African Virtual University

Education: Education EDU09

EDUCATIONAL EVALUATION AND TESTING

Ridwan Mohammed Osman

Foreword

The African Virtual University (AVU) is proud to participate in increasing access to education in African countries through the production of quality learning materials. We are also proud to contribute to global knowledge as our Open Educational Resources (OERs) are mostly accessed from outside the African continent. This module was prepared in collaboration with twenty one (21) African partner institutions which participated in the AVU Multinational Project I and II.

From 2005 to 2011, an ICT-integrated Teacher Education Program, funded by the African Development Bank, was developed and offered by 12 universities drawn from 10 countries which worked collaboratively to design, develop, and deliver their own Open Distance and e-Learning (ODeL) programs for teachers in Biology, Chemistry, Physics, Math, ICTs for teachers, and Teacher Education Professional Development. Four Bachelors of Education in mathematics and sciences were developed and peer-reviewed by African Subject Matter Experts (SMEs) from the participating institutions. A total of 73 modules were developed and translated to ensure availability in English, French and Portuguese making it a total of 219 modules. These modules have also been made available as Open Educational Resources (OER) on oer.avu.org, and have since then been accessed over 2 million times.

In 2012 a second phase of this project was launched to build on the existing teacher education modules, learning from the lessons of the existing teacher education program, reviewing the existing modules and creating new ones. This exercise was completed in 2017.

On behalf of the African Virtual University and our patron, our partner institutions, the African Development Bank, I invite you to use this module in your institution, for your own education, to share it as widely as possible, and to participate actively in the AVU communities of practice of your interest. We are committed to be on the frontline of developing and sharing open educational resources.

The African Virtual University (AVU) is a Pan African Intergovernmental Organization established by charter with the mandate of significantly increasing access to quality higher education and training through the innovative use of information communication technologies. A Charter, establishing the AVU as an Intergovernmental Organization, has been signed so far by nineteen (19) African Governments - Kenya, Senegal, Mauritania, Mali, Cote d'Ivoire, Tanzania, Mozambique, Democratic Republic of Congo, Benin, Ghana, Republic of Guinea, Burkina Faso, Niger, South Sudan, Sudan, The Gambia, Guinea-Bissau, Ethiopia and Cape Verde.

The following institutions participated in the teacher education program of the Multinational Project I: University of Nairobi – Kenya, Kyambogo University – Uganda, Open University of Tanzania, University of Zambia, University of Zimbabwe – Zimbabwe, Jimma University – Ethiopia, Amoud University - Somalia; Université Cheikh Anta Diop (UCAD)-Senegal, Université d' Antananarivo – Madagascar, Universidade Pedagogica – Mozambique, East African University - Somalia, and University of Hargeisa - Somalia

The following institutions participated in the teacher education program of the Multinational Project II: University of Juba (UOJ) - South Sudan, University of The Gambia (UTG), University of Port Harcourt (UNIPORT) – Nigeria, Open University of Sudan (OUS) – Sudan, University of Education Winneba (UEW) – Ghana, University of Cape Verde (UniCV) – Cape Verde, Institut des Sciences (IDS) – Burkina Faso, Ecole Normale Supérieure (ENSUP) - Mali, Université Abdou Moumouni (UAM) - Niger, Institut Supérieur Pédagogique de la Gombe (ISPG) – Democratic Republic of Congo and Escola Normal Superieur Tchicote – Guinea Bissau

Bakary Diallo

The Rector

African Virtual University

Production Credits

This second edition is the result of the revision of the first edition of this module. The informations provided below, at the exception of the name of the author of the first edition, refer to the second edition.

Author

Ridwan Mohamed Osman

Reviewers

Mohamed Sagayar Moussa

Augustine Mwangi

AVU - Academic Coordination

Dr. Marilena Cabral

Module Coordinator

Salomon Tchameni

Instructional Designers

Elizabeth Mbasu

Diana Tuel

Benta Ochola

Media Team

Sidney McGregor

Barry Savala

Edwin Kiprono

Kelvin Muriithi

Victor Oluoch Otieno

Michal Abigael Koyier

Mercy Tabi Ojwang

Josiah Mutsogu

Kefa Murimi

Gerisson Mulongo

Copyright Notice

This document is published under the conditions of the Creative Commons
http://en.wikipedia.org/wiki/Creative_Commons

Attribution http://creativecommons.org/licenses/by/2.5/

Module Template is copyright African Virtual University licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. CC-BY, SA

Supported By

AVU Multinational Project II funded by the African Development Bank.

Table of Contents

Foreword 2

Production Credits 4

Copyright Notice 5

Supported By 5

Prerequisite 8

Time 8

Material 8

Module Rationale 8

Overview 9

Outline 9

Unit I: Educational Evaluation 9

Unit II: Specification of Objectives 9

Unit III: Classification and Selection of Tests 9

Unit IV: Development of an Instrument 9

Unit V: Analysis, Interpretation and Use of Results 10

VII. General objectives 10

VIII. Specific learning objectives 11

IX. Pre-assessment 14

Pre-assessment 15

Pedagogical information for the Learners 19

Learning activities 26

Learning activity # 1 26

Formative Assessment 34

Activity 1 34

Activity 2 34

Activity 3 35

Learning Activity #2 36

Classification and Statement of Objectives 38

The Cognitive Domain 39

THE EFFECTIVE DOMAIN 39

The psychomotor domain 40

Formative Assessment 43

Learning Activity #3 46

Content 47

Measurement and Testing 47

Formative Assessment 61

Learning Activity #4 62

Learning Activity #5 69

Formative Assessment 76

XIII. Summative evaluation 77

XIV. Synthesis of the Module 79

XV. Author of the Module 80

XV. Reviewer of the Module 80

Prerequisite

The prerequisites for this module are General Psychology, Educational Psychology and Learning Psychology.

Time

The total time required to complete this module is 120 hours. The time is broken down as follows:

Unit I: 25 hours

Unit II: 25 hours

Unit III: 25 hours

Unit IV: 25 hours

Unit VI: 20 hours

Material

A Computer with internet connection and basic operating and application software.

Module Rationale

Evaluation is an integral part of education. It affects the way students study, their motivation, performance and their aspirations. At school levels, it enables inspectors and officials from the Ministry assess the quality of education offered in a specific school, region or throughout the country. At curriculum level, it gives us an idea of whether the curriculum we have set up – means – is leading us to our educational goals/objectives – ends. For you as a teacher, this module will give an idea of the importance of evaluation in education. It would also help you come up with appropriate means of measuring and evaluating students.

Overview

This module – Educational Evaluation and Testing – is intended to enable you have the ability to understand the principles and concepts of the different types of educational evaluation. This is introduction course, and therefore, you will not be expected to become an expert in educational evaluation and testing. However, you will be able to understand the processes of evaluation, measurement and testing in a better a manner. You will be able to plan and conduct evaluation and testing on the basis of sound principles and practices. By the end of the module, you will be able to understand the interconnections between/among curriculum, goals, objectives, testing and evaluation. You will also have new perspective on how to use evaluation results.

Outline

Unit I: Educational Evaluation

The nature of evaluation

Types of evaluation

Phases of evaluation

Unit II: Specification of Objectives

The nature of objectives

Classification and statement of objectives

Unit III: Classification and Selection of Tests

Classification of a test

Selection of a test

Unit IV: Development of an Instrument

Designing the test

Constructing test items

Test construction and reproduction

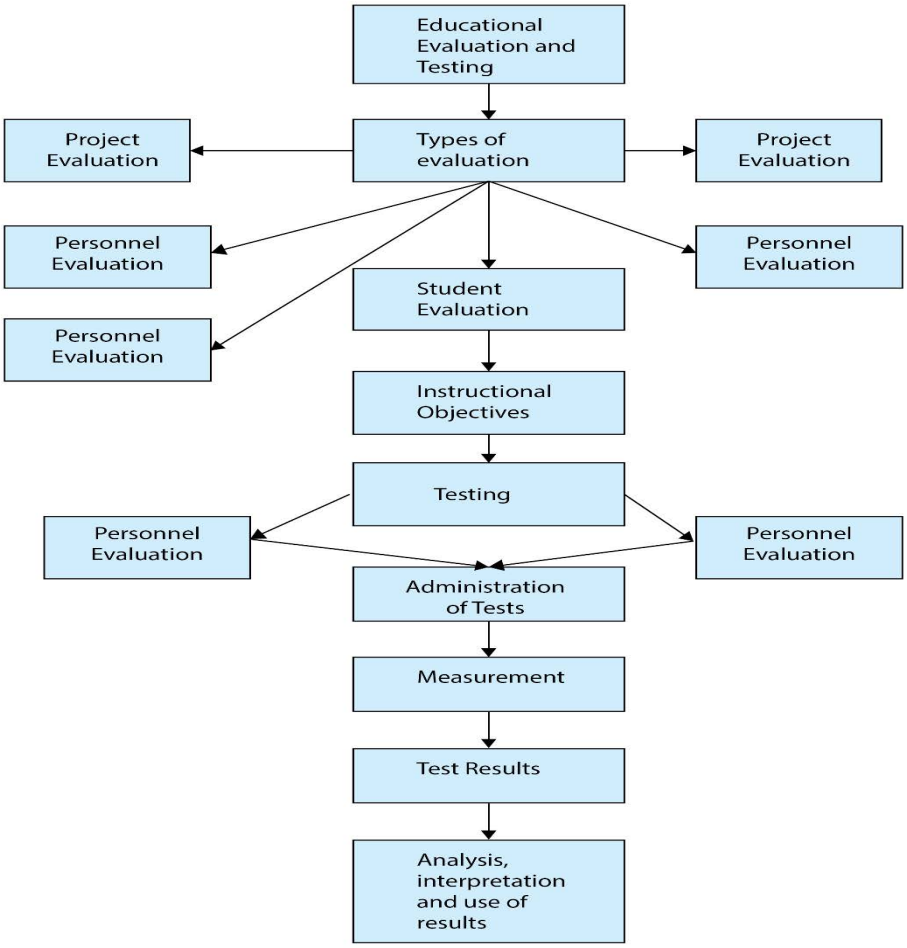
Test revision

Unit V: Analysis, Interpretation and Use of Results

Analysis

Interpretation

Using and reporting results



VII. General objectives

At the end of this course, students are expected to

Demonstrate an understanding of basic principles of educational evaluation and testing

Construct, analyse and improve assessments that adequately appraise general or specific instructional objectives

Critique and interpret standardized test results

Integrate assessment results to evaluate student performance

Interpret assessment findings for students, parents and community

Construct and interpret tests

VIII. Specific learning objectives

Unit	Unit Subtopic	Learning Objectives
1.	Educational Evaluation	
	The nature of evaluation Types of evaluation Phases of evaluation	Define and describe evaluation Describe the major steps in the evaluation process Define and describe the terms measurement, Evaluation and testing List the components that are common to all valid Models of evaluation Describe the different types of evaluation by identifying the major variables measured in each types and decisions made Define the different phases of evaluation and list the activities involved in each phase
2.	Specification of Objectives	

	The nature of objectives Classification and statement of objectives	List the basic components of any valid needs assessment model List three factors to be considered when assigning priorities to identified needs State and describe the different types of objectives State guiding principle for determining the degree of specificity required for an objective Select an area of interest and write related objective For each major category of the cognitive component Select an area of interest and write related objective For each major category of the affective component Select an area of interest and write related objective for each major category of the psychomotor component
3.	Classification and Selection of Tests	

	Classification of a test Selection of a test	State three major ways to collect data and identify one situation for which each is appropriate List four major differences between a standardized test and locally-developed test State the major ways in which tests are classified and how each is used State the characteristics of good tests List the factors which should be considered when selecting a test from a number of alternatives
4.	Development of an Instrument	

	Designing the test Constructing test items Test construction and reproduction Test revision	Identify and describe the major components of a table of specifications Identify and describe the general guidelines for item construction Describe the types of outcomes for which an essay test is most appropriate Briefly describe two approaches to the scoring of essay tests List two major guidelines for the construction of objective tests State specific objective and a corresponding item for each of the following types of tests: essay, short answer, multiple choice, true-false, and matching. Describe the major approaches to the organization of test items List major guidelines concerning test format Identify and describe the major aspects of test validation Identify and discuss the major aspects of item analysis appropriate for an NRT Discuss the results of item analysis for an NRT are interpreted Identify and discuss the major aspects of item analysis appropriate for an CRT Discuss the results of item analysis for an CRT are interpreted
5.	Analysis, Interpretation and Use of Results	

	Analysis Interpretation Using and reporting results	Describe the four scales of measurement and list three examples of each Define and describe the following types of statistics and how each is interpreted; central tendency, variability, normal distribution, percentile ranks, z score, T score and correlation Define and describe grade equivalent and age equivalents Describe and discuss the different methods of interpreting data generated from an exam Briefly state the overall major use of test results State four general principles of reporting Identify and briefly describe the major components of a formal evaluation report

IX. Pre-assessment

Rationale

This pre-assessment for the module is intended to check your level of competency of the content covered in this module. After you do the pre-assessment you will be able to know the skills and knowledge relevant to the module that you already possess or don't know. In any way, the pre-assessment will give you a hint of the level of effort that you need to invest in this module.

