COURSE GUIDE

EDU 756
MEASUREMENT AND EVALUATION

Course Team Dr. Osuji . U.S.A (Course Developer/Co-ordinator) – NOUN
Dr. Okonkwo. C.A. (Course Writers) – NOUN
Dr Nnachi R.O. (Course Editor) - Abia State University, Abia – Imo State

NATIONAL OPEN UNIVERSITY OF NIGERIA

National Open University of Nigeria
Headquarters
University Village
Plot 91, Cadasral Zone,
Nnamdi Azikwe Express way
Jabi-Abuja

Lagos Office
14/16 Ahmadu Bello Way
Victoria Island, Lagos

e-mail: centralinfo@nou.edu.ng
URL: www.nou.edu.ng

Published by
National Open University of Nigeria

Printed 2006

Reprinted 2016

ISBN: 978-058-160-3

CONTENTS **PAGE**

INTRODUCTION

This course EDU 756 Measurement and Evaluation is a one semester, two-credit unit course which is designed and made available to all students undertaking their first degree programme in education.

The course consists of 18 units. The course will involve an overview of measurement and evaluation, educational objectives, types of tests, test development and administration, Quality of classroom test and problems.

The materials have been developed to suit Nigerian students by using examples from the local environment.

There are no prerequisites for this course. But it is expected that you have studied some other courses in education.

This course guide tells you in a nutshell what the course is all about, the course materials you will require to be using and how you can work your way through these materials. Some general guidelines have been suggested to enable you complete this course without any difficulties. Some guidance on your tutor-marked assignments has also been given. Detailed information on the tutor-marked assignments will be made available to you at your study center when you start your tutorials.

There are some tutorial sessions which are linked up with this course. You are advised to attend these sessions. Details of times and venues of these tutorials will be made known to you from your study center.

WHAT YOU WILL LEARN IN THIS COURSE

The overall aim of this Course EDU 756 Educational Measurement and Evaluation is to introduce you to the fundamental principles of educational measurement and evaluation which you will need to apply as a teacher. In this course you will learn the basic elements that will enable you function effectively as a teacher, especially in the area of students assessment. You will also learn how to specify the objectives in a measurement and behavioural terms and how to evaluate the objectives.

COURSE AIMS

This course aims to give you an understanding of the fundamental principles of educational measurement and evaluation and how these could be applied in the assessment of students learning outcomes and classroom achievements.

This will be achieved by aiming to:

- Give you an overview of educational measurement and evaluation.
- Outline the educational objectives and their taxonomies.
- Give you the practical functions of tests in education.
- Explain how to measure educational outcomes in cognitive, affective and psychomotor domains.
- Explain the various types of tests and their uses.
- Introduce you to test reliability and validity.
- Explain quality control in marking system.

COURSE OBJECTIVES

Some overall objectives have been specified so as to achieve the aims set out above. In addition, each unit has specific objectives set out. These unit objectives are always included at the beginning of the unit. You should try to read them before you start working through the unit. You may also refer to them during your study of the unit to check on your progress and after completing the unit. In this way you can be sure that you have done what was required of you by the unit.

The objectives of the whole course are set out below. By meeting these objectives you should have achieved the aims of the course as a whole. On successful completion of the course, you should be able to:

1. Mention the purposes of measurement and evaluation.
2. Describe the historical development of testing and evaluation.
3. Enumerate the importance and functions of test in education.
4. Explain the concept of educational objectives
5. Discuss the taxonomy of educational objectives.
6. Describe the domains of educational objectives.
7. List the uses of classroom test.
8. List the types of test used in the classroom.
9. Enumerate the advantages and disadvantages of essay and objectives types of test.
10. Explain test administration and scoring
11. Estimate and interpret the reliability of a test.
12. Explain the validity of a test as an instrument
13. Describe the problems of marking tests.
14. Explain quality control in marking system.

WORKING THROUGH THIS COURSE

To complete this course successfully, you are required to read the study units, read other textbooks in measurement and evaluation. Each unit contains self-assessment exercises. You are also required to complete and submit the tutor-marked assignments for assessment purposes. At the end of the course or semester, you are expected to sit for a final examination. This course should take you about 15 weeks in total to complete. The components of the course have been listed below to enable you plan very well for your success in the course.

COURSE MATERIALS

The major components of this course are:

1. Course Guide
2. Study units
3. Textbooks
4. Assignment file
5. Presentation Schedule

STUDY UNITS

The study units in this course are:

Module 1 An Overview of Measurement and Evaluation

Unit 1 Definitions and purposes of Measurement and Evaluation
Unit 2 Historical development of testing and Evaluation
Unit 3 Importance and Functions of Tests in Education

Module 2 Educational Objectives

Unit 1 Educational objectives
Unit 2 Bloom's Taxonomy of Educational objectives
Unit 3 The Affective Domain
Unit 4 The psychomotor Domain
Unit 5 The Classroom Tests

Module 3

Unit 1 Types of Test
Unit 2 Essay Test
Unit 3 Objective Test
Unit 4 Test Development – Planning the Classroom Test
Unit 5 The Administration and Scoring of Classroom Test

Module 4

Unit 1	Judging The Quality Of A Classroom Test
Unit 2	Interpreting Classroom Test Scores
Unit 3	Reliability Of A Test
Unit 4	Validity Of Classroom Test
Unit 5	Problem Of Marking Test And Quality Control In Marking System

SET TEXTBOOKS

There are no compulsory set books. This material is designed to be self contained, but if you should need any other text, you can purchase any good text in measurement and evaluation, especially from among the ones in the references.

ASSIGNMENT FILE

The assignment file will be given to you on registration in this course. You will use this file to submit all the tutor marked assignments to your tutor for assessment. Note that the marks you obtain from these assignments will count towards the final mark you will obtain for this course. Further information on the assignments will be given to you by your tutorial facilitator. You are expected to complete and submit six assignments. Out of these six, four will be graded and added to the marks you will obtain to make up your total score in this course.

PRESENTATION SCHEDULE

The presentation schedule which includes the dates for the completion and submission of the assignments and attendance to tutorials will be available to you at your study center. Remember to submit all assignments by the due dates.

Try to guard against falling behind in your work.

ASSESSMENT

There are two types of assessment in this course. The first one are the tutor marked assignments (TMA) and the second is the examination at the end of the course.

In doing the assignments, you are required to apply information, knowledge and methods drawn from the course. You must submit the assignments to your tutor for assessment and grading according to the

deadlines given to you. The assignments count for 30% of your total course mark.

At the end of the course, you are required to write an examination of about three hours' duration. This examination will count for 70% of your total course mark.

TUTOR-MARKED ASSIGNMENT (TMAs)

All the units have tutor marked assignments. Your tutor will specify the six which you will complete and submit to him. You are encouraged to submit all the six assignments, out of which the best three will be graded at 10% each towards your total course mark.

You will be required to complete the assignments using materials from the study units and set books. It is desirable in all degree level education to demonstrate that you have read and researched more widely than the required minimum. Using the reference books will give you broader viewpoints and may provide a deeper understanding of the subject.

When you have completed your assignment, send it to your tutorial facilitator in the assignment file. If for any reason, you cannot submit your work on time, contact your tutor before the due date to discuss the possibility of an extension. Extensions will not be granted after the due date unless there are exceptional circumstances.

FINAL EXAMINATION AND GRADING

The final examination for this course will be for three hours' duration and will have a value of 70% of the total course grade. The examination questions will reflect the types of self-testing, activities and tutor-marked assignments which you have been meeting.

Use the time between finishing the last unit and sitting the examination to revise the entire course. You might find it useful to review yourself tests, tutor-marked assignments and comments on them before the examination. The examination covers all parts of the course.

COURSE MARKING SCHEME

The table below shows how the actual marking is broken.

Assessment	Marks
Assignments	Six assignments, best three marks of the six count 10% each = 30% of course marks.
Final examination	70% of overall course marks
Total	100% of course marks

HOW TO GET THE BEST FROM THIS COURSE

Distance learning involves the teacher and the learners working apart from each other. It means that you can read and study the self-learning materials at your own time, pace and place and at your own convenience. The self-learning material has replaced the lecturer and will guide you and direct you the same way the lecturer will do in the class.

Just as the lecturer might give you some exercises and assignments, the same way the study units provide the exercises and assignments for you to do at the appropriate points.

The study units follow that same format. They start with the contents, the introduction, the objectives, the subject matter, the conclusion and summary.

There are activities and exercises which you have to do to enable you check your progress and also to enable you achieve the objectives.

The under listed are practical strategies to help you work successfully through this course.

1. Read the course guide thoroughly.
2. Organize a study schedule to enable you study without distractions. Keep the due dates and important information especially details of your tutorials, and dates for the submission of assignments.
3. Do everything possible to stick to your study schedule. Students fail because they get behind with their course work. If you have any difficulties with your schedule contact your tutor before it is late.
4. Assemble your study materials before you start.
5. For every unit, start reading from the introduction and the objectives.
6. As you work through the unit, you will see that the content has been arranged sequentially for you to follow.
7. Keep in touch with your study center for information.
8. Try to do all the assignments and try to submit in good time. The assignments are designed to help you achieve the objectives of the course and therefore will help you to pass the examinations.
9. Review the objectives for each unit to confirm that you have achieved them. If you feel unsure about any of the objectives, review the study material or consult your tutor.
10. Proceed unit by unit through the course. Try to pace your study so as to keep yourself on schedule.

11. If you submit an assignment to your tutor for marking, go ahead and continue to read. Do not wait for the return of the assignment before you start reading.
12. Pay particular attention to the teacher's comments. If you have any questions or problem consult your tutor.
13. On completion of the last unit, review the course and prepare yourself for the examination. Check that you have achieved the objectives listed.

TUTORS AND TUTORIALS

There are tutorial sessions provided in support of this course. The dates, times and venue will be given to you at the center. You will also get the name and phone number of your tutor as soon as you register for this course.

Your tutor will mark and comment on your assignments, keep watch on your progress and/or difficulties and provide assistance to you during the course. You must submit your assignment before the due dates. They will be marked and returned to you as soon as possible.

Do not hesitate to contact your tutor by phone, e-mail or contact the course coordinator of this course if you have any problem, or you need help.

Contact your teacher if you do not understand any part of this course ort you have difficulty with the activities or exercises or you have problem with the assignment, or with the tutors' comments.

You should try to attend the tutorials. This is the only chance to have face contact with your tutor and your peers and to ask questions which are answered instantly.

SUMMARY

This course measurement and evaluation intends to introduce you to the fundamental principles of students' assessment and teaching. Upon the completion of this course, you will be equipped with the basic knowledge of the details and guiding principles in evaluation. You will be able to answer such questions as:

- What is measurement and evaluation?
- What is assessment and testing?
- What are the functions of a test?
- What are the various types of test?

- How do you measure outcomes in affective and psychomotor domains?
- How can we determine the reliability and validity of a test?
- What are the quality control measures in marking?

There are many more questions to be asked.

We wish you success in this course and hope you will find it interesting and useful.

Enjoy your study with NOUN. Good luck!

