CATHOLIC UNIVERSITY OF RWANDA

FACULTY OF EDUCATION

PGDE Programme

MODULE CODE: PGDE6362

MODULE TITLE: **THEORY AND PRACTICE OF TEACHING**

UNIT 1: MEASUREMENT AND EVALUATION

BY: DR. PHILOTHERE NTAWIHA (PhD)

Learning outcomes

By the end of this unit you should be able to:

- Distinguish between test, measurement, assessment and evaluation;
- State the purposes of assessment and evaluation in education;
- Explain and construct a valid and reliable tests;
- Analyse the test items
- Moderate a test
- Mark a test
- Compute some measures of general tendency, variability, and relationship to interpret the tests in general and continuous assessment in particular

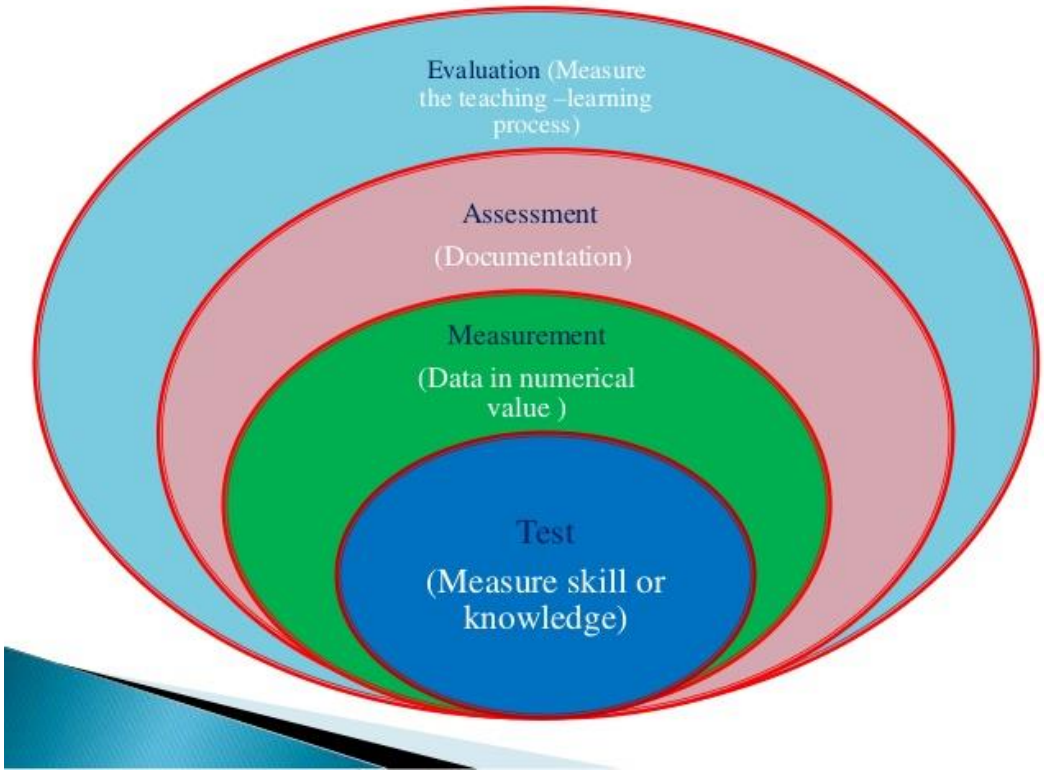
Unit content

In this course unit, you will be taken through the following points:

- The concept of Testing, measurement, assessment and evaluation in education
- Importance of educational assessment, measurement and evaluation
- Educational objectives
- Qualities of a good test: Validity, Reliability, item analysis
- Interpretation of test scores: frequency distribution, measures of central tendency, variability, and relationship.
- Tests moderation and marking
- Grading

Food for Thought

You always hear people using terms such as test, measurement, assessment and evaluation. However, it is very hard for some of them to establish a clear distinction between them. As an educationist, what are the similarities and differences between the four terms? With reference to the following figure, try to establish the existing similarities and differences between them.



Lesson one: The concept of Testing, measurement, assessment and evaluation in education

When defined within an educational setting, assessment, evaluation, and testing are all used to measure how much of the assigned materials students are mastering, how well student are learning the materials, and how well student are meeting the stated goals and objectives. However, education professionals make distinctions between assessment, evaluation, and testing. The distinctions are the following:

1. Assessment: the process of collecting, synthesizing and interpreting information in order to make a decision. Assessment in educational setting may describe the progress students have made towards a given educational goal at a point in time. However, it is not concerned with the explanation of the underlying reasons and does not proffer recommendations for action. Although, there may be some implied judgment as to the satisfactoriness or otherwise of the situation.

In the classroom, assessment refers to all the processes and products which are used to describe the nature and the extent of pupils' learning. This also takes cognizance of the degree of correspondence of such learning with the objectives of instruction.

Types of assessment

2. Evaluation: is the process of judging the quality or value of a performance or a course of action. In this perspective, evaluation adds the ingredient of value judgment to assessment. It is concerned with the application of its findings and implies some judgment of the effectiveness, social utility or desirability of a product, process or progress in terms of carefully defined and agreed upon objectives or values. Evaluation often includes recommendations for constructive action. Thus, evaluation is a qualitative measure of the prevailing situation. It calls for evidence of effectiveness, suitability, or goodness of the programme. "*It is the estimation of the worth of a thing, process or programme in order to reach meaningful decisions about that thing, process or programme*."

Types of Evaluation

In education system, there are two main levels of evaluation: programme level and student level. The former has to do with the determination of whether a programme has been successfully implemented or not, while the latter has to do with determination of how well a

student is performing in a programme of study. Each of the two levels can involve either of the two main types of evaluation: **formative evaluation** and **summative evaluation.**

1. **Formative Evaluation:** This is a type of evaluation that takes place while the program activities are happening, the course is progressing. It is designed to help both the student and teacher to pinpoint areas where the student has failed to learn so that this failure may be rectified. It provides a feedback to the teacher and the student and thus estimating teaching success e.g. daily, weekly tests, terminal examinations etc. In other words, the purpose of formative evaluation is to find out whether after a learning experience, students are able to do what they were previously unable to do. It therefore helps students perform well at the end of a programme. Formative evaluation enables both the teacher and students to:

1. Provide feedback on performance, in class or on assignments,
2. Identify the levels of cognitive process of his students;
3. Choose the most suitable teaching techniques and materials;
4. Determine the feasibility of a programme within the classroom setting;
5. Determine areas needing modifications or improvement in the teaching-learning process; and
6. Determine to a great extent the outcome of summative evaluation.