Pre-assessment

1. The systematic process of collecting and analysing data in order to make decision is called

   a. Evaluation

   b. Measurement

   c. Assessment

   d. Testing

2. The process of quantifying the degree to which someone or something possesses a given trait is called

   a. Evaluation

   b. Measurement

   c. Assessment

   d. Testing

3. Situation analysis is a component of the phase of evaluation.

   a. Planning

   b. Process

   c. Product

4. Mastery objectives are intended to communicate

   a. The relative priority of needs

   b. Behaviours believed to be required or successful performance

   c. Behaviours believed to be achieved by all students

   d. The complexity of objectives

5. Which of the following domain of educational objectives entails physical abilities?

   a. Cognitive domain

   b. Affective domain

   c. Psychomotor domain

6. The most discrete, measurable objectives are known as

   a. Vision

   b. Goals

c. Intermediate objectives

d. Specific objectives

7. Which of the following should be a component of educational objectives?

a. Behaviour

b. Condition

c. Criterion

d. All of them

8. A test item in which the respondents should compose the answers is?

a. Essay

b. Multiple choice

c. True-false

d. Matching

9. Performance tests are intended to measure

a. Attitudes

b. General intelligence

c. The process or product of an activity

d. Personality

10. Achievement tests measure

a. Student potential

b. Personal characteristics

c. Proficiency in knowledge or skills

d. Perseverance and patience

11. Norm-referenced standards interpret each score relative to

a. Scores of others on the same standard

b. Difficulty of the exam

c. Already developed criteria

d. Student background

12. Which of the following is NOT characteristic of a good test?

a. Validity

b. Reliability

c. Comprehensiveness

d. Difficulty

13. Validity of a test means

a. An individual's score is the same regardless of who is scoring

b. Detecting the small differences in the achievement of students

c. The extent to which a test is measuring what it is supposed to measure

d. The consistence of the test

14. A test that produces the same results when administered repeatedly is said to be

a. Valid

b. Reliable

c. Comprehensive

d. Discriminatory

15. Table of specifications is used to

a. Ensure that all intended outcomes are measured

b. Help students understand the instructions better

c. Promote the reliability of a test

d. Balance the difficulty of a test

16. A number that describes student's location on an achievement continuum is called

a. Raw score

b. Percent correct

c. Grade equivalent

d. Standard score

17. Which of the following processes affects the reliability of a test?

a. Administration

b. Scoring

c. Standards

d. Instruction

e. None

18. Content-related validity is a types of validity, in which

a. The test is demonstrated to be effective in predicting criterion or indicators of a construct

b. The items of a test represent the entire range of possible items the test should cover

c. An association between the test scores and the prediction of a theoretical trait is demonstrated

d. None of the above

19. Data for evaluation can be collected through

a. The administration of standardized instruments

b. The administration of locally-developed instruments

c. Observation

d. All of the above

20. Which of the following is not a purpose of testing?

a. To describe each student's developmental level within a test area

b. To identify a student's areas of relative strength and weakness in subject areas

c.To monitor year-to-year growth in the basic skills

d. To predict the future success of the individual

Answers

1. a

2. b

3. b

4. c

5. c

6. d

7. d

8. a

9. c

10. d

11. a

12. d

13. c

14. b

15. a

16. c

17. e

18. b

19. d

20. d

Pedagogical information for the Learners

As indicated earlier, the aim of the pre-assessment was to help you realize your master of the content of this module. If you answered more than 10 questions of the pre-assessment correctly, you are already have considerable mastery of the concepts and principles underlying educational testing and evaluation. Going through the activities of this module will help improve your level of knowledge and skills in the area. If you scored between 5 and 10 of the questions correctly, you are in a good position to study this module. You have good idea of the major concepts, but still you need to master the content of the module. If you answered less than 5 questions of the pre-assessment correctly, you need to spend more time and effort in mastering this module, if it is to help your level of competency in the area of educational measurement and testing.

Compiled list of Key Concepts (Glossary)

Evaluation is the systematic process of collecting and analysing data in order to make decisions.

Measurement is the process of quantifying the degree to which someone or something possesses a give trait, i.e. quality, characteristic or feature.

Pre-requisites: behaviours, knowledge and skills that are assumed to have been acquired by students

Objectives are specific statements of what is to be accomplished and how well, and are expressed in terms of quantifiable, measurable outcomes.

Needs assessment is the process of identifying needs and establishing relative priority of those needs.

Competency-based objectives are intended to communicate behaviours believed to be required for successful job performance and to have students demonstrate these behaviours prior to graduation.

Mastery objectives are objectives which should be achieved by virtually all students, regardless of ability or background

Taxonomy of educational objectives is intended to classify all objectives into a hierarchy of categories based on presumed complexity

Affective Domain deals with outcomes that are much more difficult to promote and to measure, intangibles such as feelings, attitudes, interests and values.

Psychomotor Domain entails physical abilities, those involving muscular or motor skills, manipulation of objectives, or neuromuscular coordination.

Goals: Goals are global statements of long-term outcomes.

Intermediate Objectives are more precise than general objectives but not as detailed as specific objectives.

Specific Objectives deal with the most discrete intended outcomes.

An essay test is one in which the number of questions is limited and res- ponders must compose answers,

An objective test is one for which subjectivity in scoring is eliminated, at least theoretically.

Standardized test is one that is developed by subject matter and measurement specialists,

Locally-developed test is usually developed locally by teachers for a specific purpose.

Self-report data consist of oral or written responses from individuals.

Rating scale is an instrument with a number of items related to a given variable, each item representing a continuum of categories between two extremes.

Achievement tests are tests that measure the correct status of individuals with respect to proficiency in given areas of knowledge or skills.

Character or Personality tests are designed to measure characteristics of individuals along a number of dimensions

Aptitude tests are measures of potential that are used to predict how well someone is likely to perform in a future situation.

Performance standards are the criteria to which the results of measurement are compared in order to interpret them.

Norm-referenced test is any test, standardized or locally-developed, which reports and interprets each score in terms of its relative position with respect to other scores on the same test,

Criterion-referenced test is any test which reports and interprets each score in terms of an absolute standard is criterion-referenced.

Test objectivity means that an individual's score is the same, or essentially the same, regardless of who is doing the scoring.

Discrimination is the ability of a test to detect or measure small differences in achievement or attainment.

Comprehensiveness indicates the extent to which a test samples major lesson objectives that are to be measured

Validity is the extent to which a test measures what it is intended to measure.

Content validity refers to the extent to which items on the test represent the entire range of possible items the test should cover.

Criterion-related validity refers to the extent to which a test is demonstrated to be effective in predicting criterion or indicators of a construct.

Concurrent validity occurs when the criterion measures are obtained at the same time as the test scores.

Predictive validity occurs when the criterion measures are obtained at a time after the test

Construct validity refers the extent to which a test demonstrates an association between the test scores and the prediction of a theoretical trait.

Reliability refers to the consistency of a measure. A test is considered reliable if we get the same result repeatedly.

A table of specifications is essentially a blueprint for a test with the purpose to insure that all intended outcomes, and only the intended outcomes, are measured and that the test includes the appropriate number of items for each measured item.

Test validation is the process of administering and revising over and over again until acceptable levels of validity and reliability are achieved.

A raw score is the number of questions a student gets right on a test (assuming each question is worth one point).

Percent Correct is obtained when the raw score is divided by the total number of questions and the result is multiplied by 100.

Grade equivalent is a number that describes a student's location on an achievement continuum.

Percentile rank is a score that tells the percent of students in a particular group that got lower raw scores on a test than the student did.

List of Compulsory Readings

Reading #1

Title: Designing Evaluation for Education Projects

Complete Reference: http://wateroutreach.uwex.edu/use/documents/NOAAEvalmanualFINAL.pdf. Retrieved 25/03/2016

Abstract

The reading goes through the basics of evaluation, providing discussions on everything from types of evaluations and ways of collecting information to using an outside evaluator and the ethical considerations of gathering data from program participants. This information is intended to answer questions about project evaluation and provide guidance in using evaluation as a project improvement tool.

Rationale

A considerable amount of time, effort, and other resources go into the development and implementation of education projects. Quite obviously, the goal is to create effective projects that can serve as models of excellence. Whether the project is an hour-long endangered species walk, a family day festival, marine resources monitoring, or a community forum, the aim of providing quality educational experiences remains the same. The aim of this reading is to help you – as the future teachers, administrators, project managers and policy makers – understand the process of evaluating projects to improve their effectiveness and impact.

Reading #2

Title: The Case against Standardized Tests

Complete Reference: www.alfiekohn.org/books/tcast.htm

Retrieved 25/03/2016

Abstract

The article criticizes standardized tests by outlining the major weakness found in the literature. How valid are test scores are predictors of grades? Do they have any validity as predictors of actual accomplishment? Are the tests biased against certain members of society? This article reviews the extensive critical literature on the subject of standardized tests in an attempt to answer these questions.

Rationale

The article will enable you, as the learner, see the negative side of standardized tests. Such weaknesses noted in the article will be applicable when you encounter the challenged of choosing between standardized tests and locally-developed tests.

Reading #3

Title: Test Construction: Some Practical Ideas

Reading #3

Title: Improving Essay Tests

Complete Reference: http://ideaedu.org/wp-content/uploads/2014/11/Idea_Paper_17.pdf

Retrieved 25/03/2016

Abstract

The article which focuses on the development of essay tests gives comprehensive overview of essay-type tests. It starts by describing what an essay test is. In addition, the article outlines the strengths and limitations of essay tests. Recommendations concerning when to use essay tests and how to construct them are discussed. At the end of the article, the process of scoring essay tests is given.

Rationale

The article will enable the learners develop an idea of how to develop essay-type items that are appropriate for classroom testing. It would give the reader an idea of the major weaknesses that curtail the effectiveness of essay tests as classroom testing tools.

Reading #4

Title: Designing Objective Test Questions: An Introductory Workshop Complete Reference: http://www.caacentre.ac.uk/dldocs/otghdout.pdf

Abstract

Retrieved 25/03/2016

The paper gives introduction of the different types of objectives test items and how to construct them. The types of items discussed include Multiple Choice Questions (MCQ), True-False, Assertion Reason, Multiple-Response Questions (MRQ), and matching questions.

Rationale

This article will give learners a general concept of the different types of objective test questions. In each category, examples are provided so that it would be easier for you to construct items of that style.

Reading #5

Title: How to Prepare Better Multiple – Choice Test Items: Guidelines for

University Faculty

Complete Reference: https://testing.byu.edu/handbooks/betteritems.pdf

Retrieved 25/03/2016

Abstract

The booklet is entirely dedicated to Multiple-Choice items, the most important type of objective test items. It discusses the structure of MCQs, the advantages and limitations, when they should be used and their varieties. The article discusses how MCQs can be used to measure higher-order skills. The booklet, at the end, provides some guidelines on how to construct such type of test items.

Rationale

After closely studying and applying, the booklet is expected to enable the learner distinguish between instructional objectives that can best be measured by using MCQ, evaluate MCQ items by using commonly-accepted criteria by identifying the flaws in them and recommend ways to improve them. Most important of all, the learner will be able to construct well-written MCQs that measure specific objectives.

Reading #6

Title: Best Practice in Mathematics: Using Test Results to Inform Instruction and Improve Student Achievement

Complete Reference:

http://www.ctb.com/media/articles/pdfs/resources/ctb_best_practice_in_math.pdf

Retrieved 25/03/2016

Abstract

The article discusses how teachers can use test results to inform instruction and improve academic achievement. A number of guidelines are given on the teacher can use test results. Example of a class report is given.

Rationale

The article will help teachers think of the ways they can use test results, so that their instructions is aligned with student needs and weaknesses.

Reading #7

Title: Descriptive Statistics

Complete Reference:

http://www.sagepub.com/sites/default/files/upm-binaries/9881_040143ch02.pdf

Retrieved 25/03/2016

Abstract

The article discusses with elaborate examples from social sciences the concepts underlying descriptive statistics. These include types of data, visual descriptive statistics, and numerical descriptive statistics.

Rationale

The article will give learners solid understanding concerning descriptive statistics and the skill to apply these concepts to their everyday exam.

Learning activities

Learning activity # 1

EDUCATIONAL EVALUATION

SUMMARY

The first activity in this module is intended to provide students basic understanding of educational evaluation and testing. Basic terminology in the subject will be discussed. Also, types of evaluation would be considered. In addition, the learning activity will also cover the different phases of evaluation, starting with the preparation and the finalization of the process. The readings and exercises incorporated in this learning activity shall enable the learner achieve the objectives set for the corresponding units of the module.

List of Key Concepts

Evaluation is the systematic process of collecting and analysing data in order to make decisions.

Measurement is the process of quantifying the degree to which someone or something possesses a give trait, i.e. quality, characteristic or feature.

Pre-requisites: behaviours, knowledge and skills that are assumed to have been acquired by students

CONTENT

Evaluation and Measurement

Evaluation is an integral component of all systems of education at all processes. It is what enables educators, teaches, administrators, policy makers and the community have an idea of what is missing and what is available. Evaluation can be defined in two ways, depending on what we want to achieve at end of the exercise. Evaluation is the systematic process of collecting and analysing data in order to determine whether, and to what degree objectives have or are being achieved.

Evaluation is the systematic process of collecting and analysing data in order to make decisions.

The first part of these definitions (systematic process of collecting and analysing data) is common to both of the definitions provided here. However, the basic difference between the two definitions is the issue of whether decisions or judgments are an integral component of evaluation. The second definition seems to be more inclusive, as it does not preclude the

activities implied in the first definition.

In most cases, evaluation is intended to answer questions that require some degree of valuing. Questions, such as

- Is the special program worth what it costs?
- Is the new experimental reading curriculum better than the previous one?
- Should Mr. Ahmed be place in a program for the gifted?
- The function or purpose of evaluation is to
- Determine the current status of the object of evaluation
- Compare the status with a set of standards or criteria
- Select an alternative in order to make decision

At the end of the process, there may be only two alternatives or a combination of complex activities and programs designed as intervention for the current situation. In general, the process of evaluation involves determination of the types of data which needs to be collected, determination of the individual, group or groups from which data will be obtained, collection of data, analysis of data, interpretation of the data and decision-making.

Basing decisions on valid procedures is critical because although all decisions are not equally important, each one has a consequence which directly or indirectly affects students. The more serious the consequence, the more important the decision. Every educational decision, however, should be made on rational, objective basis to the greatest degree possible and should be based on the best data available.

The data collected during the evaluation process are only as good as the measurements upon which the measurements are based. Measurement is the process of quantifying the degree to which someone or something possesses a give trait, i.e. quality, characteristic or feature.

A measurement permit is a more objective description of traits and facilitates comparisons. Thus, instead of saying that Kimani is underweight for her age and height, we can say that Kimani is 16 years old, 5'8' tall, and weighs only 85 pounds. Further, instead of saying Kimani is more intelligent than Juma, we can say Kimani has a measures IQ of 125 and Juma has a measured IQ of 88. In each case, the numerical statement is more precise, more objective, and less open to interpretation than the corresponding verbal description.

