternative. The correct answer is marked by (\*). Based on the data, determine P and D, and interpret the results. Evaluate the effectiveness of the distracters.

Item Number		Alternatives				
		A	B	C	D	Omit
1.	Upper 27%	8*	5	7	7	0
	Lower 27%	2*	9	8	8	0
2	Upper 27%	13	2	10*	2	0
	Lower 27%	5	1	9*	12	0
3	Upper 27%	2	20*	3	2	0
	Lower 27%	11	2*	8	6	0
4	Upper 27%	0	4	3	20*	0
	Lower 27%	0	7	12	8*	0

## REFERENCES

- Airasian, P.W. (1996). *Assessment in Classroom*. New York : McGraw-Hill, Book Com.
- Borich, G.D, and Tombrie, (1995). *Educational Psychology*. New York: HarperCollins College publishers.
- Borich, G.D. (1988). *Effective Teaching Psychology*. New York: Macmillan Publishing Com.
- Carey, L.M. (1994). *Measuring and Evaluating School Learning*. 2<sup>nd</sup> ed. Boston: Allyn and Bacon.
- Ebel, R.L, and Frisbie, D. A. (1991). *Essentials of Educational Measurement*. Englewood Cliffs, CA: Prentice Hall.
- Eggen, P. and Kauchak, D. (1999). *Educational Psychology: Windows on Classrooms*. 4<sup>th</sup> ed. Upper Saddle River: Prentice-Hall, inc.
- Elliot, S. et al. (2000). *Educational Psychology: Effective Teaching, Effective Learning*. 3<sup>rd</sup> ed. Boston: McGraw- Hill Companies, Inc.
- Gronlund, E. N. (1971). *Measurement and Evaluation*. 2<sup>nd</sup> ed. New York: Collier-Macmillan limited.
- Mehrens, W.A., and Lehmann, I. J. (1984). *Measurement and Evaluation*. New York: Holt Rinehart and Winston.
- Nitko, A.J (1996). *Educational Assessment of Students*. 2<sup>nd</sup> ed. Englewood Cliffs, CA: Merrill Prentice Hall.
- Oosterhof, A.( 1994). *Classroom Application of Educational Measurement*. New York: Macmillan Publishing company.
- Payne, D.A. (1997). *Applied Educational Assessment*. Belmont, CA: Wadsworth Publishing Company.
- Payne, D.A.(1992). *Measuring and Evaluating Educational Outcomes*. New York: Macmillan Publishing Company.