7. Restructure their (students) understanding /skills and build more powerful ideas and capabilities

2. Summative Evaluation: This is the type of evaluation carried out at the end of the course of instruction to determine the extent to which the objectives have been achieved. It is called a summarizing evaluation because it looks at the entire course of instruction or programme and can pass judgment on both the teacher and students, the curriculum and the entire system. It often attempts to determine the extent the broad objectives of a programme have been achieved. It is usually used for grading or certification.

3. Testing: It is the process of measuring the level of skills or knowledge that has been reached. Tests are therefore detailed or small scale task carried out to identify the candidate's level of performance and to find out how far the person has learnt what was taught or be able to do what he/she is expected to do after teaching. In other words, tests are formal, systematic procedure for gathering information to examine someone's knowledge of something in order to determine what he or she knows or has learned. They are the instruments for assessment.

Types of a test

By the aim and objective of a test, we have the following types of tests:

a) Placement test: for placing students at a particular level, school, or college.

b) Achievement tests: for measuring the achievement of a candidate in a particular course either during or at the end of the course.

c) Diagnostic tests: for determining the problems of a student in a particular area, task, course, or programme. Diagnostic tests also bring out areas of difficulty of a student for the purpose of remediation.

d) Aptitude tests: are designed to determine the aptitude of a student for a particular task, course, programme, job, etc.

e) Predictive tests: designed to be able to predict the learning outcomes of the candidate. A predictive test is able to predict or forecast that if the candidate is able to pass a particular test, he/she will be able to carry out a particular task, skill, course, action, or programme.

f) Continuous assessment tests are designed to measure the progress of students in a continuous manner. Such tests are taken intermittently and students' progress measured regularly. The cumulative scores of students in continuous assessment often form part of the overall assessment of the students in the course or subject.

g) Standardized tests: are any of the above mentioned tests that have been tried out with large groups of individuals, whose scores provide standard norms or reference points for interpreting any scores that anybody who writes the tests has attained. Standardized tests are to be administered in a standard manner under uniform positions.

h) Nonstandardized tests/Teacher-made tests: these are tests produced by teachers for a particular classroom use. Such tests may not be used far-and-wide but are often designed to meet the particular learning needs of the students.

4. Measurement: is the process of quantifying or assigning a number to a performance or trait. Measurement is therefore a process of assigning numerals to objects, quantities or events in order to give quantitative meaning to such qualities. In the classroom, to determine a child's performance, you need to obtain quantitative measures on the individual scores of the child. If the child scores 80 in Mathematics, there is no other interpretation you should give it. You cannot say he has passed or failed. Measurement stops at ascribing the quantity but not making value judgment on the child's performance.

Scales of measurement

a) Nominal scale of measurement: It is a scale in which the numbers are used to label, classify, or identify the people or objects of interest. E.g. the numbers that appear on football jerseys to identify a given player. Another example of a nominal scale of measurement is the numbers that are assigned to the seats in the football stadium. In this case the number merely distinguishes one seat from the other and in this way allows you to pick out the seat corresponding to the ticket you have purchased.

b) Ordinal scale of measurement: It is the rank order scale. It allows to make ordinal judgments, that is, it allows to determine which is better or worse than any other.

c) Interval scale of measurement: It is a scale of measurement that has equal distances between adjacent numbers in addition to ordinality. Therefore, with an interval scale of measurement you gain the ability to specify the distance that exists between the people, objects, or events of interest for the variable being measured.

d) Ratio scale of measurement: It is a scale that allows us to make ratio as well as ordinal statements because it has equal intervals between adjacent points on the scale and an absolute zero point. Any time you determine how much you weigh, how tall you are, or how many items you correctly answered on an exam, you are using a ratio scale of measurement. This is because each of these scales has a zero point that implies the total absence of the characteristic being measured, and the units above this zero point increase in equal amounts.

Purpose of educational assessment, measurement and evaluation

Assessment, measurement and evaluation in education setting are carried out from time to time for the following purpose:

- to determine the relative effectiveness of the programme in terms of students' behavioural output;
- to make reliable decisions about educational planning;
- to ascertain the worth of time, energy and resources invested in a programme;
- to identify students' growth or lack of growth in acquiring desirable knowledge, skills, attitudes and societal values;
- to help teachers determine the effectiveness of their teaching techniques and learning materials;

- to help motivate students to want to learn more as they discover their progress or lack of progress in given tasks;
- to encourage students to develop a sense of discipline and systematic study habits;
- to provide educational administrators with adequate information about teachers' effectiveness and school need;
- to acquaint parents or guardians with their children's performances;
- to identify problems that might hinder or prevent the achievement of set goals;
- to predict the general trend in the development of the teaching-learning process;
- to ensure an economical and efficient management of scarce resources;
- to provide an objective basis for determining the promotion of students from one class to another as well as the award of certificates;
- to provide a just basis for determining at what level of education the possessor of a certificate should enter a career.

CHECK YOUR PROGRESS

1. With relevant and concrete examples, distinguish between test, assessment, measurement, and evaluation in education.
2. Discuss different types of evaluation that are used in education.
3. Discuss different types of tests used in education.
4. Using good examples, distinguish between different scales of measurement.

Lesson two: Educational Objectives

In our everyday activities, objectives help us focus on what's important; they remind us of what we want to accomplish. Objectives in teaching describe the kinds of content, skills and behaviours teachers hope their students will develop through instruction.

Objectives are very important in developing lesson plans. Teachers cannot help students meet their objectives if they do not know what their objectives are. Similarly, if teachers don't identify their objectives, instruction and assessment will be purposeless.

Objectives are particularly crucial in teaching because teaching is an intentional and normative act. Teaching is **intentional** because **teachers teach for a purpose; they want students to learn something as a result of teaching**. Teaching is also **normative** because **what teachers teach is viewed by them as being worthwhile for their students to learn**.

Because teaching is both **intentional** and **normative**, it is always based on objectives. Normative teaching is concerned with selecting objectives that are worthwhile for students to learn. Intentional teaching is concerned with issues of how teachers will teach their objectives-what learning environments they will create and what methods they use to help students learn the intended objectives. Although teachers' objectives may sometimes be implicit and fuzzy, it is best that objectives be implicit, clear, and measurable.