There is more professional disagreement about whether all traits of interest can be measured. Can you really measure elusive qualities such as empathy, appreciation, motivation or interests? Can you really measure values and attitudes? The answer is yes, but it is not easy. Supporters of this position put forth the following argument: if something exists, it exists in quantities; if it exists in quantities, it can be measured. If you accept this logic, any trait of interest to educators can be measured. Thus, the purpose of educational measurement is to represent how much of 'something' is possessed by a person or entity.

Agreeing that at least theoretically all things can be measured and measuring them are two quite different processes. Valid measurements in education is not easy. The major problem is that, with the exception of measurement of physical characteristics such as height and weight, all measurement is indirect; there are no yardsticks or scales for measuring traits like intelligence, achievement or attitude. Assessment of such traits must be of necessity be based upon inference. The problem is further complicated by the lack of well-validated instruments. In some areas, the problem is not as serious as in others.

The term measurement is not synonymous with the administration of pen-and- pencil tests. Data may be collected via processes such as observation and analysis and rating of a product. In some cases, required data may already be available and retrievable from records. In many cases, however, some combination of standardized and/or self-developed tests is required.

Types of evaluation

When we speak of the types of evaluation, we are referring to the different processes, products and persons subject to evaluation. These include student, curricula, schools, school systems, large populations, special programs or projects and personnel. The fact that we speak of different types of evaluation does not mean that there are a number of different evaluation processes. The basic evaluation process is the same, regardless of what is being evaluated. What differs is what is being evaluated, how the evaluation process is applied and the types of decisions made. Depending upon what is being evaluated, different types of data will be collected, different criteria will be applied to the data, and different types of decisions will be made. But the basic evaluation process is the same and the same general concepts and principles of measurement and evaluation are applicable.

Student Evaluation

Achievement is one of the variables on which student is assessed; other major variables include aptitude, intelligence, personality, attitudes and interests. In order to assess achievement, tests, both standardized and teacher-made, are administered; projects, procedures and oral presentations are rated; and formal and informal observations are made. A teacher uses performance data not only to evaluate student progress but also to evaluate his/her own instruction. In other words, the process of evaluating students provides feedback to the teacher. Feedback on current student progress also gives direction to future instructional activities.

There are all kinds of decisions made which do directly affect the student; not all of these decisions are made by the teacher. The teacher decidesWhether the student has achieved objectives, or to what degree, and provides appropriate remedial work if necessary

Whether the student is working up to a potential.

What should be expected from a certain student or a group of students.

Whether the child has special needs which cannot be met in regular classroom and recommend that a child be placed in special environment

There are numerous other decisions that are made concerning students in which the teacher is either not involved or only makes recommendations. These decisions include

The placement of a child in a special program

Promotion of a student.

The employment of students and their admission to other educational institutions

Curriculum evaluation

Curriculum evaluation involves the evaluation of any instructional program or instructional materials, and includes evaluation of such factors as instructional strategies, textbooks, audio-visual materials, and physical and organizational arrangements. Curriculum evaluation may involve evaluation of a total package or evaluation of one small aspect of a total curriculum, such as a film. Although on-going programs are subject to evaluation, curriculum evaluation is usually associated with innovation, a new or different approach; the approach may be general or specific to a given area.Curriculum evaluation usually involves both internal and external criteria and comparisons. Internal evaluation is concerned with whether the new process or product achieves its stated objectives, that is whether it does what it purports to do, as well as with evaluation of the objectives themselves. External evaluation is concerned with whether the process or the product does whatever it does better than some other process or product.

In addition to student achievement, there are a number of other factors which should be considered during curriculum evaluation. The two most important factors are attitude and cost. Research has demonstrated that there is a proportional relationship between teacher attitude toward curriculum and its ultimate effectiveness.Curriculum evaluation has one major problem associated with it: It is very difficult to compare fairly the effectiveness of one program or approach with another. Even if two programs deal with the same subject area, they may deal with objectives which are very different, and it is very difficult to find a test or other measures which is equally fair, or valid for both programs. If one curriculum is to be compared to another, the objectives of each must be examined carefully; if no measure can be located which is equally appropriate to both, then one must be developed.

School evaluation

Evaluation of a school involves evaluation of the total educational program of the school and entails the collection of data on all aspects of its functioning. The purpose of the school evaluation is to determine the degree to which school objectives are being met and to identify areas of strength and weakness in the total program. Information from school program provides feedback to which gives direction to the future activities of the school and results in decision concerning the allocation of school resources.

One major component of school evaluation is the school testing program; the more comprehensive the testing program, the more valuable are the resulting data. A school testing program should include measurement of achievement, aptitude, personality and interest. Test selected for a school must match the objectives of the school and be appropriate for the students to be tested.

In general, school evaluation involves more than the administration of tests to students; it may require any combination of questionnaires, interviews, and observations with data being collected from all persons in the school community, including administration, teachers and counsellors.

Evaluation of large populations

Evaluation of large populations involves assessing the current status and the educational progress of large number of students, typically distributed over a large geographic region. State-wide assessment programs are generally based on the premises that the state system of education is responsible for student achievement of certain basic skills required for effective functioning in our society, and that programs designed to promote achievement of the basic skills should be as effective and economical as possible. One of the major aims of state assessment is to provide information to state and local decision makers about the adequacy of basic educational program. State-wide assessment typically involves measurement of minimum educational objectives and selected optional objectives using both criterion-references and norm-referenced

Evaluation of special projects and programs

Special projects and programs include all those organized efforts which are not, strictly speaking, part of the regular school program; they are typically innovative in nature and the duration of their existence is dependent upon their success. Whether it is required or not, conduction of evaluation is in the best interest of a project or program since it is the only valid way to very its effectiveness.

Evaluation of personnel

Evaluation of personnel (staff evaluation) includes evaluation of all persons responsible, either directly or indirectly, for educational outcomes, i.e., teachers, administrators, counsellors and so forth. It has been found out that this area of evaluation is very complicated; it is difficult to determine what behaviours are to be evaluated. The best solution to problem of personnel evaluation is to collect the best and most data possible, from as many sources as possible.

Phases of evaluation

Evaluation is a continuous process; contrary to public opinion, it is not what you do at the end. Evaluation should be planned for prior to execution of any effort and should be involved throughout the duration of the activity. There are typically a series of temporary ends in a continuous cycle. In student evaluation, for example, we start with a set of instructional objectives. Then we implement instructional strategies to facilitate their achievement. Then we measure achievement, a temporary end in the instructional cycle. Based on the results, we reassess our objectives and strategies and proceed. Thus the process is cyclic, with feedback from one cycle guiding the next. We do not just evaluate outcomes; every stage of the process is subject to evaluation beginning with the objectives.

The evaluation process entails decision-making. Any educational endeavour involves a whole host of decisions which must be made – decisions about objectives, decisions about strategies, decisions about measurement, and so forth. These various decisions can be classified in terms of when they are made, that is, during what stage of the activity under study. Thus, each phase of evaluation involves different kind of decisions. Logically we can identify three phases; the planning phase, the process phase and the product phase.

The planning phase

This initial phase of evaluation takes place prior to actual implementation and involves making decisions about what course of action will be taken and toward what ends. The planning phase involves a number of processes which are discusses below;

Situation analysis

The first step is to analyse the situation as it presently exists in order to establish the parameters of the effort. This step includes activities such as the collection of background information and assessment of existing constraints. For a teacher this may involve examination of the commutative records of his or her students in order to get a frame of reference based on their abilities and histories. After the parameters have been established, more realistic goals and objectives can be formulated.

Specification of objectives

Goals are general statements of purpose, or desired outcomes and not as such directly measurable. Each goal must be translated into one or more specific objectives which are measurable. Thus, objectives are specific statements of what is to be accomplished and how well and are expressed in terms of quantifiable, measurable outcomes. Objectives may process oriented or product oriented. Process objectives describe outcomes desired during the execution of the effort, and they related to the development and execution. Product objectives, on the other hand, describe outcomes intended as a result of the effort.

Objectives give direction to all subsequent activities and achievement of objectives is ultimately measured. Objectives, whether instructional or program objectives, form the foundation of all subsequent evaluation activities, and therefore it is critical that they themselves be evaluated in terms of relevance, measurability, substance, and technical accuracy.

Specification of pre-requisites

Objectives entail unique procedure with respect to student evaluation. In most cases, specification of a given set of instructional objectives is based on the assumption that students have already acquired certain skills and knowledge. If the assumption is incorrect, then the objectives are not appropriate. The assumed behaviours are referred to as pre-requisites or entry behaviours. Systematic instruction and evaluation require that these pre-requisites be specified and measured. Assessment of entry behaviour is specifically important at the beginning of any instructional unit. To arrive at pre-requisites, you simply ask yourself the following question: What must any students know or be able to do prior to instruction order to benefit from instruction and achieve any objectives.

Selection and development of measuring instruments

Collection of data to determine degree of achievement of objectives requires administration of one or more instruments. Such instruments must either be developed or selected. Selection of an instrument involves examining those that are available and selecting the best one. Best, in this case, means the one that is most appropriate for your objectives and users. Development of a good instrument takes considerable time, effort and skill. Training in measurement in the process is necessary for such end.

Delineation of strategies

Strategies are generally approaches to promoting achievement of one or more objectives. There may be instructional strategies, curriculum strategies, and pro- gram strategies. Each strategy entails a number of specific activities, and there are typically a number of strategies to choose from. Execution of these strategies must be planned for, to ensure the availability of necessary resources. Strategies which must be thoroughly thought of before evaluation is conducted include: task analysis, review of concepts, sequencing, provision of feedback and practice.

Preparation of time schedule

Preparation of realistic time schedule is important for all types of evaluation; rarely do we have as long as we please to conduct evaluation.

Basically a time schedule includes a list of the major activities of evaluation effort and corresponding expected initiation and completion times for each activity. You should allow yourself enough time, so that if an unforeseen minor delay occurs, you can still meet your final deadline.

The process phase

The process phase involves making decisions based upon events which occur during actual implementation of the planned instruction, program or project. The first step in the process phase is to administer pre-tests, if such are appropriate. Based on the pre-test results, decisions may be made concerning the appropriateness of the already specified objectives. Following initial testing, planned strategies and activities are executed in the predetermined sequence. Data collected during this phase provide feedback concerning whether execution is taking place as planned and whether and whether strategies and activities are being effective. The basic purposes of this phase are to determine whether the effort is being executed as intended, to determine the degree of achievement of process objectives, and to identify ways in which improvements can be made. The process phase is referred to as formative evaluation.

The product phase

The product phase involves making decisions at the end or more likely at the end of one cycle of instruction, a program or a project. Decisions made during the product phase are based on the results of the post-tests an on other cumulative types of data. The major purpose of the product phase is to collect data in order to make decisions concerning the overall effectiveness of instruction, a program or project. During this phase it is determined whether and/or what degree intended product objectives were achieved. Data analysis and interpretation is followed by the preparation of a report which describes the objectives, procedures, and outcomes of the effort. The results of the product phase of evaluation are used in at least in two major ways:

- They provide feedback and direction to all who are were involved in the effort
- They provide feedback to outside decision makers, such as parents, principals, school board members and funding sources
- Results of the product phase need to be interpreted with care. Failure to meet objectives, for example, is not necessarily fatal; degree of achievement needs to be considered. The product phase of evaluation is referred to as summative evaluation.

Formative Assessment

Activity 1

After reading the above note on the different types of evaluation, complete the chart below.

Type of evaluation	Major variables measured	Major types of decisions	What is this type of evaluation important to education
Student			
Curriculum			
School			
Large population			
Special programs			
Personnel			

Activity 2

Identify five problems associated with at-the-end evaluation.

1.	
2.	
3.	
4.	
5.	

Activity 3

Complete the chart below.

Phase of evaluation	Examples of decisions	Specific activities
Planning		
Process		
Product		

Evaluation - Case Study

Choose one of the schools in your area and interview the headmistress or anybody charged with examinations of the school about the way the conduct their exams. The following questions will serves as a guide.

What does evaluation mean to this school? Why is it conducted?

Which types of evaluation are conducted in this school? Why is it conducted?

How does the school plan and implement student evaluation? What steps are taken in each phase? How are the results used? Who has access to the results of the evaluation and how are they used?

Evaluation of Education Projects

Evaluation of education projects constitutes an integral part of the process of reviewing and evaluating the educational systems.

Study Compulsory Reading #1 (Designing Evaluation for Education Projects) and write an essay on evaluation of education projects.

- What is project evaluation?
- Why is project evaluation important in education?
- What are the steps involved in project evaluation?

- What are the different types of project evaluation?

Learning Activity #2

Specification of Objectives

SUMMARY

Objectives are the heart of all evaluation and assessment activities in education; they give direction to the process of education. Objectives help us what to measure and how to measure. They indicate whether the process of education has been successful, and what can be done to may it so. In this activity, learners will be able learn what objectives are and their importance in education. In addition, learners will master the types and components of educational objectives. Bloom's taxonomy of educational objectives will be discussed, so that they serve as guide to check whether the objective we set and measure are comprehensive enough to include all aspects of education.

Learning Activity

Video on learning objectives

https://www.youtube.com/watch?v=_woMKwBxhwU

This video explains the criterion for developing learning objectives. Listen to it and make notes.

List of Key Concepts

Objectives are specific statements of what is to be accomplished and how well, and are expressed in terms of quantifiable, measurable outcomes.

Needs assessment is the process of identifying needs and establishing relative priority of those needs.

Competency-based objectives are intended to communicate behaviours believed to be required for successful job performance and to have students demonstrate these behaviours prior to graduation.

Mastery objectives are objectives which should be achieved by virtually all students, regardless of ability or background

Taxonomy of educational objectives is intended to classify all objectives into a hierarchy of categories based on presumed complexity

Affective Domain deals with outcomes that are much more difficult to promote and to

measure, intangibles such as feelings, attitudes, interests and values.