MAIN COURSE

CONTENTS **PAGE**

MODULE 1 AN OVERVIEW OF MEASUREMENT AND EVALUATION

UNIT 1 DEFINITIONS AND PURPOSES OF MEASUREMENT AND EVALUATION

CONTENTS

1.0 INTRODUCTION

In this unit, the meaning of assessment, test and testing measurement and evaluation will be given. The purpose of carrying out measurement and evaluation will also be given since the primary purpose of educationally measuring and evaluating the learner is to utilize the results for the improvement of teaching-learning.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain the terms test and testing
- define the term assessment

- clarify the terms measurement and evaluation
- list the purposes of measurement and evaluation
- explain the types of evaluation.

3.0 MAIN CONTENT

3.1 Meaning of Terms

3.1.1 Test and testing

Simply put a test is a measuring tool or instrument in education. More specifically, a test is considered to be a kind or class of measurement device typically used to find out something about a person. Most of the times, when you finish a lesson or lessons in a week, your teacher gives you a test. This test is an instrument given to you by the teacher in order to obtain data on which you are judged. It is an educationally common type of device which an individual completes himself or herself, the intent is to determine changes or gains resulting from such instruments as inventory, questionnaire, opinionnaire, scale etc.

Testing on the other hand is the process of administering the test on the pupils. In other words the process of making you or letting you take the test in order to obtain a quantitative representation of the cognitive or non-cognitive traits you possess is called testing. So the instrument or tool is the test and the process of administering the test is testing.

3.1.2 Assessment

Now that you have learnt the difference between test and testing, Let us move on to the next concept which is assessment. As a teacher, you will be inevitably involved in assessing learners; therefore you should have a clear knowledge and the meaning of assessment.

The term assess is derived from a Latin word “asoidere” meaning “to sit by” in judgment. There are many definitions and explanations of assessment in education.

Let us look at few of them.

i. Freeman and Lewis (1998) to assess is to judge the extent of students’ learning.
ii. Rowntree (1977): Assessment in education can be thought of as occurring whenever one person, in some kind of interaction, direct or indirect, with another, is conscious of obtaining and interpreting information about the knowledge and understanding,

of abilities and attitudes of that other person. To some extent or other, it is an attempt to know the person.

iii. Erwin, in Brown and Knight, (1994). Assessment is a systematic basis for making inference about the learning and development of students… the process of defining, selecting, designing, collecting, analyzing, interpreting and using information to increase students' learning and development.

You will have to note from these definitions that

- Assessment is a human activity.
- Assessment involves interaction, which aims at seeking to understand what the learners have achieved.
- Assessment can be formal or informal.
- Assessment may be descriptive rather than judgment in nature.
- Its role is to increase students' learning and development
- It helps learners to diagnose their problems and to improve the quality of their subsequent learning.

SELF-ASSESSMENT EXERCISE 1

i. What is the major difference between test and testing?
ii. In your own words define Assessment.
iii. Give an example of a test.

3.1.3 Measurement

This is a broad term that refers to the systematic determination of outcomes or characteristics by means of some sort of assessment device. It is a systematic process of obtaining the quantified degree to which a trait or an attribute is present in an individual or object. In other words it is a systematic assignment of numerical values or figures to a trait or an attribute in a person or object. For instance what is the height of Uche? What is the weight of the meat? What is the length of the classroom? In education, the numerical value of scholastics ability, aptitude, achievement etc can be measured and obtained using instruments such as paper and pencil test. It means that the values of the attribute are translated into numbers by measurement.

SELF-ASSESSMENT EXERCISE 2

List the instruments which you can use to measure the following: weight, height, length, achievement in Mathematics, performance of students in Technical drawing, attitude of workers towards delay in the payment of salaries.

3.1.4 Evaluation

According to Tuckman (1975) evaluation is a process wherein the parts, processes, or outcomes of a programme are examined to see whether they are satisfactory, particularly with reference to the stated objectives of the programme, our own expectations, or our own standards of excellence.

According to Cronbach et al (1980) evaluation means the systematic examination of events occurring in and consequent on a contemporary programme. It is an examination conducted to assist in improving this programme and other programmes having the same general purpose.

For Thorpe (1993) evaluation is the collection analysis and interpretation of information about training as part of a recognized process of judging its effectiveness, its efficiency and any other outcomes it may have.

If you study these definitions very well, you will note that evaluation as an integral part of the instructional process involves three steps. These are

i. Identifying and defining the intended outcomes.
ii. Constructing or selecting tests and other evaluation tools relevant to the specified outcomes, and
iii. Using the evaluation results to improve learning and teaching.

You will also note that evaluation is a continuous process. It is essential in all fields of teaching and learning activity where judgment need to be made.

SELF-ASSESSMENT EXERCISE 3

Explain the difference between measurement and evaluation, assessment and testing.

3.2 Types of Evaluation

The different types of evaluation are: placement, formative, diagnostic and summative evaluations.

3.2.1 Placement Evaluation

This is a type of evaluations carried out in order to fix the students in the appropriate group or class. In some schools for instance, students are assigned to classes according to their subject combinations, such as

science, Technical, arts, Commercial etc. before this is done an examination will be carried out. This is in form of pretest or aptitude test. It can also be a type of evaluation made by the teacher to find out the entry behaviour of his students before he starts teaching. This may help the teacher to adjust his lesson plan. Tests like readiness tests, ability tests, aptitude tests and achievement tests can be used.

3.2.2 Formative Evaluation

This is a type of evaluation designed to help both the student and teacher to pinpoint areas where the student has failed to learn so that this failure may be rectified. It provides a feedback to the teacher and the student and thus estimating teaching success e.g. weekly tests, terminal examinations etc.

3.2.3 Diagnostic Evaluation

This type of evaluation is carried out most of the time as a follow up evaluation to formative evaluation. As a teacher, you have used formative evaluation to identify some weaknesses in your students. You have also applied some corrective measures which have not showed success. What you will now do is to design a type of diagnostic test, which is applied during instruction to find out the underlying cause of students persistent learning difficulties. These diagnostic tests can be in the form of achievement tests, performance test, self rating, interviews observations, etc.

SELF-ASSESSMENT EXERCISE 1

What are the major differences and similarities between formative evaluation and diagnostic evaluation?

3.2.4 Summative evaluation

This is the type of evaluation carried out at the end of the course of instruction to determine the extent to which the objectives have been achieved. It is called a summarizing evaluation because it looks at the entire course of instruction or programme and can pass judgment on both the teacher and students, the curriculum and the entire system. It is used for certification. Think of the educational certificates you have acquired from examination bodies such as WAEC, NECO, etc. These were awarded to you after you had gone through some types of examination. This is an example of summative evaluation.

3.3 The Purpose of Measurement and Evaluation.

The main purposes of measurement and evaluation are:

i. Placement of student, which involves bringing students appropriately in the learning sequence and classification or streaming of students according to ability or subjects.
ii. Selecting the students for courses – general, professional, technical, commercial etc.
iii. Certification: This helps to certify that a student has achieved a particular level of performance.
iv. Stimulating learning: this can be motivation of the student or teacher, providing feedback, suggesting suitable practice etc.
v. Improving teaching: by helping to review the effectiveness of teaching arrangements.
vi. For research purposes.
vii. For guidance and counseling services.
viii. For modification of the curriculum purposes.
ix. For the purpose of selecting students for employment
x. For modification of teaching methods.
xi. For the purposes of promotions to the student.
xii. For reporting students progress to their parents.
xiii. For the awards of scholarship and merit awards.
xiv. For the admission of students into educational institutions.
xv. For the maintenance of students.

4.0 CONCLUSION

Now that you have gone through the descriptions of the major terms used in measurement and evaluation and you can give the purposes of measurement and evaluation as well as explain the types of evaluation, you have placed yourself on a good footing for the study of this all important course which you can not do without as a teacher.

5.0 SUMMARY

In general, those practitioners in the educational system are most of the times interested in ascertaining the outputs of the educational programme. Output is counted in terms of test results which are naturally expressed in quantitative indices such as scores or marks. Test, which is a device, an instrument or a tool consisting of a set of tasks or questions, is used to obtain the results. Test can be in the form of pen and paper examination, assignments, practical etc. The process of administering this test is called testing. But an act of measurement is done when we award marks to an answer paper or assignment.

So measurement gives the individuals ability in numerical indices of scores i.e. measurement is quantitative. Assessment can be seen as the engine that drives and shapes learning, rather than simply an end of term examination that grades and reports performance. Evaluation is expressed in qualitative indices such as good, excellent pass or fail.

Value judgment is therefore attached to the measurement.
Evaluation can be placement, formative, diagnostic or summative

6.0 TUTOR-MARKED ASSIGNMENT (TMA)

1. What are the types of evaluation?
2. Give the difference between Test and testing?
3. List 5 purposes of measurement and evaluation?

7.0 REFERENCES/FURTHER READING

Obimba, F.U. (1989) Fundamental of Measurement and Evaluation in Education and Psychology. Owerri, Totan Pub. Ltd.

Ohuche, R.O. and Akeju, S.A. (1977) Testing and Evaluation in Education. Lagos: African Educational Resources (AER)

STRAIDE Handbook (2002) Assessment and Evaluation in Distance Education**.** New Delhi: A Publication of Indira Gandhi National Open University (IGNOU).

Tuchman, B.W. (1975) Measuring Educational Outcomes Fundamental of Testing. Atlanta: Harcourt Brace Jovanovich inc.

UNIT 2 HISTORICAL DEVELOPMENT OF TESTING AND EVALUATION

CONTENTS

1.0 INTRODUCTION

In the last unit, you read through important definitions in measurement and evaluation, you saw the types of evaluation and the purposes of evaluation. In this unit we shall move another step to look at the historical development of testing and evaluation. This may help you to appreciate the course the more, and also appreciate the early players in the course.

2.0 OBJECTIVES

At the end of this unit, you shall be able to

- trace the historical development of testing and evaluation
- mention some of the early players in testing and evaluation
- mention some of the testing organizations in Nigeria.

3.0 MAIN CONTENT

3.1 Early Measurement of Individual Differences

The history of measurement started with the invention of tests to measure individual differences in skills among adults. In January 1796, the astronomer royal of Greenwich observatory in England – Maskelyne, was recorded to have dismissed his assistant, Kinnebrook, for recording the movement of stars across the telescope with eight-tenths of a second later than his. According to Tuckman (1975) between 1820 and 1823, a German astronomer –Bessel improved on the work of Maskelyne by demonstrating the variability in personal equations and observations. He argued that fluctuations existed from occasion to occasion and from person to person. This means that there is a variation

in the simple reaction time or a measure of the time required to react to a simple stimulus.

In 1863, a half cousin of Charles Darwin, Sir Francis Galton worked on individual differences. In 1883 he published a book titled inquires into the Human Faculty and Development. His work was regarded as the beginning of mental tests. Have you heard of his tests? In 1884, Galton opened an anthropometric laboratory to collect the characteristic measurements of people within the same period Mckeen Cathel, an American Psychologist was also studying individual differences in primary physical terms.

These were the earliest recorded history of testing. But you will note that early measurement approaches in history both written and oral, were informal. The first written tests were the informal examinations used by the Chinese to recruit people into the civil service. This was about 2200BC.