Taxonomy of educational objectives

What is taxonomy? Taxonomy refers to classification of organisms or objects into groups based on similarities of structure or origin, etc. Since education is a cross-cutting issue, its objectives need to be classified. Hence, educational objectives are normally classified according to their **level of specificity** and to **domain of human behavior**.

A. LEVELS OF EDUCATIONAL OBJECTIVES

Objectives can range from very general to very specific. Depending on their level of specificity, objectives can be classified into one of the three levels ranging from the most broad to least broad: **global**, **educational**, and **instructional**. It is worth to note that regardless of the type or specificity, an objective should always focus on ***student*** learning and performance rather than on teacher actions or classroom activities.

1. Global Objectives: often called "Goals" are broad, complex student learning outcomes that require substantial time and instruction to accomplish. They are very general, encompassing a large number of more specific objectives. Examples include the following:

- The student will become a lifelong learner
- The student will become mathematically literate
- The students will learn to use their minds well, so that they may be prepared for responsible citizenship, further learning, and productive employment in our nation's economy.

Because global objectives are broadly inclusive, they are rarely used in classroom assessment unless they are broken down into more narrow objectives. The breadth encompassed in global objectives makes them difficult for teachers to use in planning classroom instruction. Hence narrower objectives must be identified to meet classroom needs.

2. Educational Objectives: these objectives represent a middle level of abstraction. The following are some examples:

- The student can interpret different types of social data
- The student can correctly solve addition problems containing two digits
- The student distinguishes between facts and hypotheses
- The student can read English poetry aloud.

Educational objectives are more specific than global objectives. They are sufficiently narrow to help teachers plan and focus teaching, and sufficiently broad to indicate the richness of the objectives and to suggest a range of possible student outcomes associated with the objective.

3. Instructional Objectives: these are the most specific type of objective. Examples include the following:

- The student can correctly punctuate sentences
- Given five problems requiring the student to find the lowest common denominator of a fraction, the student can solve at least four of the problems
- The student can list the names of the first five U.S. presidents.

Instructional objectives focus teaching on relatively narrow topics of learning in a content area. These concrete objectives are used in planning daily lessons.

The following table illustrates the difference in degree of breadth among the three types of objectives and compares their purposes, scopes, and time frames.

	Level of objective		
	Global	**Educational**	**Instructional**
Function/ purpose	Provide vision	Develop curriculum, plan instruction, define suitable assessments	Plan teaching activities, learning experiences, and assessment exercises
Scope	Broad	Intermediate	Narrow
Time to accomplish	One or more years	Weeks or months	Hours or days
Example of breadth	The student will know how to repair a variety of home problems	The student will use appropriate procedures to find solutions to electrical problems in the home	Given a home repair problem dealing with a malfunctioning lamp, the student will repair it.

CHECK YOUR PROGRESS

1. Explain what you understand with taxonomy of educational objectives.
2. Distinguish between different types of educational objectives according to their level of specificity.

B) THREE DOMAINS OF EDUCATIONAL OBJECTIVES

Objectives are logically and closely tied to instruction and assessment. In addition to differing in terms of level, classroom objectives (and their related instruction and assessments) differ in terms of three general types of human behavior: **the cognitive**, **affective**, and **psychomotor domains**. For the cognitive domain of human behavior, the most used taxonomy is the **Bloom's taxonomy**. For the affective domain, the most used taxonomy is that of **Krathwohl and associates** (Krathwohl, Bloom, and Masia). And for the psycho-motor domain of human behavior the most known taxonomy is the **taxonomy by Dave**.

1. Cognitive domain (Knowledge-Based)

The cognitive domain involves knowledge and the development of intellectual skills (Bloom, 1956). This includes the recall or recognition of specific facts, procedural patterns, and concepts that serve in the development of intellectual abilities and skills. According to Bloom's Taxonomy, there are six major categories of cognitive processes, starting from the simplest to the most complex. Bloom's Taxonomy was created in 1956 under the leadership of educational psychologist Dr Benjamin Bloom in order to promote higher forms of thinking in education, such as analyzing and evaluating concepts, processes, procedures, and principles, rather than just remembering facts (rote learning). Lorin Anderson, a former student of Bloom, and David Krathwohl revisited the cognitive domain in the mid-nineties and made some changes, changing the names in the six categories from noun to verb forms, and rearranging them. The following tables present the old and revised versions of the Bloom's Taxonomy.

Old version of the Bloom's Taxonomy

Taxonomy Level	Related Verbs	General Description
1. Knowledge	Remember, recall, identify, recognize	Memorizing facts
2. Comprehension	Translate, rephrase, restate, interpret, describe, explain	Explaining in one's own words
3. Application	Apply, execute, solve, implement	Solving new problems
4. Analysis	Break down, categorize, distinguish, compare	Breaking into parts and identifying relationships
5. Synthesis	Integrate, organize, relate, combine, construct, design	Combining elements into a whole
6. Evaluation	Judge, assess, value, appraise	Judging quality or worth

The revised version of the Bloom's Taxonomy

Level	Description	Verbs	Objective
1.Remembering	Recall information	Define List	Define levels of cognitive domain
2.Understanding	Understand meaning of information	Describe Explain	Explain purpose of cognitive domain
3.Applying	Utilize information to complete a task	Solve Use	Write objectives for levels of cognitive domain
4.Analyzing	Classify and relate assumptions or evidence	Contrast Examine	Compare cognitive and affective domain
5.Evaluating	Critique idea based on specific standards and criteria	Judge Justify	Judge effectiveness of writing objectives using taxonomy
6.Creating	Integrate or combine ideas into a new product or plan	Design Develop	Design way to write objectives that combines 3 domains

2. Affective domain (Emotive-Based)

The affective domain involves feelings, attitudes, interests, preferences, values, and emotions. Emotional stability, motivation, trustworthiness, self-control, and personality are all examples of affective characteristics. Although affective behaviors are rarely assessed formally in schools and classrooms, teachers constantly assess affective behaviors informally, especially when sizing up students. Teachers need to know who can be trusted to work unsupervised and who cannot, who can maintain self-control when the teacher has to leave the classroom and who cannot, who needs to be encouraged to speak in class and who does not, who is interested in science but not in social studies, and who needs to be prodded to start class work and who does not. Most classroom teachers can describe their students' affective characteristics based on their informal observations and interactions with the students. The most widely accepted and used taxonomy of affective behaviors is the taxonomy prepared by **Krathwohl and associates** (Krathwohl, Bloom, and Masia). According to this taxonomy, there are five levels in the affective domain.