Psychomotor Domain entails physical abilities, those involving muscular or motor skills, manipulation of objectives, or neuromuscular coordination. Goals: Goals are global statements of long-term outcomes.

Intermediate Objectives are more precise than general objectives but not as detailed as specific objectives.

Specific Objectives deal with the most discrete intended outcomes.

Content

The Nature of Objectives

Objectives are specific statements of what is to be accomplished and how well, and are expressed in terms of quantifiable, measurable outcomes. Objectives are executed prior to execution of an effort and are subject to technical review. An instructional objective is an intent communicated by a statement describing a proposed change in a learner – a statement of what the learner is to be like when he or she has successfully completed a learning experience. Learning is an inferred event, not directly observable one; thus it must be determined what behaviour, or performance, will constitute sufficient evidence that the desired ability exists. Whether a capability is acquired, or learned, as a result of instruction can only be determined by observing performance at two different points in time; before and after instruction. Learning per se cannot be directly observed or measured, changes in performance can be.

Objectives give direction or guide activities of an effort. The set general strategy and activities for their attainment. This means that different set of objectives will usually generate different strategies and thus activities. Some people argue that objectives produce rigid, conforming procedures, standards and behaviours which discourage creativity and spontaneity and fail to take into consideration individual differences. They also question whether all educational outcomes can be objectified. They support their argument that we tend to select objectives that are easier to measure, not necessary important. However, it should be noted that most of the desired outcomes can be expressed in observable, measurable terms.

Needs Assessment

It would be virtually impossible for the educational system to attempt to foster achievement of all possible objectives; formally or informally, each educational endeavour must select and put in priority order the goals and objectives with which it will be concerned. Conceptually if there is a gap between the way things are (current status or status quo) and the way we would like them to be (desired outcome), there is a need. Depending upon how specifically a need has been defined, it can be translated into either a goal or an objective, the intent being to eliminate the need.

Needs assessment is the process of identifying needs and establishing relative priority of

those needs. The basic goals and objectives of the educational system have remained relatively stable but the emphasis or priorities have shifted over time. Needs assessment can be applied to, and used by all levels and components of the educational system, including classrooms, programs, projects, schools, school systems, states and countries.

Classification and Statement of Objectives

Types of objectives

Objectives look basically look the same regardless of the type of evaluation, all objectives indicated intended outcomes. One basic difference is that certain types of evaluation, and their corresponding objectives, are primarily concerned with the performance of individuals, whereas others are primarily concerned with the performance of groups.Literally, thousands and thousands of objectives, covering virtually every area of instruction, have been written to express desired student behaviours. In contrast, very little exists in terms of performance objectives for teachers, even less for administrators, and essentially nothing for special program and project personnel; this is due mainly to the lack of empirical evidence linking given behaviours with job success.

Competency-based objectives

The objectives or competencies to be achieved by students in various programs are typically developed by university or college instructors, experts in their respective fields, and reviewed by appropriate public school or community groups. The in- tent of the competency-based objectives is to communicate behaviours believed to be required for successful job performance and to have students demonstrate these behaviours prior to graduation.

Astery Objectives

Certain instructional objectives are considered to be required, minimum essentials, i.e., objectives which should be achieved by virtually all students, regardless of ability or background; these are referred to as mastery objectives, or minimum skills, objectives. In mastery system, the amount learned, as indicated in the objective, is constant; instruction is individualized and the amount time required to achieve these objectives varies from one student to another.

Taxonomy of Educational Objectives

The taxonomies classify all objectives into a hierarchy of categories based on presumed complexity. Each succeeding category involves behaviours believed to be more complex than the one previous and each considered be pre-requisite to the next.

The stated purpose of the taxonomies is to facilitate communication. The existence of

taxonomies also focuses attention on a wide range of behaviour outcomes, thus making it less likely that all the objectives for a given effort will involve lower level behaviours only. It is believed that all instructional objectives can be classified as belonging to one of the three taxonomies or domains – cognitive, affective and psychomotor.

The Cognitive Domain

The taxonomy of cognitive objectives has definitely made educators aware of the wide range of abilities involved in cognitive learning. Each of the six major categories represents a different kind of learning process. The major categories of the cognitive taxonomy of educational objectives are as follows:

- Knowledge
- Comprehension
- Application
- Analysis
- Synthesis
- Evaluation

Robert Gagne has approached classification of the types of learning from a different angle. He believes that types or classes of learning can be categorized based on the conditions of instruction necessary to facilitate learning of each type. Gagne has distinguished eight classes of learning and corresponding instructional conditions for learning that are associated with them: a) signal learning b) stimulus-response learning c) chaining d) verbal association e) discrimination learning f) concept learning and g) problem solving

In the past, educators have been criticized for paying too much attention to cognitive outcomes and not being concerned enough with the feelings and attitudes of students. The affective domain deals with such outcomes.

THE EFFECTIVE DOMAIN

The taxonomy of affective objectives deals with outcomes that are much more difficult to promote and to measure, intangibles such as feelings, attitudes, interests and values. The affective category represents a hierarchy of acceptance which ranges from willingness to receive, or attend, to characterization by a value. As with all objectives, the intent is to identify observable, measurable behaviours from which we can infer learning; this is clearly more difficult for affective outcomes. Achievement of affective objectives can be determined through the administration of self-report measures which entail a number of problems.

The major categories of the affective taxonomy of educational objectives are as follows;

- Receiving
- Responding
- Valuing
- Organization
- Characterization by value

The psychomotor domain

The psychomotor domain entails physical abilities, those involving muscular or motor skills, manipulation of objectives, or neuromuscular coordination. The task for taxonomy development in the psychomotor domain is more complex than for other domains. Though there aren't widely accepted categories for this domain, the following is used in some of the literature on the area;

Reflex

Fundamental movements

Perceptual abilities

Physical abilities

Skilled movements

Non-discursive communication

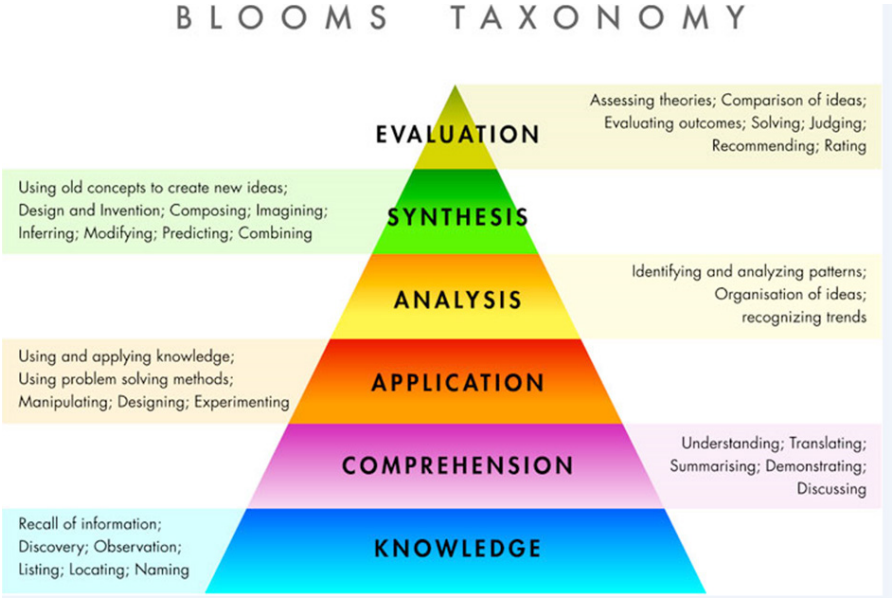


Figure 1: Blooms Taxonomy

Levels of Specificity

Objectives can be written at varying levels of specificity. At one end of the continuum we have very broad statements of long-term outcomes; these are generally referred to as goals or general objectives. At the other end of the spectrum we have very precise statements of more immediate outcomes, each representing one behaviour or outcome; these are generally referred to as specific objectives.

Types of Educational Objectives

General Objectives

Intermediate Objectives

Specific Objectives

Goals/General Objectives

Goals are global statements of long-term outcomes. The achievement of such goals or generally objectives cannot be directly measured directly. Goals of education tend to be established at the national level.

Intermediate Objectives

Goals are frequently translated into intermediate objectives. Objectives at this level are more precise than general objectives but not as detailed as specific objectives. They are in various situations referred to as cumulative objectives, course objectives and tasks. Such objectives are usually measureable and typically represent more complex behaviours expected as a result of achievement of a number of pre-requisite specific objectives.

Specific Objectives

Specific objectives deal with the most discrete intended outcomes. They are measurable outcomes upon which day-to-day activities such as instruction are based. The task analysis procedure is a method of translating cumulative objectives into specific objectives.

The table below provides summary of the terminology associated with the levels of specificity of objectives.

Level	Definition	Related terms

General objectives	Broad statements of long-term outcomes	Goals
Intermediate objectives	More precise statements of more short term outcomes	Second level objectives Module objectives Course objectives Terminal objectives Unit objectives
Specific objectives	Precise statements of immediate outcomes	Instructional objectives Behavioural objectives

An objective can be so vague that it is ambiguous or so detailed that it is trivial and impractical. A sensible guiding principle suggests that each objective should represent a distinct task. A task can be defined as the smallest component of performance which has a distinct and independent purpose. The more simply worded objective is clearly more functional.

Stating Objectives

Selection or development of an objective is really the most important part of any educational program. The objective gives direction to all other components of the program, including the content, methodology to be used for instruction and assessment. First and foremost, we need to evaluate the specificity of the objective. An objective should communicate the intended outcome well enough that any evaluator will assess it in essentially the same way and be able to tell whether it had been achieved.

Components of Objectives

Behaviour

The behaviour component of an objective indicates what will happen, or as in the case of instructional objectives, what observable measurable performance will constitute satisfactory evidence that the desired learning has occurred. An objective should state the intended learning outcome using an action verb that represents an observable outcome. Each objective should entail an achievement of only one and behaviour, otherwise measurement of achievement becomes confounded.

Condition

The conditions of an objective are the 'givens,' the context within which the behaviour will be exhibited. Conditions usually specify necessary materials, allowable resources, or imposed restrictions, and the word 'given' is not necessarily used.

Criterion

When stated, the criterion component of an objective is the part that tells us how well the behaviour must be performed. It is the standard against which actual performance is compared. If the criterion is obvious or 100%, it is not generally stated.

Formative Assessment

1. Study Compulsory Readings #2 and #3. For each of the following behaviours, check C if it is a cognitive behaviour and if it is an affective behaviour. Also, check H if it represents higher-order outcomes and L if it represents a lower-order outcome.

No.	Behaviour	Type		Level	
		C	A	H	L
1.	Recites article one of the constitution				
2.	Balances chemical equations				
3.	Pays attentions to film on genetics				
4.	Interprets a graph				
5.	Buys only Halal foods				
6.	Defines the term evaluation using his/her own words				
7.	Listens to a classical radio station on the way to and from work every day				
8.	Participates in a spelling bee				
9.	Prepares an income tax return				
10.	Knows the meaning of road signs				

2. For each of the following statements, check G if it is a goal, I if it is an intermediate objective or S if it is a specific objective.

No.	Statement	G	I	S
1.	To become a good citizen			
2.	To type a business letter in correct form			
3.	To states three food sources of iron			
4.	To makes three baskets out of ten attempts from the free throw line			
5.	To have command of basic skills			
6.	To write grammatically correct sentences			
7.	To become economically self-sufficient			
8.	To perform basic operations on fractions			
9.	To be able to give a list of 20th century novels, states the author of each with 80% accuracy			
10.	Voluntarily checks out two books from the library during the term			

3. Study the content of readings #4 and #5, and then do the following exercise.

1. Select any two of the following topics taught in secondary school syllabus.

a. The periodicity of chemicals

b. The reproductive system

c. Gravitational force

d. Rocks

e. Earthquakes

f. World Wars

2. For each of the topics you have selected, write at least 2 general objectives and between 4 specific objectives to be attained by the learners at the end of the evaluation.

3. Indicate the level of secondary education for which these objectives are appropriate.

4. For each of the objectives you have developed, classify them on the basis of the domain – cognitive, affective or psychomotor – to which they are related to.

5. Again, for each of the objectives you have developed, indicate its components such as (behaviour, condition and criterion).

Learning Activity #3

CLASSIFICATION AND SELECTION OF TESTS

SUMMARY

This learning activity is intended to enable students master some concepts in the area of educational evaluation and testing. The activity starts with some basic concepts related to measurement and testing. Then, the modes of data collection that are used in evaluation are briefly discussed. In detail, the classification schemes used to categorize different testing. The most important characteristics of a good test are later outlined in this section of the module. Lastly, you are given some highlight on the process of selecting a test (usually standardized) from a number of options.

List of Key Concepts

An essay test is one in which the number of questions is limited and res- ponders must compose answers,

An objective test is one for which subjectivity in scoring is eliminated, at least theoretically.

Standardized test is one that is developed by subject matter and measurement specialists,

Locally-developed test is usually developed locally by teachers for a specific purpose.

Self-report data consist of oral or written responses from individuals.

Rating scale is an instrument with a number of items related to a given variable, each item representing a continuum of categories between two extremes

Achievement tests are tests that measure the correct status of individuals with respect to proficiency in given areas of knowledge or skills.

Character or Personality tests are designed to measure characteristics of individuals along a number of dimensions

Aptitude tests are measures of potential that are used to predict how well someone is likely to perform in a future situation.

Performance standards are the criteria to which the results of measurement are compared in order to interpret them.

Norm-referenced test is any test, standardized or locally-developed, which reports and interprets each score in terms of its relative position with respect to other scores on the same test,

Criterion-referenced test is any test which reports and interprets each score in terms of an absolute standard is criterion-referenced.

Test objectivity means that an individual's score is the same, or essentially the same, regardless of who is doing the scoring.

Discrimination is the ability of a test to detect or measure small differences in achievement or attainment.

Comprehensiveness indicates the extent to which a test samples major lesson objectives that are to be measured

Validity is the extent to which a test measures what it is intended to measure.

Content validity refers to the extent to which items on the test represent the entire range of possible items the test should cover.