The oral examinations conducted by Socrates in the 5th century B.C were informal. In America, educational achievement tests were used for assessment through oral examinations before 1815. You have read about Galton, James Cattel etc and their roles in the history of test development. There are others, let us briefly mention very few of them in this section. Karl Pearson, developed the Pearson product-moment correlation coefficient which is useful in checking the reliability and validity of standardized tests.

Edward L. Thorndike was a former student of Cattels. He made major contributions in achievement testing.

By 1904, Alfred Binet had established himself as France's premier psychologist, expert in human individual differences. He studied the differences between bright and dull children. In 1904, he developed a test for measuring intelligence of children with his assistant Theodore Simons. This test is called Binet – Simons intelligence test. In 1916, Louis Terman and his associates at Stanford University revised the Binet-Simon scale and brought out the Stamford-Binet version.

Group-tests development started during the World War I when the need to measure the intelligence of soldiers so as to assign them to different tasks and operations arose. As a result group of psychologists including Yerkes, R.M and Otis, A. developed the Army Alpha, which is a written group intelligence test, and Army Beta, which is the individual non-verbal intelligence test. Others are: David Wechsler who developed series of individual intelligence scales from 1939 to 1967;

George Fisher, an Englishman who developed the first standardized objectives test of achievement in 1864 and J.M. Rice, an American developed the standard spelling objective scale in 1897. The list is not comprehensive yet.

Meanwhile, let us come to the Nigerian situation.

SELF-ASSESSMENT EXERCISE 1

i. Mention five people in connection with the history of test and measurement?
ii. Who developed the following:. i. The first mental test? ii.Pearson product moment correlation coefficient

SELF-ASSESSMENT EXERCISE 1

i. You would have mentioned: Meskelyne of England, Sir Francis Galton, Mckeen Cattel, Socrates, Karl Perason, E.L. Thordike, Alfred Binet, Louis Terman, George Fisher, J.M. Rice, Yerkes, R.M. and Otis, A.
ii. 1. Sir Francis Galton developed the first mental test.
2. Karl Pearson developed the Pearson product moment correlation coefficient.

3.2 Test Organizations in Nigeria

Do you know any organization in Nigeria which conduct examinations or develop examination questions? You must have taken note of WAEC, NECO, JAMB, NABTEB etc.

The West African Examinations Council was established first on 31st December 1951 in Gambia and in 1952 in Nigeria to serve as examination body for the West African countries; i.e. the Anglophone countries. WAEC conducts such examinations as school certificate (GCE), Royal Society of Arts (RSA) and the City and Guilds examinations.

The Joint Admission and Matriculation Board (JAMB) established in 1976 is charged with the responsibility of conducting common entrance examinations for the universities, U.M.E., colleges of Education and Polytechnics in Nigeria. The National Business and Technical Education Board NABTEB is charged with the responsibility of organizing examinations and certification of business and Technical or Vocational examinations. It is located in Benin.

The National Examinations Council (NECO) is established to organise examinations for school certificates both senior and junior, and the common entrance examinations to the unity schools in Nigeria. It is located in Minna, Niger state.

The International Center for Educational Evaluation (ICEE) is concerned with educational evaluation in Nigeria. It is located at the Institute of Education of the University of Ibadan. It offers higher degree programmes in:

- Evaluation of educational achievement
- Evaluation of innovative educational programmes
- Evaluation of public examinations
- Evaluation of curriculum materials
- Evaluation of teaching and learning strategies

Apart from the National bodies, there are some or most of the state education boards and local education boards that have their assessment units. Sometimes these units can develop joint examinations for secondary school students or primary school pupils as the case may be. Does your state education board organize joint examinations for students? Even the primary school certificate examinations are organized by the states.

Some states organize their own junior school certificate examinations. Even companies these days organize some aptitude tests for recruitment or employment exercises.
Some ministries organize promotion examinations for their staff. These are some types of tests and measurement.

SELF-ASSESSMENT EXERCISE 2

Mention 5 bodies that conduct examinations in Nigeria?

SELF ASSESSMENT EXERCISE 2

Your answers may have included:

- West African Examinations Council (WAEC)
- National Examinations Council (NECO)
- Joint Admissions and Matriculation Board (JAMB)
- National Business and Technical Education Board (NABTEB)
- International Centre For Educational Evaluation (ICEE)

4.0 CONCLUSION

Testing can be very helpful if its use increases the learning and `performance of children. This is why, you have seen, that, the history of testing started very early, it has grown from the test of individual differences to almost all aspects of education and human life. Hence there is no aspect of life that can be mentioned where there is no form of measurement or the other. This is because test from the best means of detecting characteristics in a reasonable objective fashion. They help us gain the kinds of information about learners and learning that we need to help students learn.

5.0 SUMMARY

In this unit, we have traced the history of test and measurement, starting from 1796 when the astronomer royal of Greenwich recheck in England – Maskelyne dismissed his assistant for recoding observations which were less by 8/10 of his own. You noted that the variation in the records was as a result of individual differences in the handling of the instruments of observation. You will recall that it was between 1820 and 1823 that a German astronomer – Bessel worked on variability in individual observations and came out with the simple reaction time. Sir Francis Galton was associated with the development of the mental test.

Galton worked in his enthropometric laboratory to collect the characteristic measurements of people. The American psychologist Mckeen Cattel studied individual differences in primary physical terms.

These measurements were all informal. Do you remember the Chinese, Socrates, Pearson, Thorndike, Binet, Terman etc, all had something to do in the early development of test and measurement. In Nigeria, organizations like WAEC, NECO, NABTEB, JAMB, ICEE are all in the activities of measurement.

We shall see the importance of tests in education in the next unit. This will help you place a value on this course.

6.0 TUTOR-MARKED ASSIGNMENT

1. Who was the first person to develop a mental test?
2. Who was the first American psychologist to study individual differences in primary physical terms?
3. Give the names of three persons involved in the early development of test and measurement in history?
4. List 3 organizations concerned with testing in Nigeria?

7.0 REFERENCES/FURTHER READING

Obimba, F.U. (1989) Fundamental of Measurement and Evaluation in Education and Psychology. Abuja, Ibadan, Lagos, Owerri: Totan Publishers Ltd.

Ohuche, R.O. and Akeju, S.A. (1977) Testing and Evaluation in Education. Lagos: Africana Educational Resources (AER).

Tuckman, B.W. (1973) Measuring Educational outcomes Fundamentals of Testing. Atlanta, Chicago, New York, San Francisco, Harcourt Brace Jovanovich Inc.

UNIT 3 IMPORTANCE AND FUNCTIONS OF TESTS IN EDUCATION

CONTENTS

1.0 INTRODUCTION

Tests, as the instruments of measurement and evaluation in education are very important as their functions constitute the rationale for the study of the course. Tests serve various purposes in the educational system. They serve both primary and secondary or supplementary purpose. The primary purpose is to make use of the results of evaluation for the improvement of teaching and learning. Whatever other uses the results are put to are regarded as secondary to this primary use. In this unit we shall be looking at the importance and functions of tests in education and the scales of measurement.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain the functions of tests in education
- explain the measurement scales.

3.0 MAIN CONTENT

3.1 Functions of Tests in Education

In section 3.3 of unit I, we discussed the purposes of measurement and evaluation, but you are aware that test is the instrument used in measurement and evaluation.

Therefore, apart from those purposes, which in one way or the other are applicable to tests, we shall explain the main functions of tests in the educational system in this section. As a teacher, when you set a test for your class, what do you do with the results? How does the test help you in the teaching and learning situation?

According to Nwana (1981) a test can fulfill a variety of functions.

These are:

(a) Motivate pupils to study.
(b) Determine how much the pupils have learned.
(c) Determine the pupils' special difficulties.
(d) Determine the pupils' special abilities.
(e) Determine the strength and weakness of the teaching Method.
(f) Determine the adequacy or otherwise of instructional Resources.
(g) Determine the extent of achievement of the objectives etc.

3.1.1 To Motivate The Pupils To Study

When you go to the church or mosque, every Sunday or Friday, you listen to the preacher as he exposes the Word, you are not subjected to any written or verbal examination on the substance of the preaching. Therefore, the effect of the preaching on you cannot be determined.

Thus whether you are sleeping or paying attention to the preacher is left for you and you alone. This is not the case in a school situation where we need to verify our efforts as soon as possible. This is why test are

regularly used to motivate pupils to learn. They study hard towards their weekly, terminal or end of the year promotion examinations.

Without these test, many of the pupils would be reluctant to make out time for private studies while some of them would be less likely to be attentive when the teacher is teaching, no matter how interesting and lively the teaching may be. You can see that listening to a teacher who is not going to give a test is like listening to the preacher in the church or mosque. But the risk here is that when the tests are too many, pupils start working to pass them.

3.1.2 To Determine How Much The Pupils Have Learned

A test can be used to find out the extent to which the contents have been covered or mastered by the testees. For instance, if you treat a topic in your class at the end you give a test and many of your students score high marks. This is an indication that they have understood the topic very well. But if they score very low marks, it implies that your efforts have been wasted. You need to do more teaching. It is the results of the test that will help you decide whether to move to the next topic or repeat the current topic.

3.1.3 To Determine The Pupils' Special Difficulties

Tests can be constructed and administered to students in order to determine particular problems of students. This is done in order to determine appropriate corrective actions. This identification of weaknesses and strength on the part of the students is the diagnostic use of tests. It helps in the desirable effort to give pupils, individuals or group remedial attention. Can you think of any of such tests? Before you continue do this activity.

SELF-ASSESSMENT EXERCISE 1

What is a major risk or disadvantage of administering too many tests much testing on students? Suggest a remedy for this risk.

SELF-ASSESSMENT EXERCISE 1

Your answers may have included: Too many tests would make students begin to work solely for passing the tests. Most teachers teach for the purpose of their students passing examinations especially public examinations. Students therefore become more examination conscious.

An ideal situation is one in which a sufficient number of tests are given to encourage the students to study and pay attention, but not so many

that they become excessively examination conscious. It takes experience and skill for the teacher to know when the right balance has been achieved.

3.1.4 To Determine The Pupils Special Abilities

Tests can be used as a measure to indicate what a person or a group or persons or students can do. These can be measures of aptitudes – capacity or ability to learn and measures of achievement or attainment. These can be done using aptitude tests and achievement tests. The major concentration of the class teacher is achievement test which he is expected to use to promote learning and bring about purposeful and desirable changes in the students entrusted to him

3.1.5 To Determine The Strength And Weakness Of The Teaching Methods

The results of classroom tests provide empirical evidence for the teacher to know how well or how effective his teaching methods are. Test results are used as self –evaluation instrument. They can be used by others to evaluate the teacher. If the results are not encouraging, the teacher may decide to review his teaching methods with a view to modifying or changing to another.

3.1.6 To Determine The Adequacy Or Otherwise Of Instructional Materials

A good teacher makes use of a variety of teaching aids for illustrations and demonstrations. Effective use of these instructional resources helps to improve students understanding of the lesson. Topics which look abstract can be brought to concrete terms by the use of these materials. Therefore to determine the effectiveness, adequacy or otherwise of these teaching aids test can be used.

3.1.7 To Determine The Extent Of Achievement Of The Objectives

There are goals and objectives set for the schools. Every school is expected to achieve the goals and objectives through the instructional programmes. The results of tests given to students are used to evaluate how well the instructional programmes have helped in the achievement of the goals and objectives.