Taxonomy Level	**Related Verbs**	**General Description**	**Objectives**
1. Receiving	Listen, notice, tolerate	Being aware of	Listen attentively the music on the cassette player
2. Responding	Comply, enjoy, follow	Showing some new behavior as a result of experience	Dance following the music's rhythm on the cassette player
3. Valuing	Carry out, express	Showing some definite involvement or commitment	Voluntary participate in a music festival
4. Organization	Choose, consider, prefer	Integrating a new value into one's general set of values relative to other priorities.	Purchase own music equipment
5. Characterization	Act on, depict, exemplify	Acting consistently with the new value; person is known by the value	Join existing groups to play music twice per week

3. Psychomotor domain (Action-Based)

The psychomotor domain includes physical and manipulative activities. Holding a pencil, using a mouse, keyboarding, setting up laboratory equipment, building a bookcase, playing a musical instrument, shooting a basketball, buttoning a jacket, and brushing teeth are good examples of activities that involve psychomotor behaviors. Although psychomotor behaviors are present and important at all levels of schooling, they are especially stressed in the pre-school and elementary grades, where tasks like holding a pencil, opening a locker, and buttoning or zipping clothing are important to master. Although there are a number of psychomotor domain taxonomies, the most accepted and used taxonomy is the one prepared by **Dave** (1975) which has five levels.

Taxonomy Level	Related Verbs	General Description	Objectives
1. Imitating	Follow, mimic, replicate	Observing and patterning behavior after someone else	Perform a skill while observing a demonstrator
2. Manipulating	Act, execute, perform	Performing actions by memory or following instructions	Build a model following instructions
3. Refining	Demonstrate, improve, master	Performing actions with a high degree of precision	Perform a skill without assistance
4. Coordinating	Adapt, combine, customize	Adapting a series of actions to achieve harmony and internal consistency	Combine a series of skills to produce a video with music, drama, color, sound, etc.
5. Naturalizing	Create, design, invent	Mastering a high level performance until it becomes second-nature or natural	Display competence while in action (i.e. Jordan playing basketball or Nancy Lopez hitting a golf ball).

As noted previously, assessments encompass the cognitive, affective, and psychomotor domains because teachers are interested in knowing about their students' intellectual, attitudinal, and physical characteristics. Notice, however, that different assessment approaches characterize the different behavior domains. For example, the cognitive domain is most likely to be assessed using paper-and-pencil tests or various kinds of oral questioning. Behaviors in the affective domain are most likely to be assessed by observation or questionnaires: for example, which subject do you prefer, English or chemistry? Psychomotor behaviors are generally assessed by observing students carrying out the desired physical activity.

CHECK YOUR PROGRESS

1. Explain different levels in the revised Bloom's taxonomy of the cognitive domain and state a clear objective for each level.
2. Explain different levels involved in the affective domain and state an objective for each level.
3. Discuss different levels involved in the psycho-motor and state an objective for each level.

Lesson three: qualities of a test

1. Validity

What is validity? The term **validity** refers to the degree to which a test is measuring what it was supposed to measure. It is the most important idea to consider when preparing or selecting a test.

Types of test validity

a) Content validity: The extent to which the content of a test correspond to variables it is designed to measure. It shows how adequately the test samples the universe of knowledge, skills, perceptions and attitudes that the test taker is expected to exhibit.

How to establish content validity of an instrument? The content validity of a test is determined by expert judgment, i.e. experts in the area covered by the test are asked to assess how well items in the test represent their intended content area. This is mostly done through moderation.

b) Criterion-related validity: The relationship between scores obtained using the instrument and scores obtained using one or more other instruments or measures. There are two types of criterion-related validity. These are concurrent validity and predictive validity.

i) Concurrent validity: a newly constructed instrument is correlated with an earlier instrument that has been constructed and established to be valid for similar purpose.

Procedures to establish concurrent validity:

The two instruments (the newly and the earlier constructed instruments) are administered to the same group at the same time or shortly thereafter then the two sets of scores are correlated. A correlation coefficient of more than 0.6 is required for the instrument to be concurrently valid.

ii) Predictive validity: the degree to which a test can predict how well an individual will do in a future situation. It is extremely important for tests that are used to classify or select individuals.

How to determine predictive validity of a test? Predictive validity of an instrument is determined in the following way:

- Administer an instrument (a test) to a group of subjects

- Wait until the behavior to be predicted occurs

- Obtain measures of criterion (behavior) for the same group

- Correlate the two sets of scores.

c) Construct validity: Used to measure a particular construct to which are attached certain meanings. Psychological concepts such as intelligence, anxiety, and creativity are considered hypothetical constructs because they are measured in terms of observable effects on behavior. Tests have been designed to measure such concepts.

How to establish construct validity of a test?

- Expert judgment: the test is given to different experts to judge whether indicators in the test correspond to the pre-determined indicators.
- Calculation of a correlation coefficient between the two tests (pre-determined and the developed one).

Threats to/ factors influencing test validity

- cultural beliefs
- Attitudes of testees
- Values: students often relax when much emphasis is not placed on education
- Maturity: students perform poorly when given tasks above their mental age.
- Atmosphere: Examinations must be taken under conducive atmosphere
- Absenteeism: Absentee students often perform poorly

2. Reliability (Types of reliability of a test and factors affecting it and methods of estimating test reliability)

The term 'Reliability' refers to the degree to which the test consistently measure whatever they are measuring.

a) Test-retest reliability/stability

The degree to which scores on the same test by the same individuals are consistent over time. The same test is twice administered to the same group of respondents within a certain period

of time then scores from the two administrations are correlated using either Pearson Product Moment Correlation Coefficient or Spearman Rank Order Correlation Coefficient.

i) The formula for Pearson Product Moment Correlation Coefficient (**r**) is:

$$r = \frac{N(\Sigma XY)-(\Sigma X)(\Sigma Y)}{\sqrt{[N(\Sigma X^2)-(\Sigma X)^2]\,[N(\Sigma Y^2)-(\Sigma Y)^2]}}$$

Where:

r= Pearson product moment correlation coefficient

Σ= Sum of

X= A raw score in test A

Y= A raw score in test B

N= Total number of scores

ii) The formula for Spearman Rank order Correlation Coefficient ($\mathbf{r_s}$) is:

$$r_s = 1-\left(\frac{6\Sigma D^2}{N(N^2-1)}\right)$$

Where:

$\mathbf{r_s}$= Spearman correlation coefficient between the rank orders

Σ= Sum of

$\mathbf{D^2}$= Squared differences between the rankings

N= Number of pairs of rankings

b) Equivalent forms reliability/parallel or alternative forms reliability

The degree to which two different forms of a test but equivalent administered to same individuals at the same time delivers the same scores. This can be shown using either Pearson Product Moment Correlation Coefficient or Spearman Rank Order Correlation Coefficient.

c) Internal consistency reliability

It tests the internal consistency of items of the tests. . The most used approaches of establishing internal consistency reliability are:

a) Split-half procedure: Breaking single test into two halves (odd items versus even items). This type of reliability can be obtained using either Pearson Product Moment Correlation Coefficient or Spearman Rank Order Correlation Coefficient.

b) Kuder- Richardson Approach:

Reliability of tests is often expressed in terms of correlation coefficients. Correlation concerns the similarity between two persons, events or things. Correlation coefficient is a statistics that helps to describe with numbers, the degree of relationship between two sets or pairs of scores. Positive correlations are between 0.00 and + 1.00. While negative correlations are between 0.00 and – 1.00. Correlation at or close to zero shows no reliability; correlation between 0.00 and + 1.00, shows some reliability; and correlation at + 1.00 shows a perfect reliability. The more the correlation coefficient approaches 1.00 the more the test is reliable.

Factors affecting the test reliability

Some of the factors that affect reliability include:

- The relationship between the objective of the tester and that of the students.
- The clarity and specificity of the items of the test.
- The significance of the test to the students.
- Familiarity of the tested with the subject matter.
- Interest and disposition of the tested.
- Level of difficulty of items.
- Socio-cultural variables.
- Practice and fatigue effects.

CHECK YOUR PROGRESS

1) With relevant examples, distinguish test validity from test reliability.
2) Discuss different types of test validity and explain how you can establish them.
3) Discuss different types of test reliability and how you can establish them.
4) The following are sets of scores obtained from the tests administered to 10 students twice. Calculate the reliability of that test.

   Test A: 12, 13, 10, 8, 7, 6, 6, 4, 3, 1.

   Test B: 7, 11, 3, 7, 2, 12, 6, 2, 9, 6.

3. Item Analysis

What is item analysis?

Item analysis is a process which examines student responses to individual test items (questions) in order to assess the quality of those items and of the test as a whole i.e. item analysis gives information about the **difficulty level** of a question and it indicates how well each question shows the difference (**discriminate**) between the bright and dull students. In essence, item analysis is used for reviewing and refining a test. The two most common statistics reported in an item analysis are the **item difficulty**, which is a measure of the proportion of examinees who responded to an item correctly, and the **item discrimination**, which is a measure of how well the item discriminates between examinees who are knowledgeable in the content area. To carry out item analysis, you need to:

- Arrange the scored papers in order of merit – highest and lowest
- Select the 1/3(i.e.33%) upper group
- Select the 1/3(i.e. 33% lower group
- Item by item, calculate the number of students that got each item correct in each group.
- Estimate Item Difficulty Index and/or Item Discrimination Index

a) Item Difficult

Item difficulty is the percentage of students who answered a test item correctly. This means that low **item difficulty** values (e.g., 28, 56) indicate difficult **items**, since only a small percentage of students got the **item** correct.

How to calculate Item Difficulty Index?

The formula for discrimination index is as follows:

$$P = \frac{U+L}{N}$$

Where:

P= Item Difficulty Index

U= The number of students that got it right in upper group

L= The number of students that got it right in the lower group

N= The number of students actually involved in the item analysis

Interpretation of Item Difficulty Index

Difficulty Index	Percentage range	Interpretation
0.75 – 1.0	75%-100%	Easy
0.25 – 0.75	25%-75%	Average
0.25 or below	0-25%	Hard

b) Item Discrimination

It is the degree to which each question shows the difference (discriminate) between the bright and dull students. In other words, item discrimination is the degree to which students with high overall exam scores also got a particular item correct. It is often referred to as **Item Effect**, since it is an index of an item's effectiveness at discriminating those who know the content from those who do not. A test with many poor questions will give a false impression of the learning situation. Usually, a discrimination index of 0.4 and above is acceptable. Items which discriminate negatively are bad. This may be because of wrong keys, vagueness or extreme difficulty.

How to calculate Item Discrimination Index?

The formula for discrimination index is as follows:

$$D = \frac{U-L}{1/2N} \quad \text{OR} \quad \frac{U-L}{0.5N}$$

Where:

D= Item Discrimination Index

U= The number of students that got it right in upper group

L= The number of students that got it right in the lower group

N= The number of students actually involved in the item analysis

Interpretation of Item Discrimination Index

Discrimination Index	Interpretation
0.30 and above	Good
0.10 – 0.30	Fair
Equal to 0	No discrimination. All students got the item right
Negative	Poor. The item was flawed or miskeyed.

OTHER CHARACTERISTICS OF A GOOD TEST

4. Practicability: the ease of administration and scoring of a test.

5. Administrability: to ensure this, a test should be designed in the way that the scores obtained will not vary due to factors other than differences of the students' knowledge and skills.

6. Scorability: a good test should be easy to score, directions for scoring should be clear and the answer sheet as well as the answer key should be provided.

7. Comprehensiveness: a good test should encompass all aspects of a particular subject of study.

8. Simplicity: A good test should be easy to understand along with the instructions and other details.

9. Objectivity: a good test should not be influenced by emotion or personal prejudice. Lack of objectivity reduces test validity.

10. Test length: a good test should not be too long.

CHECK YOUR PROGRESS

1. Discuss different characteristics of a good test.

2. In the table below, determine the P and D for items 2, 3, 4 & 5. Item 1 has been calculated as an example. Total population of Testees is 60.

Item	1/3 upper group U=20	1/3 lower group L=20	Difficulty Index(P)= U+L/N	Discrimination Index(D)= U-L/0.5N
1	15	10	15+10/40=0.625 i.e. 62.5%	15-10/20=0.25
2	18	15		
3	5	12		
4	12	12		
5	3	10		

Lesson Four: Test Moderation and Marking

1. Test Moderation

Writing assessment instruments for use in exams requires skills. Sometimes the item seems clear to the person who wrote it but may not necessarily be clear to others. That is why the assessment instruments should be **moderated** before their empirical trial i.e. they should be reviewed by a review panel (with a number of people).

What is moderation?