Criterion-related validity refers to the extent to which a test is demonstrated to be effective in predicting criterion or indicators of a construct.

Concurrent validity occurs when the criterion measures are obtained at the same time as the test scores.

Predictive validity occurs when the criterion measures are obtained at a time after the test

Construct validity refers the extent to which a test demonstrates an association between the test scores and the prediction of a theoretical trait.

Reliability refers to the consistency of a measure. A test is considered reliable if we get the same result repeatedly.

Content

Measurement and Testing

As discussed earlier, measurement is an essential component of the evaluation process. It is a critical part since resulting decisions are only as good as the data upon which the data are based.

In general sense, data collection is involved in all phases of evaluation – the planning phase, the process phase and the product phase. Measurement, however, the process of quantifying the degree to which someone or something possesses a given trait, normally occurs in the process and the product phase.

Testing is necessary at certain points and useful at others. Testing can be conducted at the end of an instruction cycle – semester, term or unit.

Such post testing is for the purpose of determining the degree to which objectives (formal or informal) have been achieved, be they instructional objectives or program objectives. Frequently, pre-test or baseline data are collected at the beginning of the cycle. Pre-tests serve several purposes, the most important being that knowledge of the current status of a group may provide guidance for future activities as well as a basis of comparison for post-test results. There are a variety of situations where testing is useful. A teacher may administer

tests of entry behaviour to determine whether assumed pre-requisites have indeed been achieved. A special project designed to reduce dropouts may administer attitude tests and tests of personality variables such as introversion, aggression and anxiety in an effort to identify potential dropouts or to better understand students having difficulties. A school may administer tests of scholastic aptitude in order to determine realistic achievement goals for students and to assist in the guidance process.

Data Collection

There are three major ways to collect data:

Administer a standardized instrument

Administer a locally developed instrument

iii) Record naturally available data (such as grade point averages and absenteeism)

Depending upon the situation, one of these ways may be most appropriate or a combination may be required. Collection of available data, requiring minimum effort, sounds very attractive. There are not very many situations, however, for which this type of data is appropriate. Even when it is appropriate – that is, will facilitate intended decision making – there are problems inherent in this type of data. For example, the same letter grade does not necessarily represent the same level of achievement, even in two different classes in the same school or two different schools in the same system. Further, the records, for which the data is taken may be incomplete and disorganized. Developing an instrument for a particular purpose also has several major drawbacks. The development of a 'good' instrument requires considerable time, effort and skill. Training at least equivalent to a course in testing and evaluation is necessary in order to acquire the skills for good instrument development.

In contrast, the time it takes to select an appropriate instrument (usually from among standardized, commercially available instruments) is inevitably less than the time it takes to develop an instrument which measures the same thing.

Further standardized instruments are typically developed by experts who possess the necessary skills. Thousands of standardized instruments are available which yield a wide variety of data for a wide variety of purposes. Major areas for which numerous measuring instruments have developed include achievement, personality and aptitude. Each of these can be in turn further divided into may subcategories. In general, it is usually a good idea to find out whether a suitable instrument is already available before jumping into instrument development. There are situations, however, for which use of available instruments is impractical or inappropriate.

A teacher-made test, for example, is more appropriate or valid for assessing the degree to which students have achieved the objectives of a given unit.

Classification Schemes

At this point it must be emphasized that a test is not necessarily a written set of questions to which an individual responds in order to determine whether he/she passes. A more inclusive definition of a test is a means of measuring the knowledge, skills, feelings, intelligence or aptitude of an individual or a group. Tests produce numerical scores which can be used to identify, classify or otherwise evaluate test takers. While in practice most of tests are in paper-and-pencil form, there are many different kinds of tests and many different ways to classify them. The various classification schemes overlap considerably, and categories are by no means mutually exclusive. Any test can be classified on more than one dimension.

Response Behaviours

The term response behaviours refers to the way in which behaviours to be measured are exhibited. While in some cases responses to questions or other stimuli are given orally, usually, they are either written or take the form of an actual performance.

Written responses

Written tests can be classified as either essays (subjective) or objective and standardized or locally-developed.

Essay Vs. Objective Tests

Essays

An essay test is one in which the number of questions is limited and responders must compose answers, typically lengthy, e.g., 'identify and discuss the major reforms in African education during the colonial period.' Determining the goodness or correctness of answers to such questions involves some degree of subjectivity.

Objective tests

An objective test is one for which subjectivity in scoring is eliminated, at least theoretically. In other words, anyone scoring a given test should come up with the same score. Examples of objective tests are multiple choice tests, true-false tests, matching tests, and short answer tests. Standardized Vs. Locally Developed Tests

Standardized Tests

A standardized test is one that is developed by subject matter and measurement specialists that is field-tested under uniform administration procedures, that is revised to meet certain criteria and scored and interpreted using uniform procedures and standards. Standardization

permeates all aspects of the test to the degree that it can be administered and scored exactly the same way every time it is given. Although other measurement instruments can be standardized, most standardized tests are objective, written tests requiring written responses. Although exceptions occur, the vast majority of standardized tests have been administered to groups referred to as the norm group. The performance of a norm group for a given test serves as the basis of comparison and interpretation for other groups to whom the test is administered. The score of a norm group are called norms. Ideally, the norm group is a large, well defined group which is representative of the group and subgroups for whom the test is intended.

Locally-developed test

The opposite of standardized test is obviously non-standardized test. Such tests are usually developed locally for a specific purpose. The tests used by teachers in the classroom are examples of locally-developed tests. Such tests do not have the characteristics of standardized tests. A locally-developed test may be as good as a standardized test, but not that often. Use of locally-developed test is usually more practical and more appropriate. A locally-developed test would more likely reflect what was actually taught in the classroom to a greater degree than standardized tests.

Performance Tests

For many objectives and areas of learning, use of written tests is an inappropriate way of measuring behaviour. You cannot determine how well a student can type a letter, for example, with a multiple choice test or open question. Performance is one of these areas. Performance can take the form of a procedure or a product. A procedure is a series of steps, usually in a definite order, executed in performing an act or a task. Examples include, adjusting a microscope, passing a football, setting margins on type writer, drawing geometric figures or calculating sum of figures in Excel. A product is a tangible outcome or result. Examples of a product include typed letter, a painting, a poem, and a science project. In either case the performance is observed and rated in some way. Thus, performance is one which requires the execution of an act or the development of a product in order to determine whether or to what degree a given ability or trait exists.

Data Collection Methods

There are many different ways to collect data and classifying data collection methods is not easy. However, a logical way to categorize them initially is in terms of whether the data are obtained through self-report or observation.

Self-Report

Self-report data consist of oral or written responses from individuals. An obvious type of self-report data is that resulting from the administration of standardized or locally-developed

written tests, including certain achievement, personality and aptitude tests. Another type of self-re-

port measure used in certain evaluation efforts is the questionnaire, an instrument with which you are probably familiar. Also, interviews are sometimes used.

Observation

When observation is used, data are collected not by asking but by observing. A person being observed usually does not write anything;

he or she does something and that behaviour is observed and recorded. For certain evaluation questions, observation is clearly the most appropriate approach. To use an example, you could ask students about their sportsmanship and you could ask teachers how they handle behaviour problems, but more objective information would probably be obtained by actually observing students at sporting events and teacher in their classrooms. Two types of observation which are used in evaluation efforts are natural observation and observation of simulation. Certain kinds of behaviours can only be observed as they occur naturally. In such situations, the observer does not control or manipulate anything, and in fact works very hard at not affecting the observed situation in any way. As an example, classroom behaviour can best be addressed through observation. In simulation observation, the evaluator creates the situation to be observed and tells participants what activities they are to engage in. This technique allows the evaluator to observe behaviour which occurs infrequently in natural situations or not at all.

Rating scale

Is an instrument with a number of items related to a given variable, each item representing a continuum of categories between two extremes.

Persons responding to items place a mark to indicate their position on each item.

Rating scales can be used as self-report or as an observation instrument, depending on the purpose for which they are used.

Behaviours Measured

Virtually all possible behaviours that can be measured fall into one of three major categories: achievement, character and personality, and aptitude. All of these can be standardized or locally-developed. These categories also apply equally well to the three domains of educational outcome, namely cognitive, affective an psychomotor.

Achievement

Achievement tests measure the correct status of individuals with respect to proficiency in given areas of knowledge or skills. Achievement tests are appropriate for many types of

evaluation besides individual student evaluation. Achievement test can be standardized (which are designed to cover content which are common to many classes of the same type) or locally-developed (designed to measure particular set of learning outcomes, set by specific teacher). Standardized tests are, in turn, available for individual curriculum areas or in the form of batteries, which measure achievement in several different areas.

A diagnostic test is a type of achievement test which yields multiple scores for each area of achievement; these scores facilitate identification of specific areas or deficiency or learning difficulty. Items in a diagnostic test are intended to identify skills and knowledge that students must have achieved before they proceed to another level. Ideally, diagnosis should be an on-going process and the teacher must design them in a way that such tests help him/her find out the problems that students encounter as they proceed in the learning process.

Character and Personality

Tests of character and personality are designed to measure characteristics of individuals along a number of dimensions and to assess feelings and attitudes toward self, others, and a variety of activities, institutions and situations. Most of the tests of character and personality are self-report measures and ask an individual to respond to a series of questions or statements. There are instruments in this category which are designed to measure personality, attitudes, creativity, and interest of students.

Aptitude

Aptitude tests are measures of potential. They are used to predict how well someone is likely to perform in a future situation. Tests of general aptitude are variously referred to as scholastic aptitude tests, intelligence tests, and tests of general mental ability. Aptitude tests are also available to predict a person's likely level of performance after following some specific future instruction or training.

Aptitude tests are available in the form of individual test on specific subject or content or in the form of batteries. While virtually all aptitude tests are standardized and administered as part of school testing program, the results are useful to teachers, counsellors and administrators. Readiness aptitude tests (or prognostic tests) are administered prior to instruction or training in a specific area in order to determine whether and to what degree a student is ready for, or will profit from, an instruction. Readiness tests, which are part of aptitude tests, typically include measurement of variables such as auditory discrimination, visual discrimination and motor ability.

Performance Standards

Performance standards are the criteria to which the results of measurement are compared in order to interpret them. A test score in and of itself means nothing. If I tell you that Ahmed

got 18 correct, what does that tell you about Ahmed's performance? Absolutely nothing. Now if I tell that the average score for the test was 15, at least you know that he did better than the average. If instead I tell you that a score of 17 was required for a master classification, you don't know any- thing about the performance of the rest of the class, but you know that Ahmed attained mastery. These are the two ways with which we can interpret the results of a test, first by comparing it to other students in the class (that is Norm-Referenced Measurement) or by comparing it to a pre-determined criteria (that is Criterion-Referenced Measurement).

Norm-referenced standards

Any test, standardized or locally-developed, which reports and interprets each score in terms of its relative position with respect to other scores on the same test, is norm-referenced. If your total IQ score is 100, for example, the interpretation is that your measured intelligence is average, average compared to scores of a norm group. The raw scores resulting from administration of standardized test are converted to some other index which indicates relative position. One such equivalent technique familiar to you is the percentile. A given percentile indicates the percentage of the scores that were lower than the percentile's corresponding score. For example, Mr. Ahmed might have scored on the 42nd percentile in a standardized math test, which means 42% of the students who took that test scored below Ahmed. In such way of communicating student scores, there is no indication of what Mr. Ahmed knows or does not know. The only interpretation is in terms of Ahmed's achievement compared to the achievement of others.

Norm-referenced tests are based on the assumption that measured traits involve normal curve properties. The idea of the normal curve is that measured traits exist in different amounts in different people. Some people have a lot of it, some people have little of it, and most have some amount called the 'average' amount. For example, if you administer a math test to a class of 100 students, given that the test of appropriate level – that is not too easy or too difficult – a small portion of the class will perform high and another equal portion will perform low, while the majority will perform around the average score.

The normal curve is figured below.

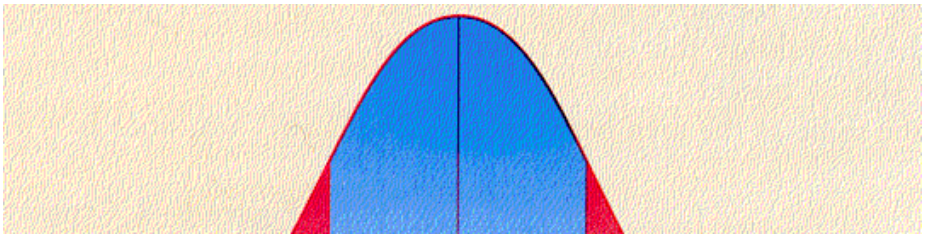


Figure 3 : Normal curve

Most of the scores are (the highest frequency) are average, fewer scores are above or below average, and very few are way above or way below the average. The average group contains approximately 65% of the scores. Norm-referenced measurement is used in this hypothetical distribution of scores is applied to real classroom or assessment situations. That is, when the top 2-3% are given As (whatever their scores be), the next 15% are given Bs, the next 65% receive Cs, the next 15% are given D, and the bottom 2-3% receive F grades. The letter grades are assigned irrespective of the actual performance of the student.

There is one problem with norm-referenced standards. Some people can get good grades (or conversely bad grades) without necessarily exhibiting the corresponding performance level. Such method may encourage the students, but it may cause a special problem when the class is more homogeneous.

Criterion-referenced standards

Any test which reports and interprets each score in terms of an absolute standard is criterion-referenced. In other words, interpretation of one person's score has nothing to do with anybody's score. A score is compared with a standard or performance, not with scores of other people. When criterion-referenced tests are used, everyone taking the test may do well or everyone may do poorly. In this context, criterion can be defined as a domain of behaviours measuring an objective.