SELF ASSESSMENT EXERCISE 1

List five functions of tests in education.

SELF-ASSESSMENT EXERCISE 1

Your answers would have included the following:

- To motivate the pupils to study.
- To determine how much the pupils have learned.
- To determine the pupils' special difficulties.
- To determine the pupils' special abilities.
- To determine the strengths and weaknesses of the teaching methods.
- To determine the adequacy or otherwise of instructional materials.
- To determine the extent of achievement of the objectives.

3.2 Measurement Scales

Measurement exists in several levels depending on what is to be measured, the instrument to be employed, the degree of accuracy or precision desired and the method of measurement. There are four levels of measuring scales. These are nominal, ordinal, interval and ratio scales.

3.2.1 Nominal Scale

Some of the times, you are involved in the classification of objects, things, human beings, animals, etc. For instance, you can classify human beings as males and females, you can also classify living things as plants and animals.

Have you ever assigned your students into classes or groups? Nominal scale which is the simplest of the measurement scales involves only the assignment to classes or groups and does not imply magnitude. It is used when we are interested in knowing if certain objects belong to the same or different classes. We may decide to assign teachers in the school system into two groups of graduate teachers = I and non-graduate teacher = 0. You will note that these numbers are just codes, as they do not indicate magnitude. We may group students into classes A, B, C D, etc.

These are also codes. There is no order involved here, there is nothing like greater than or less than. Any letter, number or numerals used in a nominal scale have no quantitative significance; they are used for convenience only. Nominal scale does not lend itself to some of the useful arithmetical and statistical operations such as addition, subtraction, multiplication and division. Do you think that this is a shortcoming or an advantage?

3.2.2 Ordinal Scale

In the last section (i.e. 3.2.1) you learnt about the nominal scale, which involves simple classification of objects, things, people or events. The next in simplicity is the ordinal scale. In this case there is order. One real number may be greater than or equal to or less than another real number.

For instance, the number 5 is greater than the number 3. There is classification, as well as indication of size and rank. In your class, you rank your students using their test result from first, second, third to nth positions.

This is ranking. It is the order. It is the ordinal scale. You will note that what is important here is the position of the individuals or things or events in a group. The ranks cannot be compared. For instance, you can not say that a student who is 2^{nd} in a test is twice as good as the student who is ranked 4^{th} and half as good as the student ranks 1t, in spite of the fact that 2 is half of 4 and 1 is half of 2. let us illustrate this with an example. Take the scores of four students ranked 5^{th}, 6^{th}, 7^{th} and 8^{th} as 70%, 68%, 50%, and 49% respectively. You note that the difference between 5^{th} and 6^{th} positions is only 2%, while between 6^{th} and 7^{th} positions, we have, not 2% but 18%.

And the difference between 7^{th} and 8^{th} positions is just 1%. It means therefore that equal intervals on the scale do not represent equal quantities. Therefore, only limited arithmetical operations can be used on the ordinal scale.

SELF-ASSESSMENT EXERCISE 1

What is the major difference between nominal scale and ordinal scale of measurement?

SELF-ASSESSMENT EXERCISE 1

You may have answered that nominal scale involves only classification of objects, persons, things, events etc into groups, while ordinal scale involves both classification and order or ranks of objects, persons, things, events etc.

3.2.2 Interval Scale

This is the next scale after ordinal scale. You remember that we said that in the ordinal scale equal intervals do not represent equal quantities. But the reverse is the case in the interval scale. In other words equal intervals represent equal quantities here. The amount of difference

between adjacent intervals on the scale is equal. Take the calendar as an example, days on it, represent equal amounts of time. You will observe that equal amount of time separates 2, 4, 6, and 8 days.

In an examination, a student who scored 80% is 10% greater than the student who scored 70% and is 10% less than one who scored 90%. Because the data here are continuous, some arithmetic operations like addition, subtraction, can take place here. For instance, if you say John's height is 1.56m, Olu's height is 1.66m and Ibe's height is 1.70m. it follows that Ibe is the tallest. He is 0.04m taller than Olu and 0.14m taller than John. It implies that you can rank as well as get the differences between the intervals. But there is no absolute zero in the interval scale.

3.2.4 Ratio

This is the highest level or scale of measurement. It has all the characteristics of the others. In addition there is an absolute zero here. Most of the arithmetic operations like addition, subtraction, multiplication and division are involved in this scale. Let us use the meter rule as an example of ratio scale. This scale has equal intervals as in the intervals scale. It has a zero mark, which implies that at this point there is complete absence of what the meter rule measures. If it is used to measure height, it means that at zero, there is no height at all. This is not the case with the interval. If you give a test in your class and a student scores zero, does it imply that the student does not know anything at all? Take the case of the calendar again as an example. Look at the calendars you have from different sources and places. Is there any one with zero point? Can there be any zero day or time?

Let us use another example to drive this point home. If you use a weighing balance to measure the weights of objects, you will discover that at zero mark there is a complete lack of weight. There is no weight at all. But if you use a test to measure intelligence of students, any student who scores zero does not imply he has no intelligence at all. The ratio scale is appropriately used in the physical sciences, while the interval scale is used in education, social science and psychology.

The table below illustrates the four scale levels.

S/N	Scale level	Illustration	Possible Arithmetic Operation
1.	Nominal	Numbers are used only as names or groups or classes e.g. female = 0, Male = 1,	Counting only.

		Adults = A, Youths = B, Children = C, Doctors = 1, Teachers = 2, Lawyers = 3	
2.	Ordinal	Numbers are used as rank order. E.g. 1^{st}, 2^{nd}, 3^{rd} etc Mary = 1^{st}, Joy = 2^{nd}	Counting, Ranking method greater than, less than.
3.	Interval	Intervals between any two numbers are equal both in interval and quantities. E.g. Scores, calendar etc.	Counting, Ranking, Addition, and subtraction.
4.	Ratio	Each number is a distance from zero. There is absolute zero. E.g. ruler, weighing balance, speedometer etc	Counting, Ranking, Addition, Subtraction, Multiplication and division.

SELF-ASSESSMENT EXERCISE 4

i. List the four types of measurement scales?
ii. What are the major difference between interval scale and Ratio?

SELF ASSESSMENT EXERCISE 1

i. Nominal, ordinal, interval and Ratio scale.
ii. The main differences between intervals scale and Ratio scale are
(a) there is no absolute zero in interval scale where as there is absolute zero in the ratio scale.
(b) In the ratio scale arithmetic operations like addition, subtraction, multiplication and division are involved while in the interval scale, we have only addition and subtraction.

4.0 CONCLUSION

From what you have learnt so far in this unit, you should be able to deduce that good measurement is no accident. It comes about as a result of conscious effort on the part of those who are willing to play the game according to the rules. There is no substitute for insight and good judgment.

An examination will produce results usually in the form of marks but it is human beings who make the decisions. This implies that the data we collect must be from defined scale of measurement so as to make analyses and interpretation very easy and valid.

5.0 SUMMARY

In this unit we have looked at the major functions of test in education. These include * motivation of students to learn

- Determination to how much the student have learned
- Determination of the students' special difficulties
- Determination of students' special abilities
- Determination of the strength and weaknesses of teaching methods.
- Determination of the adequacy or otherwise of instructional materials.
- Determination of the extent of achievement of the objectives etc

We also looked at the four scales of measurement. These are: Nominal scale which involves, simple classification, ordinal scale which involves rank order, interval scale which involves equal intervals and equal quantities but without absolute zero, and the ratio scale which has absolute zero and is especially used in the physical sciences.

6.0 TUTOR-MARKED ASSIGNMENT

1. Mention the four levels of measurement, their main features and their possible arithmetic operations.

7.0 REFERENCES/FURTHER READING

Nwana, O.C. (1981) Educational Measurement for Teachers Ikeja: Thomas Nelson Africa.

Onuche, R.O. and Akeju S.A. (1977) Testing and Evaluation in Education, Lagos, Monrovia Owerri African Educational Resources (AER).

MODULE 2 EDUCATIONAL OBJECTIVES

UNIT 1 EDUCATIONAL OBJECTIVES

CONTENTS

1.0 INTRODUCTION

Objectives generally indicate the end points of a journey. They specify where you want to be or what you intend to achieve at the end of a process. An educational objective is that achievement which a specific educational instruction is expected to make or accomplish. It is the outcome of any educational instruction. It is the purpose for which any particular educational undertaking is carried out. It is the goal of any educational task. In this unit we shall look at the difference role educational objectives assume at different settings.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- list the different levels of educational objectives
- give the differences between aims and objectives
- write specific objectives at the instructional level
- explain the importance of feedback in instructional situation.

3.0 MAIN CONTENT

3.1 Various Level of Educational objectives

Educational objectives can be specified at various levels.
These levels include the national level, the institutional level and the instructional level. In this unit we shall look at the various levels.

3.2 At the National Level

At this level of educational objectives, we have merely policy statements of what education should achieve for the nation. They are in broad outlines reflecting national interests, values, aspirations and goals. The objectives are general and somewhat vague. At this level, they may be interpreted. It can be in the form of the National Policy on Education. Have you seen the Federal Republic of Nigeria, National Policy on Education?

The latest edition is the 2004 edition and another edition will soon be released. The national educational aims and objectives are stated, you will try to get a copy for your use. The goals of education include:

i. To motivate national consciousness and national unity.
ii. To inculcate the right type of values and attitudes for the survival of the individual and the Nigerian society.
iii. To train the mind in the understanding of the world around.
iv. To acquire the appropriate skills, abilities and competences both mental and physical as equipment for the individual to live in and contribute to the development of the society.

Apart from these goals of education in Nigeria, the National Policy also specifies objectives for early childhood education, primary education, secondary education, adult, technical, teacher, special education, university and other forms of education.

These educational goals can also be specified in the National development plan. Let us look at one of such goals specified in the Nigerian National development plan. It says that the development of (i) a strong and self-reliant nation (ii) a great and dynamic economy (iii) a just and egalitarian society (iv) a land of bright and full opportunity for all citizens and (v) a free and democratic society. You will recall that we have earlier said that these are called goals or aims. They are broad and too vague to give focused direction to curriculum development.

SELF-ASSESSMENT EXERCISE 1

Get a copy of the National Policy on Education and read the goals or aims of secondary education, primary education, Technical education and Pre-Primary education.

3.3 At the Institutional level

This is the intermediate objectives level. The aims are logically derived and related to both the ones at the national level and the one's at the instructional levels. What are the objectives of your university? By the time you look at the educational objectives of three or four institutions, you would have noticed that the educational objectives at this institution have been established. They are narrowed to achieve local needs like the kinds of certificate to be awarded by the institutions. These institutional objectives are usually specified by an act or edict of the house of assembly if it is a state government institution, otherwise by an act of the national assembly.

The National Open University of Nigeria (NOUN) was established by an act of the parliament or National assembly. You can read it again in your handbook (know your university). These objectives are often in line with those stipulated for that kind of institution at the national level. These objectives are not specific. They are also broad aims. For instance, an objective for a college of education can be to produce intermediate level teachers to teach in the Pre-primary, primary and secondary schools. What do you think would be the objectives for a monotechnic which awards diploma in Agriculture?