Moderation is the process of teachers sharing their expectations and understandings of standards with each other in order to improve the consistency of their decisions about student learning and achievement. Moderation process involves teachers sharing evidence of learning and collaborating to establish a shared understanding of what quality of evidence looks like. It therefore helps teachers to either confirm or adjust their initial judgments. According to Maxwell (2002) “Moderation is concerned with the consistency, comparability and fairness of professional judgements about the levels demonstrated by students."

Guiding questions

For effective test moderation, consider the following questions:

- Is the test (assignment/exam) clear in each item? Is it likely that the person attempting a question will know what is expected?
- Are the items expressed in the simplest possible language?
- Are there unintended clues to the correct answer?
- Is each item a fair item for assessment at this level of education?
- Is the wording appropriate to the level of education where the item will be used?
- Is the format reasonably consistent so that students know what is required from item to item?
- Is there a single clearly correct (or best) answer for each item?
- Is the type of item appropriate to the information required?

- Are the items representative of the skills/knowledge to be assessed?

Benefits of moderation

a) For teachers

For teachers moderation has the following benefits:

- ✓ Moderation brings together collective wisdom, resulting in greater consistency of judgment, and focused teaching.
- ✓ Moderation provides greater confidence in teacher judgments, and assurance that judgments are consistent with those of other professionals.
- ✓ Moderation leads to shared expectations of learning, and understandings of standards and progression of learning.
- ✓ Moderation improves quality of assessment.
- ✓ Moderation aligns expectations and judgments with standards or progressions, and hence improved teaching and learning.
- ✓ Moderation assures parents and others that interpretations of students' achievements are in line with those of other professionals.

b) For the leadership teams

For the leadership moderation has the following benefits:

- ✓ Greater confidence in teachers' judgments and assurance that judgments are consistent within and across schools.
- ✓ Provides useful, dependable information for target setting.
- ✓ Provides information that can shape future professional development needs for teachers.

2) Marking

What is marking?

Marking or scoring is the process of awarding a number (usually), or a symbol to represent the level of student learning achievement. According to Sadler (2005) the most common method is by adding up the number of correct answers on a test, and assigning a number that

correlates. Higher numbers reflect better quality work. As a rule, **marking applies to students' level of performance in individual assessment tasks**, not to overall achievement in a course.

How to minimise discrepancies, biases and subjectivity in marking?

- ✓ Always refer to the marking scheme
- ✓ Mark one item/question for all copies before proceeding to the next item
- ✓ Marking should be when someone is in good conditions both physically and psychologically
- ✓ Do not change your standard as you mark. There may be a tendency of marking first papers objectively and last ones with flexibility
- ✓ Do not mark for more than two hours at a time and do not mark when you are tired.

Consistency in Marking

In order to attain some level of consistency in marking, the following have to be considered:

- ✓ All markers must understand the marking scheme and award marks consistently
- ✓ Alternative methods and answers not in marking scheme should be discussed before amending the marking scheme accordingly
- ✓ All markers should note amendments made in marking scheme

What is a 'Marking scheme'?

A marking scheme is a tool for marking/assessing students' work through indicators. To standardize the conditions of marking students' work, the marker should be guided by a marking scheme.

Functions of marking scheme

As a tool for marking students' work, a marking scheme is primarily used for the following reasons:

- ✓ Ensuring maximum objectivity in the marking;
- ✓ Standardizing the marking;
- ✓ Allowing a possible highest inter-examiner agreement

- ✓ Providing support for new teachers not for depersonalizing them with regard to marking but giving them some tools to get them change their views about the students' production

What should be considered when setting up a marking scheme?

The marking scheme can be considered qualitatively or quantitatively.

a) In a qualitative perspective, it provides examiners/markers with a list of qualitative indicators. Example: **clarity of writing**, **absence of spelling errors**, etc.

b) In a quantitative sense, it establishes the link between production and marks by setting thresholds: So, one has to bear in mind the following aspects:

- ✓ Is complete information extracted/retrieved?
- ✓ Is relevant information extracted/retrieved from documents?
- ✓ The marking scheme must be accurate, practical and simple
- ✓ It must not be heavy; it should only target some "key" benchmarks

CHECK YOUR PROGRESS

1. Differentiate moderation from marking.
2. Discuss the benefits of moderation for both the teacher and the school leadership
3. Discuss different ways you can ensure objectivity in marking students' work.
4. Explain different roles (functions) of a marking scheme.

Lesson Five: Grading system

According to Sadler (2005) Grading is the grouping of student academic work into bands of achievement. Grading usually occurs at a larger level, for example: significant assessment tasks, entire modules or courses and again is represented by a symbol. The most common grading symbols are A,B,C,D etc and HD, D, C, P (High Distinction, Distinction, Credit, Pass) etc. In fact, the most precious and valuable records after evaluation are the marked scripts and the transcripts of a student. At the end of every examination e.g. semester examination, the marked scripts are submitted through the head of department or faculty to the Examination Officer. Occasionally, the Examination Officer can round off the marks carrying decimal, either up or down depending on whether or not the decimal number is greater or less than 0.5

The marks so received are thereafter translated/interpreted using the Grade Point (GP), Weighted Grade Point (WGP), Grade Point Average (GPA) or Cumulative Grade Point Average (CGPA).

CREDIT UNITS

Courses are often weighed according to their credit units in the course credit system. Credit units of courses often range from 1 to 4. This is calculated according to the number of contact hours as follows:

1 credit unit = 15 hours of teaching.

2 credit units = 15 x 2 or 30 hours

3 credit units = 15 x 3 or 45 hours

4 credits units = 15 x 4 or 60 hours

Number of hours spent on practicals is usually taken into consideration in calculating credit loads.

GRADE POINT (GP)

This is a point system which has replaced the A to F Grading System as shown in the summary table below.

WEIGHTED GRADE POINT (WGP)

This is the product of the Grade Point and the number of Credit Units carried by the course i.e. WGP= GP x No of Credit Units.

GRADE POINT AVERAGE (GPA)

This is obtained by multiplying the Grade Point attained in each course by the number of Credit Units assigned to that course, and then summing these up and dividing by the total number of credit units taken for that semester (total registered for).

$$GPA = \frac{\text{Total Points Scored}}{\text{Total Credit Units registered}}$$

$$= \frac{\text{Total WGP}}{\text{Total Credit Units registered}}$$

CHECK YOUR PROGRESS

1. Discuss the difference that exists between the Grade Point (GP), Weighted Grade Point (WGP), and Grade Point Average (GPA).
2. Discuss the purpose of grading.