Characteristics of a good test

There are a number of characteristics which are desirable for all tests. Standardized tests, which are developed by experts in both subject matter and evaluation, are more likely to live up to these standards. These characteristics, which are discussed in a more detailed manner, include;

- Test objectivity
- Discrimination
- Comprehensiveness
- Validity
- Reliability
- Specification of conditions of administering
- Direction of scoring and interpretation

Test Objectivity

Test objectivity means that an individual's score is the same, or essentially the same, regardless of who is doing the scoring. A test is objective when instructor opinion, bias, or individual judgment is not a major factor in scoring. Tests may be scored by more than one person, at different or the same time. In education, we wonder to what extent is the scoring of these individual scorers is the same. If the possible difference between the people scoring the same test high, that test is low in objectivity. Though individual's may naturally be different in the way they perceive information, we assume that the more objective a test is, the more aspires to the high quality evaluation we all envision in education. This does not mean that tests which do not have high degree of objectivity (such as subjective tests) are not of quality. Even subjective tests, though they are designed to measure information that can be looked at from different angles, certain level of objectivity is necessary. The individual designing must have something in mind that constitutes good performance on such test. That thing can be multiple, but criteria must be developed to make sure any scoring is fair enough to contribute to discriminating of students on such basis. Objectivity is a relative term.

Discrimination

The test should be constructed in such a manner that it will detect or measure small differences in achievement or attainment. This is essential if the test is to be used for ranking students on the basis of individual achievement or for assigning grades. It is not an important consideration if the test is used to measure the level of the entire class or as an instructional quiz where the primary purpose is instruction rather than measurement. As is true with validity, reliability, and objectivity, the discriminating power of a test is increased by concentrating on and improving each individual test item. After the test has been administered, an item analysis can be made that will show the relative difficulty of each item and the extent to which each discriminates between good and poor students.

Often, as in obtaining reliability, it is necessary to increase the length of the test to get clear-cut discrimination. A discriminating test:

Produces a wide range of scores when administered to the students who have significantly different achievements. Will include items at all levels of difficulty. Some items will be answered correctly only by the best students; others will be relatively easy and will be answered correctly by most students. If all students answer an item correctly, it lacks discrimination.

Comprehensiveness

For a test to be comprehensive, it should sample major lesson objectives. It is neither necessary nor practical to test every objective that is taught in a course, but a sufficient number of objectives should be included to provide a valid measure of student achievement in the complete course.

Validity

The most important characteristic of a good examination is validity; that is, the extent to which a test measures what it is intended to measure. It is vital for a test to be valid in order for the results to be accurately applied and interpreted.

Validity isn't determined by a single statistic, but by a body of research that demonstrates the relationship between the test and the behaviour it is intended to measure. There are three types of validity:

a) Content validity

When a test has content validity, the items on the test represent the en- tire range of possible items the test should cover. Individual test questions may be drawn from a large pool of items that cover a broad range of topics.

In some instances where a test measures a trait that is difficult to define, an expert judge may rate each item's relevance. Because each judge is basing their rating on opinion, two independent judges rate the test separately. Items that are rated as strongly relevant by both judges will be included in the final test.

b) Criterion-related Validity

A test is said to have criterion-related validity when the test is demonstrated to be effective in predicting criterion or indicators of a construct. There are two different types of criterion validity: Concurrent Validity occurs when the criterion measures are obtained at the same time as the test scores. This indicates the extent to which the test scores accurately estimate an individual's current state with regards to the criterion.

For example, on a test that measures levels of depression, the test would be said to have concurrent validity if it measured the current levels of depression experienced by the test taker.

Predictive Validity occurs when the criterion measures are obtained at a time after the test. Examples of test with predictive validity are career or aptitude tests, which are helpful in determining who is likely to succeed or fail in certain subjects or occupations.

Construct Validity

A test has construct validity if it demonstrates an association between the test scores and the prediction of a theoretical trait. Intelligence tests are one example of measurement instruments that should have construct validity.

The instructor can ensure whether his/her test items are valid by following accepted test construction procedures that include:

Use of the lesson objectives as a basis for the test requirements. An examination so

constructed will tend to measure what has been taught.

Review of the test items and the completed examination by other instructors.

Selection of the most appropriate form of test and type of test item. Thus, if the instructor desires to measure "ability to do," he must select that form of the test that will require the student to demonstrate his "ability to do." If another less desirable form is used, it must be recognized that the validity of the measurement has been reduced.

Presentation of test requirements in a clear and unambiguous manner. If the test material cannot be interpreted accurately by the student, he or she will not realize what is being covered; hence, he or she will be unable to respond as anticipated. Such a test cannot be valid.

Elimination, so far as is possible, of those factors that are not related to the measurement of the teaching pints. A test that is not within the capabilities of the students as to time or educational level may fail to measure their actual learning in the course.

Reliability

Reliability refers to the consistency of a measure. A test is considered reliable if we get the same result repeatedly. For example, if a test is designed to measure a trait (such as introversion), then each time the test is administered to a subject, the results should be approximately the same. Unfortunately, it is impossible to calculate reliability exactly, but there several different ways to estimate reliability.

Test-Retest Reliability

To gauge test-retest reliability, the test is administered twice at two different points in time. This kind of reliability is used to assess the consistency of a test across time. This type of reliability assumes that there will be no change in the quality or construct being measured. Test-retest reliability is best used for things that are stable over time, such as intelligence. Generally, reliability will be higher when little time has passed between tests.

Inter-ratter Reliability

This type of reliability is assessed by having two or more independent judges score the test. The scores are then compared to determine the consistency of the ratters estimates. One way to test inter-ratter reliability is to have each ratter assign each test item a score. For example, each ratter might score items on a scale from 1 to

10. Next, you would calculate the correlation between the two ratings to determine the level of inter-ratter reliability. Another means of testing inter-ratter reliability is to have ratters determine which category each observations falls into and then calculate the percentage of agreement between the ratters. So, if the ratters agree

8 out of 10 times, the test has an 80% inter-ratter reliability rate.

c) Parallel-Forms Reliability

Parallel-forms reliability is gauged by comparing to different tests that were created using the same content. This is accomplished by creating a large pool of test items that measure the same quality and then randomly dividing the items into two separate tests. The two tests should then be administered to the same subjects at the same time.

Internal Consistency Reliability

This form of reliability is used to judge the consistency of results across items on the same test. Essentially, you are comparing test items that measure the same construct to determine the tests internal consistency. When you see a question that seems very similar to another test question, it may indicate that the two questions are being used to gauge reliability. Because the two questions are similar and designed to measure the same thing, the test taker should answer both questions the same, which would indicate that the test has internal consistency.

The following factors will influence the reliability of a test:

- Administration. It is essential that each student have the same time, equipment, instructions, assistance, and examination environment. Test directions should be strictly enforced.
- Scoring. Objectivity in scoring contributes to reliability. Every effort should be made to obtain uniformity of scoring standards and practices.

- Standards. The standards of performance that are established for one class should be consistent with those used in other classes. A change in grading policies not based upon facts, uniform standards, and experience factors gained from other classes will affect the reliability of test results.
- Instruction. The reliability of tests results will be affected if the instruction presented to a class tends to overemphasize the teaching points included in the examination. This is often known as "teaching the test" and is undesirable. When the instructor gives students obvious clues as to the test requirements, he not only affects the reliability of the test, but he insults the intelligence of his class.
- Length. The more responses required of students, the more reliable will be the test or measuring device.

Specification of conditions of administration

A good test must specify the conditions under which the test must be conducted. The conditions must give all students a fair chance to show what their performance. This will improve the reliability of the test. Standardized must come along with specification on the conditions in which students must perform. Administering the test with highly varied conditions will highly interfere with the results of the test. In general, when administering a test, the following must be kept in mind:

Physical conditions

- Light, ventilation, quiet, etc.
- Psychological conditions
- Avoid inducing test anxiety
- Try to reduce test anxiety
- Don't give test when other events will distract
- Suggestions
- Don't talk unnecessarily before the test
- Minimize interruptions
- Don't give hints to individuals who ask about items
- Discourage cheating
- Give students equal time to take the test.

Direction for scoring and interpreting test results

Good test must come with direction on how to score and interpret the results of a test.

This is especially important for standardized tests, which are used by individuals other than those who developed the test in the first place.

Such directions are also important for locally-developed tests, since other individuals may get involved in the process of scoring and interpreting the test, due to unforeseen circumstances.

The following guidelines contribute to the development of clear direction for tests.

- Provide clear descriptions of detailed procedures for administering tests in a standardized manner.
- Provide information to test takers or test users on test question formats and procedures for answering test questions, including information on the use of any needed materials and equipment.
- Establish and implement procedures to ensure the security of testing materials during all phases of test development, administration, scoring, and reporting.
- Provide procedures, materials and guidelines for scoring the tests, and for monitoring the accuracy of the scoring process. If scoring the test is the responsibility of the test developer, provide adequate training for scorers.
- Correct errors that affect the interpretation of the scores and communicate the corrected results promptly.
- Develop and implement procedures for ensuring the confidentiality of scores.

Selection of a Test

Usually, when we talk of selection of measuring instruments or tests, we are referring to standardized instruments. In most of the situations, you may have only one option you are compelled to choose. However, in some cases, you may be confronted with the challenge of comparing instruments and choosing the one that best suits your needs and circumstances. In such scenario, you must be your decision on objective facts and sound evaluation principles.

First and foremost, you have to think of the measurement technique that meets the purpose of the testing and the objectives to be measured. For example, the evaluation plan for an experimental reading curriculum would involve a combination of locally-developed and standardized instruments. It may also involve both norm-referenced and criterion-referenced formats. On the other hand, school testing program would require the sole use of standardized tests.

When you are required to select a standardized test from a number of options, you also need to consider the following factors completeness of the information concerning characteristics of the test. Surprisingly, there is a relationship between the amount of information supplied by the publisher and the quality of the test. The most important information required is the validity of the test. In addition, data concerning the reliability of the test should be available and meet the standards we set for our testing exercise. Other data which should be reported include information concerning test administration, scoring and interpretation.

Formative Assessment

Answer the following questions.

1. What types of evaluation data is collected in secondary schools in your area? What is the purpose of each of these data? What data do you think is missing and what will be the use of such data?

2. Which of the following types of tests is used in your area? At what level is each used and for what objectives?

   a. Standardized tests

   b. Locally-developed tests

   c. Performance tests

3. Study Compulsory Reading #6 and write a short essay (about 500 words) to defend standardized tests against the arguments posed in the article.

4. Which performance standard (criterion-referenced or norm-referenced) is used in your area? What are its advantages and disadvantages?

Learning Activity #4

DEVELOPMENT OF INSTRUMENTS

SUMMARY

This learning activity is designed to help you the principles underlying the development of test instruments. The unit starts with the process of restricting, defining and selecting the content that should be covered by a test. This process is simplified by the use of table specifications, which enable the test-developer, see the content to be measured, the level at which it should be measured and how it should be measured. After that, the guidelines concerning the construction of test items are discussed. This includes the process of coming up with effective objective and essay type items. Summarized recommendations for the design and format of tests are provided at the end of the learning activity. Lastly, it is emphasized that test instruments be revised thoroughly before being administered to tests.

List of Key Concepts

A table of specifications is essentially a blueprint for a test with the purpose to insure that all intended outcomes, and only the intended outcomes, are measured and that the test includes the appropriate number of items for each measured item.

Test validation is the process of administering and revising over and over again until acceptable levels of validity and reliability are achieved.

Content

Restriction, Definition and Selection of Content

The first step in designing a test is to put boundaries on what is to be covered in a test. Objectives usually provide the blueprint for the topics or clusters of topics to be included in a test, and a test should represent meaningful unit of instruction. A test will normally represent a range from several days to several weeks of instruction. Narrowing test coverage prevents cognitive overload on the part of students, and permits a large more representative sampling of the behaviours represented by the selected objectives. Of course certain cumulative tests such final exams require broader coverage and may represent weeks or even months of instruction. If individual tests have been carefully designed, the task of developing cumulative tests is facilitated since the outcomes measured by such tests represent a sampling of the performances measured on all previous tests.

The next thing in designing a test is to make a detailed content outline (or syllabuses) if such is not already available. The outline needs to be detailed enough to include knowledge and skill outcomes desired of students. While each and every outcome may not be measured on a test, they should all at least be candidates for inclusion. If specific objectives have been formulated, then desired performances associated with each entry in the outline will already

be specified. If not these behaviours now have to be identified and we must keep in mind that we are concerned with more than just the possession of knowledge. This process of content description is called 'domain description' and is intended to identify all possible items by delineating rules for generating items.

The third step in designing a test is to devise a scheme for sampling from the domain of behaviours. With the exception of certain mastery tests, a test invariably represents a sample of behaviours. It is just not possible to measure each and every aspect of each and every outcome. Further, if we carefully select the ones we really measure, we can generalize to the total domain with a reasonable degree of confidence. It is a requirement that test items constitute a representative sample of desired behaviours. This is what we have called content validity. The more structured and well-defined the content is, the easier this is to accomplish. Random selection, which gives each outcome an independent and equal chance of being included in the test, leads to the process of attaining content validity in a test. You can also select the outcomes to be selected for the test by stratifying the content on the basis of criteria inherent to the content (or some other logical criteria).

One systematic approach to test design is involves the development of a table of specifications. Once test content has been identified and defined, and all the behavioural outcomes specified, a logical next step is the construction of a table of specifications.

Table of Specifications

A table of specification is essentially a blueprint for a test. At the name implies, it specified the content of the test. Its basic purpose is to insure that all intended outcomes, and only the intended outcomes, are measured and that the test includes the appropriate number of items for each measured item. A table of specification is a two-way table with one axis being essentially a content outline and the other axis indicating the behaviours desired with respect to content.

Content outline

Along one axis of the table of specifications, the major headings and subheadings of the content outline are listed. The table of specifications may include all topics and subtopics, or it may include only those that have been specially specified. The more detailed this axis is, the more likely it is that the test will include items for all desired topics. Also, to greater extent, the complexity of the outline is a function of the complexity of the content area being measured.

Behavioural outcomes

Along the axis of the table the behaviours or performances for each topic are given. This approach focuses attention on higher order outcomes and insures that the test will not

contain predominantly or totally knowledge level items. The level of the items will mainly depend on the level of students and the objectives already set. However, it should be noted that teachers are more likely to neglect items that measure higher order skills, such application, synthesis and analysis. Some tables of specifications may go as specific indicating the sub-domain the outcome is related to. It should be noted that it is easier to develop table of specifications for criterion-referenced tests than for norm-referenced tests, since intended outcomes for the former have already been thought through and delineated.