3.4 At the Instructional Level

You have seen that objectives specified at both the national and institutional level are all broad goals and aims. These can be realized in its and pieces at the instructional level. Here, educational objectives are stated in the form in which they are to operate in the classroom.

They are therefore referred to as instructional objectives, behavioural objectives or learning outcomes. They are specified based on the intended learning outcomes. These objectives state what teaching is expected to achieve, what the learner is expected to learn from the instruction, how the learner is expected to behave after being subjected to the instruction and what he has to do in order to demonstrate that he has learnt what is expected from the instruction. These instructional objectives are therefore stated in behavioural terms with the use of action verbs to specify the desirable behaviour which the learner will exhibit in order to show that he has learnt. Now, look at the objective at

the section 2.0 of this unit or any other unit or any other course material. How are the objectives specified?

They are example of the instructional objectives. They are learner-centered not teacher centered. They can easily be assessed on observed. They are specified according to each unity, lesson etc.

SELF-ASSESSMENT EXERCISE 2

Write 5 instructional objectives to any course unit of your choice using action verbs**.**

Now look at the following objectives

At the end of this lesson or unit you should be able to:

i. Define the mean
ii. Calculate the mean in a given distribution of scores
iii. Calculate the mean using the assumed mean (AM) method.
iv. Explain the median of a set of scores.
v. Explain the procedure for finding the median in a set of a score
vi. Calculate the median in a grouped set of scores.

Now look at the action verbs, define, explain, calculate, specify, construct, design, state, mention, list, draw, choose, find, prepare, paint, apply, analyze, deduce, differentiate etc.

These can be easily evaluated. These are the type of verbs you will be using when you specify your objectives. You will learn more about them in other courses in education.

3.5 Importance of Instructional Objectives.

So far .you can see that instructional objectives are very important component of teaching system. Let us mention some of the importance especially as a feedback mechanism in the education system. Learning outcomes as displayed by the learners serve as feedback on how much the instructional objectives have been achieved. They also show how appropriate the curriculum of the institution is. These instructional objectives can be used as a feedback on how much the institutional objectives have been achieved and how appropriate these objectives are. You remember that objectives start from broad goals at the national level to the instructional level. In the same way when evaluating these objectives, we use the instructional level objectives.

From this, evaluation goes to the institutional to the national levels. In other words, the feedback got from the assessment of the instructional objectives is translated into finding and how much the national educational objectives have been achieved in respect to the particular type of institution, and their appropriateness. In the final analysis, the findings may lead to revising the objectives of any level or all the levels. They may lead to curriculum modification at the institution level. At the instructional level, they may lead to the adjustment of teaching methods or provision of instructional materials. You see, from the small things, activities, tests, examinations, projects, assignments, exercises, quizzes, home works etc done in the classroom setting, we can use to evaluate, in a general process, the national policy at the national level.

Apart from the feedback instructional objectives are also important because the teacher's plans of what to teach and how to teach it is based on the objectives specified to be achieved. The evaluation of pupils' learning outcome will make him know whether the objectives are being achieved or not. It means therefore that the instructional objectives give meaning and direction to the educational process.

SELF-ASSESSMENT EXERCISE 3

List the importance of instructional objectives in the education system.

SELF ASSESSMENT EXERCISE 3

Your answers may have included

i. They serve as feedback on the appropriateness of the curriculum.
ii. They serve to find how much the instructional objectives are achieved and their appropriateness.
iii. They serve to find how much or the extent of achievement of the national objectives.
iv. The recruit of the evaluation of the objectives can be used to modify the curriculum, teaching methods and instructional materials.

4.0 CONCLUSION

A test is designed to measure the level of the attainment of the specified instructional objectives. The intention of an achievement test is to translate well defined subject matter content into test questions that will elicit the types of behaviour we intend the learners to develop. Such behaviours are spelt out or implied in our course or instructional objectives. These objectives are operational translations of the goals of education and they provide guides to ensure desirable changes in the

learners behaviour in a given subject mater area. The realization of these objectives will contribute towards the achievement of the ultimate goal.

5.0 SUMMARY

In this unit you learned the various levels of objectives. These are the national level, the institutional level and the instructional level. We have said that objectives at the national level are very broad and vague. They are referred to as goals, aims or policy such as the National Policy on Education (NPE). You also learned that the objectives at the institutional level are also broad.

But the objectives at the instructional level or classroom level are specific and stated in behavioural terms, specifying what the learners are expected to do at the end of the lesson, using action verbs. Objectives at the national level can only be realized in bits and pieces through the contributory influence of the many formal and informal learning experiences at school, home and in the society.

For the nation to achieve her goal, they must be realized after formal education has been concluded.

Since the schools are institutions consciously created to ensure desirable changes in human behaviour towards the ultimate realization of the national goals, they have to make conscious efforts to ensure the attainment of the goals. This can be done through a systematic translation of these objectives, and then to instructional objectives. These should be stated in a way that they should be observed or measured.

Objectives are very important as they provide the necessary feedback for the adjustments of curriculum, teaching method and teaching aids among others.

6.0 TUTOR-MARKED ASSIGNMENT (TMA)

1. What are the 3 main levels of objectives in education?
2. What is a major difference between objectives and aim?
3. Give an importance of feedback in educational instruction?

7.0 REFERENCES/FURTHER READING

Nenty, H.J. (1985) Fundamental of Measurement and Evaluation in Education. Calabar: University of Calabar Press.

Obimba, F.U. (1989) Fundamentals of Measurement and Evaluation in Education and Psychology**.** Owerri: Totan publishers Ltd.

UNIT 2 BLOOM'S TAXONOMY OF EDUCATIONAL OBJECTIVES

CONTENTS

1.0 INTRODUCTION

Bloom' Benjamin's has put forward a taxonomy of educational objectives, which provides a practical framework within which educational objectives could be organized and measured. In this taxonomy Bloom et al (1956) divided educational objectives into three domains. These are cognitive domain, Affective domain and psychomotor domain.

In this unit, we shall look at the cognitive domain in some details, later in other units we shall look at the others.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- draw the hierarchical diagram of the Boom's cognitive domain.
- explain the level of cognitive domain.
- specify objectives corresponding to each of the domains.

3.0 MAIN CONTENT

3.1 Cognitive domain

The cognitive domain involves those objectives that deal with the development of intellectual abilities and skills. These have to do with

the mental abilities of the brain. The domain is categorized into six hierarchical comprehension, application, analysis, synthesis and evaluation. These levels are of hierarchical and increasing operational difficulties that achievement of a higher level of skill assumes the achievement of the previous levels. This implies that a higher level of skill could be achieved only if a certain amount of ability called for by the previous level has been achieved. For instance, you cannot apply what you do not know or comprehend, can you now understand what it means to be hierarchical.

Now let us look at the components of the cognitive domain.

EVALUATION
SYNTHESIS
ANALYSIS
APPLICATION
COMPREHENSION
KNOWLEDGE (MEMORY)

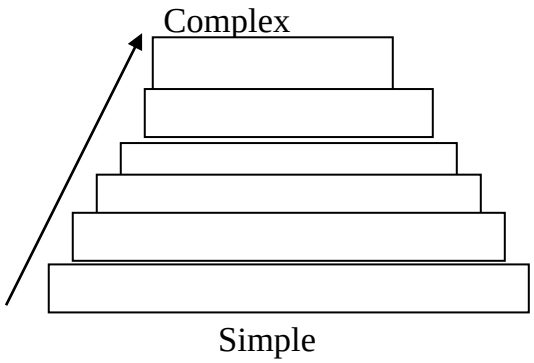


Fig.5.1. Hierarchical levels of Bloom's taxonomy

3.2 KNOWLEDGE (or Memory)

If you have studied the figure 5.1 above, you would have noticed that knowledge or memory is the first, the lowest and the foundation for the development of higher order cognitive skills. It involves the recognition or recall of previous learned information. There is no demand for understanding or internalization of information. For measurement purposes, memory or knowledge involves bringing to mind the appropriate material. This cognitive level emphasizes the psychological process of remembering. Action verbs which can serve as appropriate signals, cues and clues that can effectively bring out stored information in the mind include: define, name, list, tell, recall, identify, remember, who, which, where, when, what, recognize, how much, how many etc.
You can now use these verbs to formulate instructional objectives.

SELF-ASSESSMENT EXERCISE 1

Using these action verbs, specify some objectives in your subject area.

Memory can also be classified into the following:

i. Memory of specifics (taxonomy, facts, definitions etc)
ii. Memory of ways and means of dealing with specifics (rules)
iii. Memory of conventions (styles, symbols practices, allegories)
iv. Memory of Trends and sequences (order or sequence)
v. Memory of classification and categories (classes, sets, divisions)
vi. Memory of criteria (established facts and criteria)
vii. Memory of Techniques and procedures or Methodology.
viii. Memory of universals and abstractions.
ix. Memory of principles and generalizations (laws, formulas)
x. Memory of theories and structures (models, philosophies)

3.3 Comprehension Level

You remember that we said that memory is concerned with the accumulation of a repertoire of facts, specifics, ways and means of dealing with specifics, universals, abstractions etc. it implies that memory involvers verbalization and role learning. Comprehension is all about internalization of knowledge. It involvers making memory out of what is stored in the brain file. It is on this basis that what is stored in the brain can be understood and translated, interpreted or extrapolated. It is only when you have known something that you can understand it. Again it is only when you know and understand that you can re-order or re-arrange. Actions verbs here include explain, represent, restate, convert, interpret, re-arrange, re-order, translate, rephrase, transform etc.

Comprehension level is made up of the following:

i. Translation: which involves the ability to understand literal messages across communication forms, changing what is known from one form of communication to another e.g. from words to numbers, graphs, maps, charts, cartoons, pictures, formulas, symbols, models, equations etc.
ii. Interpretation: which goes beyond mere literal translation to identification of inter-relationships among parts and components of communication and interpreting and relating these to the main components e.g. to interpret a chart or graph etc.
iii. Extrapolation: which involves the ability to draw implications and ability to identify and continue a trend, isolate or detect consequences, suggest possible meaning and estimate possible effect.

3.4 Application Level

In the last section we noted that you can not understand what you have not known. It also means that you cannot apply what you do not understand. The use of abstractions in a concrete situation is called application. These abstractions can be in the form of general ideas, rules, or procedures or generalized methods, technical terms, principles, ideas and theories which must be remembered, understood and applied.

You must understand before correct application. Ability to apply what is learned is an indication of a more permanent acquisition of learning. Application skills are developed when the learner uses what he knows to solve a new problem, or in a new situation. Application involves the ability to the learner to grasp exactly what the problem is all about and what generalization or principles are relevant, useful, or pertinent for its solution. Some action verbs here include: apply, build, explain, calculate, classify, solve, specify, state, transfer demonstrate, determine, design, employ, predict, present, use which, restructure, relate, employ, organize etc. it involves the principles of transfer of learning.

SELF-ASSESSMENT EXERCISE 2

i. What are the components of comprehension level?
ii. What is application?

SELF-ASSESSMENT EXERCISE 2

You may have given these answers.