Lesson seven: Interpretation of classroom test score

1. Frequency distribution and percentages

A frequency distribution represents a count of the frequency with which each score occurs within a group of scores. A frequency distribution is constructed by first listing all the possible scores from highest to the lowest. Then you tally the frequency of occurrence of each score and record the total frequency to the right.

Test scores	Tally	Frequency
43	I	1
42	0	0
41	II	2
40	I	1
39	0	0
38	0	0
37	0	0
36	I	1
35	0	0
34	0	0
33	I	1
32	I	1
31	I	1
30	I	1
29	I	1
28	I	1
27	I	1
26	II	2
25	II	2
24	IIII	4
23	I	1
22	~~HII~~	5
21	~~HII~~	5
20	III	3
19	~~IIII~~ II	7
18	~~IIII~~	5
17	~~IIII~~	5
16	IIII	4
15	~~HII~~	5

It is worthy to mention that a large number of scores can even be more efficiently represented if they are displayed in a grouped frequency distribution. A grouped frequency distribution consists of counting the frequency of scores that occur in each of small number of class intervals. A class interval refers to a range of values smaller than the overall range of values under consideration. A grouped frequency distribution is constructed by counting the scores

that occur in each of the class intervals. To this end, to construct a grouped frequency distribution you must first identify the size of the class interval. This decision is very important because the size of the class interval determines how efficiently the scores are represented. Thus selection of the class interval size determines how well a group of scores will be summarized. Unfortunately, it impossible to tell you exactly how many class intervals you should use with a given set of scores. The ideal number of class intervals is one that economically presents the scores and at the same time allows you to obtain a clear picture of the data. Most of the data that social and behavioral scientists collect can be accommodated by 10 to 20 class intervals. However, this is merely a guideline. You must still decide on the exact number of class intervals to use. In general, you should select fewer class intervals as the number of scores decreases and more class intervals as the number of scores increases.

Class interval size

Once you have decided on the number of class intervals to use, you must determine the size of each class interval. The size of the class interval refers to the number of intervals of a given unit size that will constitute a given class interval. The best way to determine the size of the class interval is to subtract the lowest score in your set of data from the highest score and add 1 score unit. This gives number of intervals in the entire distribution with a unit size equal to 1. For the test scores in previous table, the number of intervals would be calculated as follows:

Number of intervals = 43-15+1

= 28+1

= 29

You now know that 29 intervals exist in the distribution. This number is then divided by the estimate of the appropriate number of class intervals, which gives the class interval size. Since you had previously determined that about 10 class intervals were needed, the class interval size would be as follows:

Class interval size: = 29/10 =2.9. This rounds to 3 so the class interval size is 3, as illustrated in the following table:

Class interval	Real limits	Tally	Frequency
42 - 44	41.5 – 44.5	I	1
39 – 41	38.5 – 41.5	III	3
36 – 38	35.5 – 38.5	I	1
33 – 35	32.5 – 35.5	I	1
30 – 32	29.5 – 32.5	III	3
27 – 29	26.5 – 29.5	III	3
24 – 26	23.5 – 26.5	𝍸 III	8
21 – 23	20.5 – 23.5	𝍸 𝍸 I	11
18 – 20	17.5 – 20.5	𝍸 𝍸 𝍸	15
15 – 17	14.5 – 17.5	𝍸 𝍸 IIII	14

Class Interval Real Limits

After you have identified the size of the class interval and the number of class intervals, you must specify the real limits of each class interval. The term 'real limits of a number' represents those points falling half a unit above and half a unit below the number. Therefore, if the number you are interested in is 6, the real limits of the number 6 are 5.5 and 6.5. For grouped data, the upper real limit of a specific class interval represents the upper real limit of the largest number in the class interval and the lower real limit of a specific class interval represents the lower real limit of the smallest number in the class interval. E.g. for class interval 42-44, the lower real limit is 41.5 and the upper real limit is 44.5.

Cumulative Frequency Distribution

It is a distribution that represents the cumulative frequency of scores below the upper real limit of the class interval of interest. To do this, you must sum the frequencies below the upper real limits of each class interval and obtain the cumulative frequency for each class interval.

Relative Frequency Distribution

It is a distribution that represents the proportion of cases at each score value or class interval. The advantage of using the relative frequency is that you can identify the pattern of scores independent of the number of cases involved. . Thus, using the relative frequency data, you can state that 25% of the testees obtained marks between 18 and 20.

Cumulative relative Frequency Distribution

It is a distribution that represents the cumulative proportion of cases at each score value or class interval. The advantage of the cumulative relative frequency is that you can identify the proportion of cases below the upper real limits of a class interval. Thus, from table following you can immediately tell that 66% of the testees obtained marks equal to or less than 23.5.

Class interval	**Real limits**	**Tally**	**Frequency**	**Cumulative Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
42 - 44	41.5 – 44.5	I	1	60	0.02	1.00
39 – 41	38.5 – 41.5	III	3	59	0.05	0.98
36 – 38	35.5 – 38.5	I	1	56	0.02	0.93
33 – 35	32.5 – 35.5	I	1	55	0.02	0.91
30 – 32	29.5 – 32.5	III	3	54	0.05	0.89
27 – 29	26.5 – 29.5	III	3	51	0.05	0.84
24 – 26	23.5 – 26.5	~~IIII~~ III	8	48	0.13	0.79
21 – 23	20.5 – 23.5	~~IIII~~ ~~IIII~~ I	11	40	0.18	0.66
18 – 20	17.5 – 20.5	~~IIII~~ ~~IIII~~ ~~IIII~~	15	29	0.25	0.48
15 – 17	14.5 – 17.5	~~IIII~~ ~~IIII~~ IIII	14	14	0.23	0.23

2. Measures of central tendency

a) Mode: It is the most frequently occurring score in a distribution of scores. If you want to identify the modal score, you merely identify the score that occurred most frequently. For example in the following set of scores: 75, 60, 78, 75, 76, 75, 88, 75, 81, 75, the the most occurring score i.e. the mode is **75**.

N.B. Sometimes a distribution of scores is multimodal, which means that more than one mode exists in a distribution. The mode is the poorest measure of central tendency measures because it is not affected by scores outside the modal interval.

b) Median: It is the measure of central tendency that divides a distribution of scores exactly in half so that 50% of scores fall below the median and 50% of the scores fall above the median.