Number of items

The intersection of content outline headings and behavioural outcome categories forms cells into which appropriate number (or percentage) of items is placed. The number of items allocated for each cell should be based on the number of objectives or their weight (or importance). This means, we have to think thoroughly about 'what it takes' to measure a particular behavioural outcome for a given topic.

Developing a table of specifications is a planning activity and ideally should be executed during the planning phase of evaluation. The logical time to do so is soon after developing specific objectives. If objectives are laid out for an entire term, a logical next step is to divide them into sets representing manageable units of instruction. As with objectives, if developed prior to instruction, a table of specifications serves as a guide for instruction since we know what outcomes are intended for what topics, and the desired emphasis to be placed on each.

Constructing test items

Overview

There are a number of guidelines for developing a number of different types of items. When students read an item, we want the only factor affecting whether they answer it correctly to be whether they possess the behaviour being measured, that is, whether they have achieved the objective. And yet, there are a number of characteristics of an item that can either prevent knowledgeable student from getting it correct or permit a unknowledgeable student to get it correct. Factors such as the way a question is worded, the nature of the alternatives, and the directions for responding can affect a students' ability to demonstrate a student achievement or lack of it.

There are basically two types of test items, essay and objective. Objective items include short answer, multiple choice, true false, and matching. Test items can also be classified as supply or selection items. Essay and short answer items are supply items because the answer must be supplied by the student. Multiple choice, true and false and matching items are referred to as selection items since the student selects an answer among provided alternatives. Which item type or types are appropriate for a given test depends mostly upon the nature of behaviours being measured and the nature of instruction. There are no item types that are specific for some specific kind of outcomes; any item type can be used to measure any behavioural outcome regardless of its taxonomy level.

General Guidelines

It is generally recommended against the development of a test in a hurry. The teacher, or the person responsible the evaluation, should get enough time to plan, develop and revise items. Certain tests, such as posts, must be developed prior to instruction. Unit tests, on the other hand, can be developed following instruction. It is advisable to develop tests ahead of time and revise as necessary, or you can develop the test as you go along. This approach has a number of advantages, the most important being the fact that it spreads the difficulty task of test construction over a period of time.

Items should directly correspond to with the intended behavioural outcome or objective. In other words, the behaviour elicited by the item should be precisely the behaviour specified in the objective. If an objective states that the student should be able handle microscope, the corresponding item should not ask him/her to list the parts of a microscope or its uses; rather, it should require students to handle it. This is called objective – item correspondence.

Closely related to the concept of objective-item correspondence is the item sufficiency. Item sufficiency means that a correct response to an item (or items) constitutes sufficient evidence that the student has demonstrated the intended behaviour, or has achieved the objective. Depending upon the nature of the objective, it might not be possible to develop such an item; several items may well be required to meet the sufficiency guideline. It is, however, a goal to work toward.

One major requirement of an item is that it should clearly communicate to the test takers the meaning intended by the test developer. In other words, the item should be as free from ambiguity as possible. The way the student interpreted the item should be the way the test developer intended it to be interpreted. The item should also clearly communicate to any competent scorer.

Item difficulty is an item characteristic that can be manipulated by the test developer. Appropriate levels of difficulty are a function of both the type of test and the objectives upon which the test is based. In any case, test developers should balance item difficulty by including items from all levels of the continuum.

Item novelty is a concept that applies to primarily to measurement of higher-order outcomes. It refers to conditions where the situation in the item is one that is totally unfamiliar to the student, that is, one that was not previously encountered during instruction. If this condition is not met, an item that is intended to measure high order outcome may in fact be a knowledge item.

Essay Tests

Essay tests are primarily used by teachers for student evaluation. Some standardized tests do contain essay items, and a formal evaluation effort may occasionally utilize such items but their use is generally confined to the classroom. An essay test is one which contains items which require the student to compose responses, usually lengthy ones. Essay test is

appropriate for measuring outcomes that involve high-order skills, such as application and synthesis. It should be noted that the scoring of objectives tests is generally subjective. This means that there is no only one and answer; but there can be a number of alternative ways that would correctly answer the question involved. Because essay responses are lengthier than those of objective tests, fewer questions can be asked in a given amount of time.

Objective items are more difficult to construct but easier to score; essay tests, on the other hand, are tough to score but relatively easy to construct. The general guidelines for item construction – objective-item correspondence, sufficiency, communication, difficulty and novelty – apply with special emphasis on communication, since there are no response alternatives to help clarify the question, it is crucial that the desired response alternatives be delineated as clearly as possible. Students should be instruction on what is required and how the question should be answered. In addition, students should be given general guidelines concerning length, time and scoring.

As already indicated, the scoring of essay tests is difficult and time-consuming process. While it is potentially a very subjective process involving low scorer reliability, the degree of subjectivity can be considerably minimized by carefully planning and scoring. We can objectify essay tests to the degree that we can specify appropriate responses and alternative acceptable responses, and then any knowledgeable scorer should be able to determine whether the responses are there. There are two basic approaches to the process of scoring essay tests, the analytical method and the global method.

The analytical approach involves identifying all the aspects or components of a perfect answer and assigning a point value to each.

The global method also referred to as the rating method or the holistic approach, results in more subjective, less reliable scoring, but it takes less time. It involves identifying all the aspects of or components of a perfect answer, but point values are not assigned to each; instead, each response is judged as a whole, a total unit, and an overall rating is made. Based, on the perceived completeness of the responses, points are assigned.

As a general guideline for essay tests, it is recommended to score each response without knowledge of who wrote it that is anonymously. This prevents scores being biased by extraneous factors, such as feelings about individual students. It is also good idea to read and score essay questions twice, before points are assigned. Ideally, this should be done by two different persons.

Objective Tests

Objective tests, in which students are not required to compose responses but rather to select from among a number of given alternatives, are more valid and reliable than essay items, and scorer reliability is much higher. Objective tests can be used to measure objectives at all levels, and are not confined to knowledge level items, as argued by many. The general guidelines for item construction, which we have previously discussed will generally apply. The

following recommendations can help you avoid common pitfalls.

- Be careful not to provide clues to the right answer
- Dependent items (when correct answer for one item is necessary in order to answer another item) should be avoided.
- Irrelevant difficulty should not be promoted, either intentionally or unintentionally, by using complex vocabulary or by making the item more complicated than it needs to be.
- Avoid negatives, especially double negatives.
- Direct quotes should be avoided.
- Trivia should not be measured.
- An item should have only one correct or best answer, unless otherwise specified.

Objective items include;

- Short answer items
- Multiple choice items
- True-false items
- Matching items

General Guidelines for the Organization, Directions and Format of Tests

Organization of tests items

If there is one type of item, items of the same type should be together.

Items can also be arranged on a continuum, from least difficult to most difficult.

Another logical way to arrange items is by subject matter topic. All items related to the same objective or outcome should be grouped together.

General Directions

If the purpose and importance of the test are not self-evident, they should be discussed with the students in order to orient them and motivate them to do their best.

In additions to directions for selecting (or supplying) answers, students should be instructed as how to record the answers.

If directions are complicated, or involve procedures unfamiliar to the students, an example or sample item should be presented to the students.

It is good idea that the test take develops scorer key.

If some topics are more important than others, it is better to have more items for these topics than to assign some items a higher value.

If test taking time is not ample, students should be given some guidelines for using their time effectively.

If there is not penalty for guessing, students should be encouraged to answer every item, even if are not sure of the correct answer.

If there is penalty for guessing, students should be instructed to respond only to those items for which are reasonably sure concerning the correct answer.

Format considerations

Items should be numbered consecutively, and directions should be prominently positioned and separated from the actual items.

Answer spaces should be of equal size and large enough to accommodate the longest response, especially on short-answer tests, and answer spaces should be placed in a vertical column in the right or left hand margin or a separate answer sheet may be used.

If an item is accompanied by any type of illustration, such as diagrams, it should be accurate and placed adjacent to the item, directly above it if possible or parallel to it.

Test revision

Test revision may actually occur prior to its intended use as well as following its formal administration. Instruments developed for specific, formal evaluation may be administered and revised over and over again until acceptable levels of validity and reliability are achieved. The process is called field testing or test validation.

Formative assessment

Choose any two areas of your specialization or interest from secondary school syllabus. You can use the same topics you have chosen in Learning Activity No. 2.

Outline the subtopics of the topics you have chosen. That should come from the syllabus (but if the syllabus does not provide such information, you can develop with the aid of secondary school teacher or by using the relevant textbooks).

Study Readings #7 and develop a table of specifications for the two topics. A sample design for appropriate table of specifications is provided below.

Develop actual tests questions which conform to the guidelines specified in Readings #8, #9, #10 and #11.

Table Specifications

Topic: _____ Class:_____ ___

No.	Topic/ Sub-topic		Level of objectives				
		Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation
1.							
2.							
3.							
4.							
5.							
6.							
7.							
8.							
9.							
10							
11							
12							

Learning Activity #5

ANALYSING, INTERPRETING AND USING TEST RESULTS

SUMMARY

As indicated by the title, this activity is designed to cover concepts related to the process of analysing, interpreting and using test results are outline. In the analysis process, it is pointed out that descriptive statistics are of great help for the teacher to summarize student scores in a way that is comprehensible. We have already indicated that a score itself is meaningless unless interpreted with the use of criteria. Different methods of interpreting test results are outlined in this learning activity. The unit ends with brief discussion on the different ways in which test results are used by different stakeholders in the education process.

List of key Concepts

A raw score is the number of questions a student gets right on a test (assuming each question is worth one point).

Percent Correct is obtained when the raw score is divided by the total number of questions and the result is multiplied by 100.

Grade equivalent is a number that describes a student's location on an achievement continuum. Percentile rank is a score that tells the percent of students in a particular group that got lower raw scores on a test than the student did.

Content

Analysing test Results

The analysis of test results enables the teacher and all users have summarized data which can be easily interpreted instead of long sheets containing student names and test scores. The first step in the data analysis is to is to describe or summarize the data using descriptive statistics. A teacher for example may be interested in only in describing the performance of the class and the achievement level of individual in relation to average class performance. Descriptive statistics permits us to meaningfully describe many scores with a small number of indices by providing answers to questions such as: what was the average score on the test? Were most of the scores close to the average or were considerably spread out? And so forth. Descriptive statistics also permits us to make interpretative statements (both norm-referenced and criterion-referenced, but mainly the former) about individual scores and answer questions such as; is this particular score above the average or below it? How much higher lower? What percentage of the scores was lower than this score? And so forth. Whether locally-developed or standardized instruments are administered, descriptive statistics is important tool in interpreting the results of the results.

The following are types of descriptive statistics:

1. Measurement of central Tendency

    - The mode
    - The median
    - The mean

2. Measures of variability

    - The range
    - The quartile deviation
    - The standard deviation
    - The normal curve

3. Measures of relative position

- Percentile ranks
- Standard scores
- Grade and age equivalents

Interpreting Test Results

Three of the fundamental purposes for testing are (1) to describe each student's developmental level within a test area, (2) to identify a student's areas of relative strength and weakness in subject areas, and (3) to monitor year-to-year growth in the basic skills. To accomplish any one of these purposes, it is important to select the type of score from among those reported that will permit the proper interpretation. Scores such as percentile ranks, grade equivalents, and standard scores differ from one another in the purposes they can serve, the precision with which they describe achievement, and the kind of information they provide. A closer look at these types of scores will help differentiate the functions they can serve and the meanings they can convey.

Types of Scores

Raw Score (RS)

The number of questions a student gets right on a test is the student's raw score (assuming each question is worth one point). By itself, a raw score has little or no meaning. The meaning depends on how many questions are on the test and how hard or easy the questions are. For example, if Olouch got 10 right on both a math test and a science test, it would not be reasonable to conclude that her level of achievement in the two areas is the same.

This illustrates why raw scores are usually converted to other types of scores for interpretation purposes.

Percent Correct (PC)

When the raw score is divided by the total number of questions and the result is multiplied by 100, the percent-correct score is obtained. Like raw scores, percent- correct scores have little meaning by themselves. They tell what percent of the questions a student got right on a test, but unless we know something about the overall difficulty of the test, this information is not very helpful. Percent-correct scores are sometimes incorrectly interpreted as percentile ranks, which are described below. The two are quite different.

Grade Equivalent (GE)

The grade equivalent is a number that describes a student's location on an achievement continuum. The continuum is a number line that describes the lowest level of knowledge or skill on one end (lowest numbers) and the highest level of development on the other end

(highest numbers). The GE is a decimal number that describes performance in terms of grade level and months. For example, if a sixth-grade student obtains a GE of 8.4 on the Vocabulary test, his score is like the one a typical student finishing the fourth month of eighth grade would likely get on the Vocabulary test. The GE of a given raw score on any test indicates the grade level at which the typical student makes this raw score. The digits to the left of the decimal point represent the grade and those to the right represent the month within that grade. Grade equivalents are particularly useful and convenient for measuring individual growth from one year to the next and for estimating a student's developmental status in terms of grade level. But GEs have been criticized because they are some- times misused or are thought to be easily misinterpreted. One point of confusion involves the issue of whether the GE indicates the grade level in which a student should be placed. For example, if a fourth-grade student earns a GE of 6.2 on a fourth-grade reading test, should she be moved to the sixth grade? Obviously the student's developmental level in reading is high relative to her fourth-grade peers, but the test results supply no information about how she would handle the material normally read by students in the early months of sixth grade. Thus, the GE only estimates a student's developmental level; it does not provide a prescription for grade placement. A GE that is much higher or lower than the student's grade level is mainly a sign of exceptional performance.

In sum, all test scores, no matter which type they are or which test they are from, are subject to misinterpretation and misuse. All have limitations or weaknesses that are exaggerated through improper score use. The key is to choose the type of score that will most appropriately allow you to accomplish your purposes for testing. Grade equivalents are particularly suited to estimating a student's developmental status or year-to-year growth. They are particularly ill-suited to identifying a student's standing within a group or to diagnosing areas of relative strength and weakness.