1. The components of comprehension level are:-
   Translation, interpretation and Extrapolation
2. Application is the use of abstractions in a concrete situation. It is the ability to apply what is learnt to solve a new problem or in a new situation. It is all about transfer of learning.

3.5 Analysis Level

This is the breaking down of communication into its constituent parts or elements in order to establish the relationship or make the relations between ideas expressed to be clear or explicit. It means breaking a learnt material into parts, ideas and devices for clearer understanding.

It goes beyond application and involves such action verbs as analyse, detect, determine, establish, compare, why, discriminate, distinguish, check consistency, categories, establish evidence etc.

The components here include:

i. Analysis of Elements: which is concerned with the ability to identify the underlying elements such as assumptions, hypothesis, conclusions, views, values, arguments, statements etc and to determine the nature and functions of such elements?
ii. Analysis of Relationship: which involves trying to determine how the elements identified are related to each other? For instance, how does the evidence relate to the conclusion?
iii. Analysis of Organizational principles: which involves determining the principles or system of organization which holds the different elements and parts together? It involves finding the pattern, the structure, systematic arrangements, point of view, etc.

3.6 Synthesis Level

In sub-section 3.4, you learnt that analysis involves breaking down of materials, communication, object etc. but in synthesis you build up or put together elements. Parts, pieces and components in order to form a unique whole or to constitute a new form, plan, pattern or structure. In other words, synthesis is concerned with the ability to put parts of knowledge together to form a new knowledge. It involves categorizing of items, composing of poems, and songs, writing etc. it involves divergent thinking. It calls for imaginative, original and creative thinking. You will note that creative – though process results in discovery of knowledge that is new or something that is tangible. It calls for creative answers to problems and for the development of questioning mind, spirit of inquiry or inquisitive mind.

It requires fluency of novel ideas and flexible mind. It allows students great freedom at looking for solutions, using many possible approaches to problem solving. Action verbs includes: plan, develop, devise, write, tell, make, assemble, classify, express, illustrate, produce, propose, specify, suggest, document, formulate, modify, organize, derive, design, derive, create, combine, construct, put together, constituted, etc synthetic can be sub divided into:

(a) Production of unique communication: which is concerned with the ability to put together in a unique organizational form a piece of written or oral communication to convey a novel idea, feeling or experience to others?
(b) Production of a plan or proposed set of operations: this is concerned with the ability to develop a plan or to propose procedures for solving problem or dealing with others.

(c) Derivation of a set of Abstract Relation: this is based on the result of the analysis of an experimental data, observation or other specific. It is the ability to form concepts generalizations, deduce propositions, predictions or relationship based on classification of experiences or observations.

SELF-ASSESSMENT EXERCISE 3

What are the sub-divisions of analysis and synthesis?

SELF ASSESSMENT EXERCISE 1

Your answers may have included:

Analysis is sub-divided into

(i) Analysis of elements
(ii) Analysis of Relationships
(iii) Analysis of organizational principles.

Synthesis is sub-divided into

(i) Production of unique communication
(ii) Production of plan or proposed set of operations
(iii) Derivation of a set of Abstract Relations.

3.7 Evaluation Level

Your knowledge of the meaning of evaluation is not very different from this level of cognitive domain. It is the highest in the hierarchy. It involves making a quantitative or qualitative judgment about a piece of communication, a procedure, a method, a proposal, a plan etc. Based on certain internal or external criteria alternatives abound, choice depends on the result of judgment which we make consciously or unconsciously based on values we held. Every day, we make judgments such as good or bad, right or wrong, agree or disagree, fast or slow etc.

These are simple evaluations. They may not base on logical or rational judgment. In education, evaluation as a cognitive objective involves the learners' ability to organize his thought and knowledge to reach a logical and rational decision which is defendable.

Evaluation is the most complex of human cognitive behaviour. It embodies elements of the other five categories. What are the other categories? Can you name them? They are knowledge, comprehension, application, analysis and synthesis. Action verbs here include: agree,

assess, compare, appraise, choose, evaluate, why, validate, judge, select, conclude, consider, decide, contract etc. evaluation can be subdivided into (a) judgment in terms of internal criteria and (b) judgment in terms of external criteria.

4.0 CONCLUSION

The taxonomy of educational objectives from the work or Professor B.S. Bloom and his colleagues has given us a sound base for the formulation of our objectives. A sound system of education should be able to produce graduates who have a wealth of knowledge and are able to comprehend much of it, and are also able to apply their knowledge and engage in those mental activities variously referred to as critical thinking, reflective thinking, divergent thinking, inductive-deductive processes, problem solving etc. these activities are concerned with the higher categories of the cognitive domain. These areas should form the bulk of the objectives because they are the ones from where we expect the highest educational dividends.

5.0 SUMMARY

In this unit, you have learnt that bloom classified educational objectives of an intellectual nature or the cognitive domain into six groups, which form a hierarchy of mental skills from the lowest and easiest level, knowledge or memory to the highest and most difficult level, evaluation. Knowledge and comprehension are regarded as the low cognitive objectives, while application, analysis, synthesis and evaluation are regarded as the higher cognitive objectives.

You have seen that objectives under knowledge require simple straight recall or recognition of facts and principles based on the accuracy of the memory. Comprehension implies a thorough or reasonable familiarity with principles and facts and their recognition when stated in different words or when they appear in different circumstances. The learner should be able to give explanations to them. Application implies transfer of learning. Learners should be able to deduce results, underlying principles and environmental conditions given the effects on the system. Analysis involves learning apart, breaking into parts etc in order to see the relationship, while synthesis is about building, constructing, assembling, etc. It involves creativity, divergent thinking, critical thinking etc. but evaluation which is the highest in the hierarchy is the ability to judge the value of subject-matter that is learnt in relation to a specific purpose. In the next unit we shall look at the other aspects of the educational objectives – Affective and Psychomotor.

6.0 TUTOR-MARKED ASSIGNMENT

1. What are the six categories of educational objectives in kothe cognitive domain of Bloom's taxonomy?
2. With examples of action verbs, explain evaluation?

7.0 REFERENCES/FURTHER READING

Nenty, H.J (1985). Fundamental of Measurement and Evaluation in Education, Calabar: University of Calabar.

Obimba, F.U. (1989). Fundamentals of Measurement and Evaluation in Education and Psychology. Owerri: Totan Publishers Ltd.

Nwana. O.C. (1981) Educational Measurement For Teacher. Ikeja – Lagos, Hong Kong, Ontario, Kenya, Thomas Nelson Africa

UNIT 3 AFFECTIVE DOMAIN

CONTENTS

1.0 INTRODUCTION

In the past, people were not very happy about emotionalism in education. They argue that intellectualism had little or nothing to do with the learner's interests, emotions or impulses.

Today, people have recognized that the learner's feelings and emotions are important considerations in education. This is why a group of interests led by Tanner and Tanner (1975) insist that the primary goals of learning are affective. They are of the opinion that learners should not learn what is selected for them by others. This is because it amounts to imposition on the learners of other people's values and purposes. This of course defys learners' own feelings and emotions. You can see that this argument is in contrast to what happens in our schools today where most schools hold that fundamental objectives are cognitive.

As a matter of fact, what we have in our school systems is the discipline-centred curriculum projects which focus on the cognitive learning to the neglect of affective processes. Although the primary goal of a good teacher is to help students learn, not to make them feel good yet it is an important role of a good teacher to make students feel good about their efforts to learn and their success in learning. This will help to create a balance and interdependence between the cognitive and the affective processes of learning. In the last unit we shall be concerned with the affective domain and its characteristics. Before we do that, let us look at what you will achieve at the end of the unit.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain the meaning of affective domain.
- describe the levels of affective domain.
- state objectives in the affective domain.
- mention the characteristic features of affective domain.

3.0 MAIN CONTENT

3.1 Characteristic Features of Affective Domain:

While you were reading the information in this unit, you learnt that affective domain has to do with feelings and emotions. These are the emphatic characteristic of this domain of acceptance or rejection. It is concerned with interests, attitudes, appreciation, emotional biases and values. The function of the affective domain in the instructional situation pertains to emotions, the passions, the dispositions, the moral and the aesthetic sensibilities, the capacity for feeling, concern, attachment or detachment, sympathy, empathy, and appreciation.

Can you now see that your feeling, emotion, appreciation, the value you place on this course, together form your affective disposition of the course. It shows your personal-social adjustment in this course and indeed in the programme. Basically you, as a learner, have internalized and appreciated /what you have been taught or what you have learnt are demonstrated in attitudes, likes and dislikes etc. Affective domain is generally covert in behaviour. The educational objectives here vary from simple attention to complex and internally consistent qualities of character and conscience.

Examples of learning outcomes in the affective domain are:

- The learner will be able to appreciate the use of frawing instruments in the construction of objects in Technical drawing.
- The learner will be able to show awareness of the rules and regulations in a technical workshop to prevent accidents.
- The learner should be able to show his likeness for neatness and accuracy in the use of measurement instruments etc.

Affective domain has five hierarchical categories. You remember that the cognitive domain has six hierarchical levels.

Specifically, the levels in affective domain fall into these levels: receiving, responding, valuing, organization and characterization

3.1 Receiving

This is the lowest level of the learning outcomes in the affective domain. It means attending. It is the learner's willingness to attend to a particular stimulus or his being sensitive to the existence of a given problem, event, condition or situation. It has three sub-levels.

These are:

i. Awareness: which involves the conscious recognition of the existence of some problems, conditions, situations, events, phenomena etc. take for instance as a teacher, you come into your class while the students are making noise. You will notice that the atmosphere will change. This is because the students have become aware of your presence. They are merely aware.
ii. Willingness: This is the next stage which involves the ability to acknowledge the object, event, problem instead of ignoring or avoiding it. The students in your class kept quite because they noticed and acknowledged your presence. If they had ignored your presence they would continue to make noise in the class.
iii. Controlled or selected attention: This involves the learner selecting or choosing to pay attention to the situation, problem, event or phenomenon. When you teach in the class, the learner is aware of your saying or the points you are making. In that case he will deliberately shut off messages or speeches or sounds as noises. Receiving in a classroom situation involves getting, holding and directing the attention of the learners to whatever the teacher has to say in the class.

3.2 Responding

In this case the learner responds to the event by participating. He does not only attend, he also reacts by doing something. If in your class you set a test for your students, first the students become aware of the test, they are willing to take the test, they now select to do it and they react by doing it. Responding has three sub-levels too. These are:

i. acquiescence in responding: which involves simple obedience or compliance.
ii. Willingness to respond: This involves voluntary responses to a given situation.
iii. Satisfaction in response: if he is satisfied with the response he enjoys reacting to the type of situation.
If in the school situation you give a project to your class, they comply by doing the project very well. They are satisfied with

what they have been able to produce. They will be happy and would wish to have that type of project again and again.
That shows that their interest is now awakened.

SELF-ASSESSMENT EXERCISE

1. What are the sub sections of Receiving and Responding?
2. Using your experience as a teacher, give an example of an event related to receiving or responding.

SELF-ASSESSMENT EXERCISE

Your answers may have included:

i. (a) Receiving = Awareness, willingness and controlled or selected attention
ii. (b) Responding = acquiescence in responding, willingness to respond and satisfaction in response.
iii. Give an example of an event you have experience.