Calculation methods:

- For **odd numbers**, it is determined by first arranging numbers/scores in order and then picking the middle score or number. e.g. for the following set of scores: 9, 14, 5, 7, 8, 3, 16. First, you rank the scores in order: 3, 5, 7, 8, 9, 14, 16, then you pick the middle score i.e. 8 as the median. Hence, the median of the distribution is **8.**
- For even numbers, it is determined by first arranging the scores in order, then picking the 2 middle scores, summing them, and dividing them by 2. e.g. for the following set of scores 9, 14, 5, 7, 8, 3, 16, 11, first rank order all the scores in the distribution: 3, 5, 7, 8, 9, 11, 14, 16, then sum up the two middle scores i.e. 8 & 9 and divide them by 2. Thus, the median for that distribution of scores equals 8+9/2= **8.5**

c) Arithmetic Mean: It is the arithmetic average of group of scores.

Calculation method:

The formula for calculating the arithmetic mean is as follows

$$\bar{X} = \frac{\Sigma X}{N}$$

Where:

Σ = Sum of

$\dot{X}$ = Arithmetic mean

N = Total number of scores

X= A score in distribution

Mode, Median, and Mean for grouped data

a) Mode = $Lmo + C \frac{(d1)}{(d1 + d2)}$

Where:

Lmo= lower real limit of the modal class

C= Class interval

d1= difference between frequency of the modal class and freq. of before the modal class

d2= difference between frequency of the modal class and freq. after the modal class.

b) Median = $LRL + W \frac{[N/2-CFB]}{FW}$ **Where:** LRL= lower real limit of the median class

CF= Cumulative frequency below the median class

FW= Frequency within the interval containing median

W= Width of the class interval

N= Number of cases

c) Mean = $\frac{\Sigma fx}{\Sigma f}$ Where: Σfx= Sum of frequencies times midpoint

Σf= sum of frequencies

Now, consider the following set of scores:

67, 76, 69, 68, 72, 68, 65, 63, 75, 69, 66, 72, 67, 66, 69, 73, 64, 62, 71, 73, 68, 72, 71, 65, 69, 66, 74, 72, 68, 69.

Class interval	Real limits	Mid points (x)	Frequency	fx	CF
74 – 76	73.5 – 76.5	75	3	225	30
71 – 73	70.5 – 73.5	72	8	576	27
68 – 70	67.5 – 70.5	69	9	621	19
65 – 67	64.5 – 67.5	66	7	462	10
62 – 64	62.5 – 64.5	63	3	189	3
			Σf = 30	Σfx = 2073	

a) Calculate the Mode for this distribution of score

b) Calculate the median for this distribution of scores

c) Calculate the mean for this distribution of scores

3. Measures of variability/dispersion

a) Range: as an index of variability the range is merely the distance between the highest and the lowest scores in any distribution of scores. Although the range does measure the variability of a group of scores, it is used very seldom because it is so unstable. This is due to the fact that the range depends on only the two extreme scores in the distribution, and if these two extreme scores change, the range changes.

Calculation method:

Range = Highest score – Lowest score

b) Variance: It is a measure of variability obtained by computing the average of the sum of squared deviations of scores about their mean.

Calculation method:

$$S^2 = \frac{\Sigma(X_i - \overline{X})^2}{N}$$

Where:

$\mathbf{S^2}$= Variance

N= Number of cases

c) Standard Deviation: It is a measure of variability obtained by taking the square root of the average of the squared deviations scores about their mean, i.e. it is the square root of the average of the squared deviations of scores from their mean. Therefore, it is merely the square root of the variance. A small standard deviation means that the group has small variability (deviation from the mean) or relatively homogeneous. A big standard deviation means that the group has large variability (deviation from the mean).

Calculation method:

$$S = \sqrt{\frac{\Sigma(Xi - \overline{X})^2}{N}}$$

Considering the following set of scores from a chemistry test: 1, 2, 3, 7, 8, 9,

a) Compute the range,
b) Compute the variance,
c) Compute the standard deviation

4. Measures of relationship

a) Pearson product-moment correlation coefficient (r)

The formula for calculation of Pearson product-moment correlation coefficient is as follows:

$$r = \frac{N(\Sigma XY)-(\Sigma X)(\Sigma Y)}{\sqrt{[N(\Sigma X^2)-(\Sigma X)^2]\,[N(\Sigma Y^2)-(\Sigma Y)^2]}}$$

b) Spearman rank order correlation coefficient (r_s)

Formula for calculation of Spearman rank order correlation coefficient

$$r_s = 1-\left(\frac{6\Sigma D^2}{N(N^2-1)}\right)$$

CHECK YOUR PROGRESS

1. Explain the difference between the frequency, cumulative frequency, relative frequency, and cumulative relative frequency in a distribution of scores.
2. Discuss the three measures of central tendency
3. Discuss the range, variance and standard deviation as measures of variability.
4. Explain how Pearson product moment correlation coefficient differs from the Spearman Rank order correlation coefficient.
5. Considering the following set of scores obtained from two tests administered to 5 students, using both Pearson and Spearman correlation coefficients compute the relationship between the two test and interpret it.

	A	**B**	**C**	**D**	**E**
Test 1: Chemistry	1	3	4	5	2
Test 2: Physics	2	3	5	4	3

REFERENCES

Allen M.J. & Yen W.M. (1976) *Introduction to Measurement Theory*. Belmont California: Wadsworth Inc.

Amin, E. M. (2005). *Social Science Research: Conception, Methodology and Analysis.* Kampala. Makerere University Printery.

Brown, F.G. (1970) *Principles of educational and psychological testing. 2nd ed*. New York: Holt, Rinehart & Winston.

Christensen, L.B. & Stoup, C.M. (1991). *Introduction to statistics for the social and behavioral sciences*. 2nd Ed. California: Brooks/Cole Publishing Company

Creswell, J.W. (2012). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research,* 4th Ed., Boston: Pearson.

Fraenkel, J.R. & Wallen N.E. (2009). *How to Design and Evaluate Research in Education.* 7th ed. New York: McGraw-Hill.

Glass, G.V. & J.C. Stanley (1970) *Statistical Methods in Education and Psychology*. New Jersey: Prentice-Hall.

Ingule, F. & Gatumu, H. (1996) *Essentials of Educational Statistics*. Nairobi: E.A. Educational Publishers.

Mehrens, W.A. & Lehmann, I.J. (1978) *Measurement and Evaluation in Education and Psychology*. New York: Holt, Rinehart and Winston.

Russell, M. K. & Airasian, P.W. (2012). *Classroom Assessment concepts and Application*. 7th Ed. New York, NY: McGraw-Hill.