Developmental Standard Score (SS)

Like the grade equivalent (GE), the developmental standard score is also a number that describes a student's location on an achievement continuum. The main draw- back to interpreting developmental standard scores is that they have no built-in meaning. Unlike grade equivalents, for example, which build grade level into the score, developmental standard scores are unfamiliar to most educators, parents, and students. To interpret the SS, the values associated with typical performance in each grade must be used as reference points.

The main advantage of the developmental standard score scale is that it mirrors reality better than the grade-equivalent scale. That is, it shows that year-to-year growth is usually not as great at the upper grades as it is at the lower grades. (Recall that the grade-equivalent scale shows equal average annual growth -- 10 months -- between any pair of grades.) Despite this advantage, the developmental standard scores are much more difficult to interpret than grade equivalents. Consequently, when teachers and counsellors wish to estimate a student's annual growth or current developmental level, grade equivalents are the scores of choice.

The potentials for confusion and misinterpretation that were described in the previous

subsection for the GE are applicable to the SS as well. Relative to the GE, the SS is not as easy to use in describing growth, but it is equally inappropriate for identifying relative strengths and weaknesses of students or for describing a student's standing in a group.

Percentile Rank (PR)

A student's percentile rank is a score that tells the percent of students in a particular group that got lower raw scores on a test than the student did. It shows the student's relative position or rank in a group of students who are in the same grade and who were tested at the same time of year (fall, midyear, or spring) as the student. Thus, for example, if Toni earned a percentile rank of 72 on the Language test, it means that she scored higher than 72 percent of the students in the group with which she is being compared. Of course, it also means that 28 percent of the group scored higher than Toni. Percentile ranks range from 1 to 99.

A student's percentile rank can vary depending on which group is used to determine the ranking. A student is simultaneously a member of many different groups: all students in her classroom, her building, her school district, her state, and the nation.

Types of Score Interpretation

An achievement test is built to help determine how much skill or knowledge students have in a certain area. We use such tests to find out whether students know as much as we expect they should, or whether they know particular things we regard as important. By itself, the raw score from an achievement test does not indicate how much a student knows or how much skill she or he has. More information is needed to decide "how much." The test score must be compared or referenced to something in order to bring meaning to it.

That "something" typically is (a) the scores other students have obtained on the test or (b) a series of detailed descriptions that tell what students at each score point know or which skills they have successfully demonstrated. These two ways of referencing a score to obtain meaning are commonly called norm-referenced and criterion-referenced score interpretations.

Norm-Referenced Interpretation

Standardized achievement batteries are designed mainly to provide for norm-referenced interpretations of the scores obtained from them. For this reason they are commonly called norm-referenced tests. However, the scores also permit criterion-referenced interpretations, as do the scores from most other tests. Thus, norm-referenced tests are devised to enhance norm-referenced interpretations, but they also permit criterion-referenced interpretation.

A norm-referenced interpretation involves comparing a student's score with the scores other students obtained on the same test. How much a student knows is determined by the student's standing or rank within the reference group. High standing is interpreted to mean the student knows a lot or is highly skilled, and low standing means the opposite. Obviously, the overall competence of the norm group affects the interpretation significantly. Ranking

high in an unskilled group may represent lower absolute achievement than ranking low in an exceptional high performing group

Most of the scores on standardized test reports are based on norm-referencing, i.e., comparing with a norm group. In the case of percentile ranks, stanines, and normal curve equivalents, the comparison is with a single group of students in a certain grade who tested at a certain time of year. These are called status scores because they show a student's position or rank within a specified group. However, in the case of grade equivalents and developmental standard scores, the comparison is with a series of reference groups. For example, the performances of students from third grade, fourth grade, fifth grade, and sixth grade are linked together to form a developmental continuum. (In reality, the scale is formed with grade groups from kindergarten up through the end of high school.) These are called developmental scores because they show the students' positions on a developmental scale. Thus, status scores depend on a single group for making comparisons and developmental scores depend on multiple groups that can be linked to form a growth scale.

An achievement battery is a collection of tests in several subject areas, all of which have been standardized with the same group of students. That is, the norms for all tests have been obtained from a single group of students at each grade level. This unique aspect of the achievement battery makes it possible to use the scores to determine skill areas of relative strength and weakness for individual students or class groups, and to estimate year-to-year growth. The use of a battery of tests having a common norm group enables educators to make statements such as "Asha is better in mathematics than in reading" or "Ayan has shown less growth in language skills than the typical student in his grade." If norms were not available, there would be no basis for statements like these.

Norms also allow students to be compared with other students and schools to be compared with other schools. If making these comparisons were the sole reason for using a standardized achievement battery, then the time, effort, and cost associated with testing would have to be questioned. However, such comparisons do give educators the opportunity to look at the achievement levels of students in relation to a nationally representative student group. Thus, teachers and administrators get an "external" look at the performance of their students, one that is independent of the school's own assessments of student learning. As long as our population continues to be highly mobile and students compete nationally rather than locally for educational and economic opportunities, student and school comparisons with a national norm group should be of interest to students, parents, and educators.

A common misunderstanding about the use of norms has to do with the effect of testing at different times of the year. For example, it is widely believed that students who are tested in the spring of fourth grade will score higher than those who are tested in the fall of fourth grade with the same test. In terms of grade- equivalent scores, this is true because students should have moved higher on the developmental continuum from fall to spring. But in terms of percentile ranks, this belief is false. If students have made typical progress from fall to spring of grade 4, their standing among fourth-grade students should be the same at both

times of the year. (The student whose percentile rank in reading is 60 in the fall is likely to have the same percentile rank when given the same test in the spring.) The reason for this, of course, is that separate norms for fourth grade are available for the fall and the spring. Obviously, the percentile ranks would be as different as the grade equivalents if the norms for fourth grade were for the entire year, regardless of the time of testing. Those who believe students should be tested only in the spring because their scores will "look better" are misinformed about the nature of norms and their role in score interpretation.

Scores from a norm-referenced test do not tell what students know and what they do not know. They tell only how a given student's knowledge or skill compares with that of others in the norm group. Only after reviewing a detailed content outline of the test or inspecting the actual items is it possible to make interpretations about what a student knows. This caveat is not unique to norm-referenced interpretations, however. In order to use a test score to determine what a student knows, we must examine the test tasks presented to the student and then infer or generalize about what he or she knows.

Criterion-Referenced Interpretation

A criterion-referenced interpretation involves comparing a student's score with a subjective standard of performance rather than with the performance of a norm group. Deciding whether a student has mastered a skill or demonstrated minimum acceptable performance involves a criterion-referenced interpretation. Usually percent-correct scores are used and the teacher determines the score needed for mastery or for passing.

When making a criterion-referenced interpretation, it is critical that the content area covered by the test -- the domain -- be described in detail. It is also important that the test questions for that domain cover the important areas of the domain.

In addition, there should be enough questions on the topic to provide the students ample opportunity to show what they know and to minimize the influence of errors in their scores.

Some of the criterion-referenced tests cover such a wide range of content or skills that good criterion-referenced interpretations are difficult to make with the test scores. However, in most tests the separate skills are defined carefully, and there are enough questions measuring them to make good criterion-referenced interpretations of the skill scores possible.

The percent-correct score is the type used most widely for making criterion-referenced interpretations. Criterion scores that define various levels of performance on the tests are generally percent-correct scores arrived at through teacher analysis and judgment.

Using test Results

Evaluation is the systematic process of collecting and analysing data in order to make decisions. Thus, the results of data analysis are used for decision making. The types of decision that follow evaluation are highly varied. They range from the determination that Student X is not making satisfactory progress this term, to the conclusion that the educational

system of Country Y is not achieving its objectives. The following types of people, both inside and outside the school system, are interested in test results:

- Legislators
- School boards
- Administrators
- Parents
- Students
- Policy makers

Formative Assessment

In the table below, you are given the data for the results of math and english exams, taken at the end of the academic year. Students who took the exams were in form II class in their secondary school.

Do the following activities using the data provided.

1. Study Reading #13 and understand the different methods of analysing exam results and then do the following.

Plot a bar chart to show the distribution of the students. b. Calculate the mean, median and mode of the data.

Calculate the standard deviation of the data.

Calculate the correlation between students' scores on the two exams.

2. Interpret the data by using the two major ways in which exam results are interpreted.

a. First use the Norm-referenced approach.

Plot the normal curve to show student distribution.

How many students fall in each category of the curve?

What can you say about the score or student number 19? How was that students' performance?

How can these results be used?

What are the weaknesses in this method of interpreting test results?

b. Secondly use the Criterion-referenced approach to interpret the same data.

Develop hypothetical criteria to interpret the data.

Plot histogram to show the number of students who fall in each category of your criteria.

Student ID	Mathematics	English I
1.	62	56
2.	45	57
3.	68	78
4.	90	8
5.	89	56
6.	67	88
7.	55	99
8.	44	90
9.	68	98
10.	98	80
11.	78	60
12.	79	40
13.	20	56
14.	49	78
15.	33	88
16.	88	56
17.	99	88
18.	100	77
19.	34	66
20.	56	55
21.	100	23
22.	77	67
23.	78	88
24.	65	34
25.	54	100
26.	21	90
27.	23	99
28.	47	100
29.	67	87
30.	68	77

The overall score for both exams is 100.

XIII. Summative evaluation

This activity is intended to assess what you have been able to learn from this module. It is time we should put what you have so far learned in educational testing and evaluation to use. Therefore, in collaboration with your teachers and your institution, you should have access to teach (secondary school is recommended, but if not possible, primary or college level will be ok).

Before you start the actual teaching, you should get the necessary syllabuses and textbooks. Also, before you start teaching, you should prepare the mode of assessment and evaluation you would conduct after you finish teaching. After you have finished teaching and conducted the exam, you are required to write a report that is between 1000– 2000 words long.

Contents of the Report

- Cover page
- Summary
- Table of contents
- Background
- Planning phase
- Situation analysis
- Specification of objectives
- Specification of pre-requisites
- Selection and development of measuring instruments e.g Delineation of strategies
- Preparation of time schedule
- Process phase
- Pretesting
- Conducting the evaluation
- Product phase
- Test results
- Analysis of results
- Interpretation of results
- Recommendations to stakeholders
- Self-valuation
- Conclusion

Grading Scheme

The report constitutes total marks of 100%. The following guideline will help the teacher of the module mark the report and give an idea of what is expected.

- Cover page (5 marks)
- Summary (5 marks)
- Table of contents (5 marks)
- Background (5 marks)

- Planning phase (30 marks)
- Situation analysis
- Specification of objectives
- Specification of pre-requisites
- Selection and development of measuring instruments
- Delineation of strategies
- Preparation of time schedule
- Process phase (15 marks)
- Pretesting
- Conducting the evaluation
- Product phase (25 marks)
- Test results
- Analysis of results
- Interpretation of results
- Recommendations to stakeholders
- Self-valuation
- Conclusion (10 marks)

XIV. Synthesis of the Module

The module was covered the most important concepts underlying educational evaluation and testing. It was the expectation of the module that students who mastered the content discussed in this module will be able to understand the process of conducting effective evaluation in education and conduct one. The module has five units. Unit I (Educational Evaluation) introduces the concept, nature, types and phases of evaluation. This unit serves as introduction for the module, since it gives learners a picture of evaluation and its role in education. The second unit (Specification of Educational Objectives) highlights in detail one of the most important steps in the process of evaluation. Evaluation starts with objectives and aims to verify the level of attainment of these objectives. In this unit, learners are taught how to develop educational objectives that re- present the whole spectrum of domains of learning and that are measurable. In the third unit, the process of classifying tests and selecting one(s) that meet the needs of the teacher are discussed. In this unit, the learner is given a chance to see how he/she can take advantage of existing standardized tests in his/her tests. The fourth Unit is dedicated to the process of developing instruments to be used in the evaluation process. The learner is exposed to various techniques of designing the test, constructing the items, and revising them. The fifth and last unit focuses on how to analyse, interpret and use test results. This unit is intended to help the learner think of the use of test results and get skills that help him/her use the results appropriately.

XV. Author of the Module

Mr. Ridwan Mohamed Osman is currently the Dean of the Faculty of Education at Amoud University, Somaliland. He is also the Coordinator of AVU Teacher Education Program in the University. He has been a lecturer of Research Methods and Education in the same university and other affiliated institutions for the last five years. Mr. Osman got his Bachelor's Degree in Education from Amoud University. He has also received his Master's Degree in Education from Egerton University in Kenya. Currently his research interest lies in the areas of teaching science in African Secondary schools, classroom management, educational testing and evaluation and indigenous education. For any further information, you can contact in the following email address.

Email: ridwaanxaaji@hotmail.com

Tel: +25224457020

XV. Reviewer of the Module

Augustine Mwangi Holds a Bachelor of Education Arts (B.Ed. Arts); Master of Education M.Ed Egerton University. Currently, he is a PhD candidate at the University of Nairobi. He has taught at the University of Nairobi, School of Continuing and Distance Education for the last 10 yrs. He has been a researcher with the Panafrican Research Agenda on the Pedagogical Integration of ICT in Education. He has vast experience in instructional design. He has worked as educational and Instructional design consultant in South Sudan (2010-2011) Pre-service teacher Training, Zambia (2013 and 2015) for the ZNLTP project, Malawi (2014) and Lesotho 2015 ICAP/Nurse Education Partnership Initiative (NEPI)

E mail: augustine.gatotoh@uonbi.ac.ke

augmwa2002@gmail.com

The African Virtual University
Headquarters
Cape Office Park
Ring Road Kilimani
PO Box 25405-00603
Nairobi, Kenya
Tel: +254 20 25283333
contact@avu.org
oer@avu.org

The African Virtual University Regional Office in Dakar
Université Virtuelle Africaine
Bureau Régional de l'Afrique de l'Ouest
Sicap Liberté VI Extension
Villa No.8 VDN
B.P. 50609 Dakar, Sénégal
Tel: +221 338670324
bureauregional@avu.org

2017 AVU