3.3 Valuing

This is concerned with the worth or value or benefit which a leaner attaches to a particular object, behaviour or situation.

This ranges in degree from mere acceptance of value or a desire to improve group skills to a more complex level of commitment or an assumption of responsibility for the effective functioning of the group. As usual, there are three sub-levels of valuing:

i. Acceptance of a value: This is a situation where the learner believes tentatively in a proportion, doctrine, condition or situation a denomination says that women should be ordained priest in the e.g churches; the members accept the doctrine.
ii. Preference for a value: In this case the learner believes in the desirability or necessity of the condition, doctrine, proposition etc. and ignores or rejects other alternatives and deliberately looks for other people views where the issues are controversial, so as to form his own opinion.
iii. Commitment to a value: In this stage the learner is convinced and fully committed to the doctrine, principle or cause. In consequence, the learner internalizes a set of specific values, which consistently manifest themselves in his event behaviour, attitudes and appreciation.

Now let us go to the next level of affective domain.

3.2 Organization

In this level the learner starts to bring together different values as an organized system. He determines the interrelationships and establishes the order of priority by comparing, relating and synthesizing the values. He then builds a consistent value system by resolving any possible conflicts between them. If the learner tries to successfully internalize the value, he may encounter some situations which may demand more than one value. In this case, he has to organize the values into a system in order to decide which value to emphasis.

There are two sub-levels of organization. These are:

i. Conceptualization of a Value

This involves the understanding of the relationship of abstract elements of a value to these already held or to new values which are gaining acceptance. In other words you may have to evaluate the works of arts which you appreciate or to find out in order to clarify some basic assumptions about codes of ethnics. It may be in the area of music where you may have to identify the characteristics of two types of music such as classical and hip up music, which you admire or enjoy in relation to the others such as jazz or highlife which you do not like. We have used works of arts and music as examples. Can you think of any other examples? Think of different types of vehicles, songs, colours, designs, prints etc.

ii. Organization of Value System?

This involves the development of a complex value system, which includes those values that cannot be compared for the purpose of making choices in order to promote public welfare, instead of the sheer aggrandizement of special personal interest. For instance, you may be in a position to relate your interest in the works of arts against other value.

You may be in situation where you compare alternative social policies and practices against the standards of public welfare. It is this level that leads individuals to develop vocational plan which can satisfy their needs for economic security and social welfare. It leads the individual to develop philosophy of life, which helps him to avoid dependence upon others, especially to avoid a situation where one becomes a public nuisance. You can see that this level is a very important one.

SELF-ASSESSMENT EXERCISE

Complete the statements below.

i. The situation where a learner believes tentatively in a doctrine, proposition of condition is called
ii. The situation in which the learner believes in the necessity of the doctrine, proposition or situation is called ……..
iii. The situation in which the learner is convinced and fully committed to the doctrine, proposition or situation is called
iv. The total sub-level of organization are and

SELF-ASSESSMENT EXERCISE

You may have given these answers:

i. Acceptance of value
ii. Preference to a value
iii. Commitment to a value
iv. Conceptualization of a value and organization of value system.

3.5 Characterization by a Value or a Value Complex

At this stage the value system is so internalized by the people of individuals so that they act consistently in accordance with such values, beliefs or ideals that comprise their total philosophy or view of life. A life-style which reflects these beliefs and philosophy are developed. The behaviour of such individuals or groups can be said to be controlled by the value system.

This is true, as one can almost predict with accuracy how an individual would behave or respond. There are two levels here:

i. Generalized set: This involves a situation where the orientation of the individual enables him to reduce to order a complex environment and to act consistently and effectively in it. There may be room for the individual to revise his judgements and to change his behaviour as a result of available new and valid evidence.
ii. Characterization: In this case, the internalization of a value system is such that the individual is consistently acting in harmony with it.

The value system regulates the individual's personal and civil life according to a code of behaviour based on ethical principles.

You now notice that the level of value complex **individual develop typical or particular** behaviour. Instructional objectives here should be concerned with the learners' general pattern of personal, social, or emotional adjustment.

4.0 CONCLUSION

From this discussion you can conclude that cognitive and affective Processes are interrelated. The way people think can be said to be Virtually directed by an organic interaction of cognitive and affective Processes.

Affective processes are inseparable from the cognitive processes of Learning. Good thinking or mental ability is dependent on the attitudes, Values and motivations that animate one's character. You have also seen that the affective domain is very important and can be neglected in the Formulation of instructional objectives especially as it relates to character formation and internalization of values.

5.0 SUMMARY

In the unit you have gone through the affective domain where you learnt its characteristic features as something to do with feelings, emotions, degree of acceptance or rejection, interests, attitudes, appreciation, emotional biases and value etc.

You also learnt that the affective domain, like the cognitive domain has hierarchical categories. But while the cognitive domain has six levels, the affective domain has five levels.

These levels are:

(1) Receiving which is the lowest and has three sub-levels of awareness, willingness and controlled or selected attention.
(2) Responding: which also has three sub-levels of acquiescence in responding, willingness to respond and satisfaction in response.
(3) Valuing: which has also three sub-levels of acceptance of a value, preference for a value and commitment to a value.
(4) Organization: This has two sub-levels of conceptualization of a value and organization of value system.
(5) Characterization by a value or value complex: This has two sub-categories of general set and characterization.

You also learnt that the cognitive and affective processes are interrelated.

6.0 TUTOR-MARKED ASSIGNMENT

1. Mention 5 characteristic features of the affective domain.
2. What are the 5 sub-levels of affective domain.
3. Explain the levels, characterization by a value or value complex.

7.0 REFERENCES/FURTHER READING

Obimba., F.U. (1989) Fundamental of Measurement and Evaluation in Education and Psychology. Owerri: Totan Pub. Ltd.

Onwuka, U. (1990) Curriculum Development for Africa.Onitsha: Africana – FEP Publishers Ltd.

UNIT 4 PSYCHOMOTOR DOMAIN

CONTENTS

1.0 INTRODUCTION

In the last two units you have worked through our discussions of the cognitive and affective domains of the educational objectives. You have to note that emphasis on either the cognitive or the affective domain will not develop the interdependence between both. They cannot operate without the third domain – psychomotor. We can only produce an educated individual when the three domains come to play objectives cannot be complex without the psychomotor domain. This is why we have discussed this domain of educational objectives in this unit.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain the psychomotor domain of instructional objectives
- discuss the six levels of psychomotor domain
- give examples of the activities needed in each level.

3.0 MAIN CONTENT

3.1 Psychomotor domain

In unit 5.0 you learnt that cognitive domain has to do with mental abilities. You also noted in the last unit that the affective domain deals with feelings, emotions, values etc. In the same way the psychomotor domain has to do with motor skills or abilities. It means therefore that

the instructional objectives here will make performance skills more prominent. The psychomotor domain has to do with muscular activities. It deals with such activities which involve the use of the limbs (hand) or the whole of the body. These tasks are inherent in human beings and normally should develop naturally.

Can you think of such abilities or skills. Consider the skills in running, walking, swimming, jumping, eating, playing, throwing, etc. One would say that these skills are material.

Yet, for affective performance or high level performance of a wide variety of tasks, it is necessary for educators to develop various skills of more complex nature in addition to the inherent ones. For instance, more complex skills can be developed through learning in such areas as driving, drawing, sports, etc. like the cognitive and affective, psychomotor domain is sub divided into hierarchical levels. From the lowest, we have (i) Reflex movements (ii) Basic Fundamental movements (iii) Perceptual abilities (iv) Physical abilities (v) Skilled movements and (vi) Non-discursive communication. Now let us take them one after the other and discuss them briefly.

3.2 Reflex Movements

At the lowest level of the psychomotor domain are the reflex movements which every normal human being should be able to make. The movements are all natural, except where the case is abnormal, in which case it may demand therapy programmes.

Apart from the abnormal situations, educators are not concerned with these movements. Now let us think of some examples.

Can you mention some of them? Your mind may have gone to; the twinkling of the eyes, trying to dodge a blow or something thrown at you, jumping up when there is danger, swallowing things, urinating or stooling by a child, etc.

3.3 Basic Fundamental Movements

Like the case of reflex movements, these are basic movements which are natural. Educators have little or nothing to do with them, except in an abnormal case where special educators step in to assist. There are three sub-categories at this stage. These are:

i. Locomotor movement: which involves movements of the body from place to place such as crawling, walking, leaping, jumping etc?

ii. Non-locomotor movements: which involves body movements that do not involve moving from one place to another? These include muscular movements, wriggling of the trunk, head and any other part of the body. They also include turning, twisting etc of the body.
iii. Manipulative movements: which involves the use of the hands or limbs to move things to control things etc.

SELF-ASSESSMENT EXERCISE

Give five examples of each of reflex movements and basic fundamental movements.

Your answers may have included:

i. Reflex movement examples = twinkling of the eyes, swallowing things, urinating or stooling by a chilkd, dodging a blow, jumping up when there is danger, falling asleep, etc.
ii. Basic fundamental movements: examples are, jumping, leaping, crawling, walking, muscular movements, wriggling of the trunk, head, turning, twisting, etc.

3.5 Perceptual abilities

This has to do with the senses and their developments. It means therefore that educators have not much to do here except to direct the use of these sense in association with certain conditions. Perceptual abilities are concerned with the ability of the individuals to perceive and distinguish things using the senses. Such individuals recognise and compare things by physically tasting, smelling, seeing, hearing and touching. You can identify the sense organs associated with these activities. With the use of particular taste, smell, sound, appearance and feeling, you can associate and understand certain objects or situations will and feelings in your mind. These senses with now help you to determine conditions and necessary course of action.

3.6 Physical abilities

These abilities fall in the area of health and physical education. You know that in athletics and games or sports in general, you need physical abilities and that these abilities can be developed into varying degrees of perfection with the help of practices. This is why sports men and women always practice in other to improve on their skills of endurance, strength, flexibility and agility. For instance, if you are a goal keeper, you will need to improve on these skills to perform.

3.7 Skilled Movements

This is a higher ability than the physical abilities. Once you have acquired the physical abilities, you now apply various types of these physical abilities in making or creating things. You can combine skills in manipulative, endurance and flexibility in writing and drawing. You can combine the neuromuscular movements together with flexibility to help you in drawing. An individual can combine strength, endurance, flexibility and manipulative movements in activities like combat sports such as wrestling, boxing, karate, tackwando, Judoka, weight lifting, etc.

For skills like drumming, typing or playing the organ or the keyboard in music, you will need a combination of manipulative movements and some perceptive abilities and flexibility.

There are three sub-levels of the skilled movements. These are simple adaptive skills, compound adaptive skills and complex adaptive skills.

SELF-ASSESSMENT EXERCISE

Mention some examples of activities which involve skilled movement.

SELF-ASSESSMENT EXERCISE

Your mind may have gone to sports where you have football, basket ball, long jump, lawn tennis, table tennis, badminton, squash, long jump, hokey, cricket, high jump, pole vault, javelin, shot put, discuss, hammer throw, relay, mile race, cycling, driving, swimming, etc. There are so many of them.

3.8 Non-discursive Communication

This is the highest level which demands a combination of all the lower levels to reach a high degree of expertise.

For instance, one can use the keyboard to play vibes and sounds but it requires a good deal of training, practice and ability to combine certain movements of the fingers in order to relay a message or to play a classical music like that of Handel.

At the same time, it will also require certain level of perceptive abilities in order to be able to interpret or decode the messages or the music. It means that both the player and interpreter must be operating at the same level of skills.

Everybody that is normal can move his limbs and legs. But you must have some level of training, practice and the ability to combine a variety of movements and some perceptive abilities in order to do diving, swimming, typing, driving, cycling etc.

At the same time, you will also need these in other to read or interpret different writings, long or shorthand. You need them to be able to manipulate your computer set accurately to give you what you want. You need training to be able to browse on the Internet and to derive maximum benefits from the use of modern information and communication technologies.

There are two sub-levels of the non-discursive communication. They are expressive movement and interpretive movement.

SELF-ASSESSMENT EXERCISE

Using your own experience describe the skills and expertise involved in the use of your handset in sending a text message.

SELF-ASSESSMENT EXERCISE

This depends on your experience and type of handset.

4.0 CONCLUSION

In this unit, you have gone through the psychomotor domain and its hierarchical levels. But remember that we have said that none of the three domains of educational objectives can function independently of the other two. Note that whether you are improving on your swimming, how to type, learning hand writings, using the talking drum, learning the technique of surgical operations, how to send a message by the talking drum, or handset, or the computer, or radio, etc, playing a musical instrument, piano, weaving basket or cloth, constructions on paper, planning wood, assembling components in metal work or electricity, learning a dancing style, driving, cycling or in sports, etc the three domains must combine. These three domains overlap in some degrees in any learning outcomes. Their separation is arbitrary and only for the purpose of discussions. They must interact in order to produce the educated person.

5.0 SUMMARY

In this unit you read that the psychomotor domain has to do with motor skills, performance skills or muscular activities. It is divided into hierarchical categories from the lowest level of reflex movements which

are natural (ii) Basic fundamental movement which is sub divided into locomotor movement, non-locomotor movement and manipulative movements (iii) Perceptual abilities: which has to do with the sense (iv) Physical abilities which has to do with endurance, strength, flexibility and agility. (v) Skilled movement: which are in three levels: simple adaptive skills, compound adaptive, flexibility and agility. (vi) Non-discursive communication: this is the highest and combines all other levels. It has two sub-levels, which are the expressive movement and the interpretive movement.

6.0 TUTOR-MARKED ASSIGNMENT

1. What is the psychomotor domain of instructional objectives.
2. Briefly explain the physical abilities using examples.
3. What are the two sub levels of non discursive communication.

7.0 REFERENCES/FURTHER READING

Obimba, F.U (1989) Fundamentals of Measurement and Evaluation in Education and Psychology. Owerri: Totan Publishers Ltd.

Onwuka, U. (1990) Curriculum Development for Africa. Onitsha: Africana – FEP Publishers Limited.

UNIT 5 CLASSROOM TESTS

CONTENTS

1.0 INTRODUCTION

The classroom test, which is otherwise called teacher-made test is an instrument of measurement and evaluation. It is a major technique used for the assessment of students learning outcomes.

The classroom tests can be achievement or performance test and/or any other type of test like practical test, etc prepared by the teacher for his specific class and purpose based on what he has taught.

The development, of good questions or items writing for the purpose of classroom test, cannot be taken for granted. An inexperienced teacher may write good items by chance. But this is not always possible. Development of good questions or items must follow a number of principles without which no one can guarantee that the responses given to the tests will be relevant and consistent. In this unit, we shall examine the various aspects of the teacher's own test.

As a teacher, you will be faced with several problems when it comes to your most important functions – evaluating of learning outcomes. You are expected to observe your students in the class, workshop, laboratory, field of play etc and rate their activities under these varied conditions.

You are required to correct and grade assignments and home works. You are required to give weekly tests and end of term examinations. Most of the times, you are expected to decide on the fitness of your students for promotion on the basis of continuous assessment exercises, end of term examinations' cumulative results and promotion examination given towards the end of the school year. Given these conditions it becomes very important that you become familiar with the planning construction and administration of good quality tests. This is because in the next few years when you graduates as a teacher your tests will be used to play a very important role in the growth and progress of Nigerian Youth?

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- list the different types of items used in classroom test
- describe the different types of objectives questions
- describe the different types of essay questions
- compare the characteristics of objectives and essay tests.

3.0 MAIN CONTENT

3.1 Some Pit Falls In Teacher – Made Tests

In unit 4.0, you were told that testable educational objectives are classified by Bloom et al (1956) into recall or memory or knowledge comprehension, application, analysis, synthesis, and evaluation. It means that you do not only set objectives along these levels but also test them along the levels. The following observations have been made about teacher-made tests. They are listed below in order to make you avoid them when you construct your questions for your class tests.

i. Most teacher-test is not appropriate to the different levels of learning outcomes. The teachers specify their instructional objectives covering the whole range simple recall to evaluation. Yet the teachers items fall within the recall of specific facts only
ii. Many of the test exercises fail to measure what they are supposed to measure. In other words most of the teacher-made tests are not valid. You may wonder what validity is. It is a very important quality of a good test, which implies that a test is valid if it measures what it is supposed to measure. You will read about in details later in this course.
ii. Some classroom tests do not cover comprehensively the topics taught. One of the qualities of a good test is that it should

represent all the topic taught. But, these tests cannot be said to be a representative sample of the whole topic taught.

iv. Most test prepared by teacher lack clarify in the wordings. The questions of the tests are ambiguous, not precise, not clear and most of the times carelessly worded. Most of the questions are general or global questions.

v. Most teacher-made tests fail item analysis test. They fail to discriminate properly and not designed according to difficulty levels.

These are not the only pit falls. But you should try to avoid both the ones mentioned here and those not mentioned here. Now let us look at the various types of teacher-made tests items.

3.2 Types of Test Forms used in The Classroom

There are different types of test forms used in the classroom. These can be essay test, objectives test, norm-referenced test or criterion referenced test. But we are going to concentrate on the essay test and objectives test. These are the must common tests which you can easily construct for your purpose in the class.

3.2.1 Essay Test

This is the type of test introduced in this country by the colonial education. In this case the testees or the students have the responsibility of thinking out the answers to the questions asked. They have the freedom to express or state the answers in their own words. It is a free-answer kind of test.

It is used by teachers to measure achievement, performance etc from classroom instruction. In this case you will have the opportunity to write what you can. At this stage, you have a good idea of Essay writing. Since you have studied it in your GST 107, GST 101 & 102 etc.

Before we look at the two types of essay test, let us see the distinctive features of essay tests. These are:

i. Students answer small number of questions. Because of time limits, usually not more than 2 or 3 hours of examination, students are required to answer in their own words nor more than 5 or 6 questions. It is not always that all the topics covered are covered by the tests.

ii. The script are written in the students own style, words and usually in his own handwriting. In some cases spelling errors as well as poor language and handwriting affect students results.

iii. The students are considerably free to organize their own answers. this implies that answers with varying degrees of accuracy and completeness. It encourages creativity by allowing their own unique way. It discourages guess-work and encourages good study habits in students.

Essay tests are of two variations. These are Extended – response type and Restricted response type.

3.2.2 Extended Response or Free Response Type of Essay Test

In this type, questions are asked in a way that the answers demand that the student is not limited to the extent to which he has to discuss the issues raised or question asked. The student has to * plan and organize his thoughts in order to give his answer. * Put his ideas across by expressing himself freely, precisely and clearly using his won words and his own writing. * Discuss the questions at length, giving various aspects of his knowledge on the question asked or issue raised.

SELF-ASSESSMENT EXERCISE

Give 5 examples of free response or Extended responses questions

SELF-ASSESSMENT EXCERCISE

Your answers may have followed this pattern.

i. Describe the sampling technique used in research studies.
ii. Discuss the social, religious and economical implication of religious riots in Nigeria.
iii. Explain the various ways of preventing accident in a school workshop or laboratory.
iv. How can the menace of touting and area boys be eliminated in the country?
v. What are the implications of fuel price hike in our economy.
vi. Describe the processes of producing or cutting screw threads in the school technical workshop.
vii. Describe the processes involved in cement production.
viii. Discuss the problem of cultism in our higher institutions.
ix. Strike should be the last option for workers to show their displeasure". Discuss.
x. Why should the classroom teacher state his instructional objectives to cover the three domains of educational objectives.

3.2.3 Restricted – Response Type

In this type, the questions are so structured that the students are limited, the scope of the response is defined and restricted. The answers given are to some extent controlled.
Now let us give some examples.

- Give three advantages and two disadvantages of essay tests.
- State four uses of tests in education.
- Explain five factors which influence the choice of building site.
- Mention five rules for preventing accident in a workshop.
- State 5 technical drawing instruments and their uses.
- Describe four sources of energy.
- Define Helix and give two applications of helix.

You can give more examples.

3.2.4 Uses of Essay test

The essay test has got some advantages in the business of measurement and evaluation. Some of these advantages are:-

i. The essay test permits a freedom of response, which in turn allows the students to prevent their ideas as detailed as they choose so as to show how deep knowledge they have in the subject area covered by the question.
ii. The free response form allows the student to express himself in his own words making use of his proficiency in the language being used to his advantage.
iii. Essay tests promotes the development of problem – solving skills. This is because the student has to think out the answer himself and put it down in organized form.
iv. It helps students to improve their writing skills such as writing speed, legibility etc because they write in their handwriting.
iv. The essay test is easy to construct. This is why it is very popular.

3.2.5 Limitations of the essay test

i. Subjectivity in scoring is the greatest disadvantage here. The scoring is not reliable because different examiners can grade the score answer differently. In fact, the same examiner can grade the same question differently at different times.
ii. Grading of essay tests is time-consuming.
iii. Essay questions do not cover the course content and the objectives as comprehensively as possible.

iv. Good command of language places individual students at an advantage while poor command places some students at a disadvantage.

3.2.6 How To Make Essay Test Less Subjective

In the last section we said that subjectivity is a major limitation of the essay tests. But you can reduce this to the barest minimum following these tips.

i. Avoid open-ended questions.
ii. Let the students' answer the same questions. Avoid options/choices.
iii. Use students' numbers instead of their names, to conceal their identity.
iv. Score all the answers to each question for all students at a time.
v. Do not allow score on one question to influence you while marking the next. Always rearrange the papers before you mark.
vi. Do not allow your feelings or emotions to influence your marking.
vii. Avoid distractions when marking.

SELF-ASSESSMENT EXERCISE

How would you make essay questions less subjective?

SELF-ASSESSMENT EXERCISE

See section 3.2.6 above.

3.3 Objective tests

The objective test otherwise regarded as the new type of test derives its name from the fact that the marking is done with a standard key. This key concept is that the students are provided with a problem to which a limited numbers of choices are presented for them to select the wanted answer.
It is so much structured that even where the student is to supply the answers, he is strictly limited to give specific and short answers. But students are given the opportunity to react to a large sample of questions which may cover the entire content area. The figure below shows the sub-categories of the objective test.

From the figure shown here in fig. 8.1 you can see that objective test items are divided first into two-supply test items and selection test items. These two are then sub-divided into:

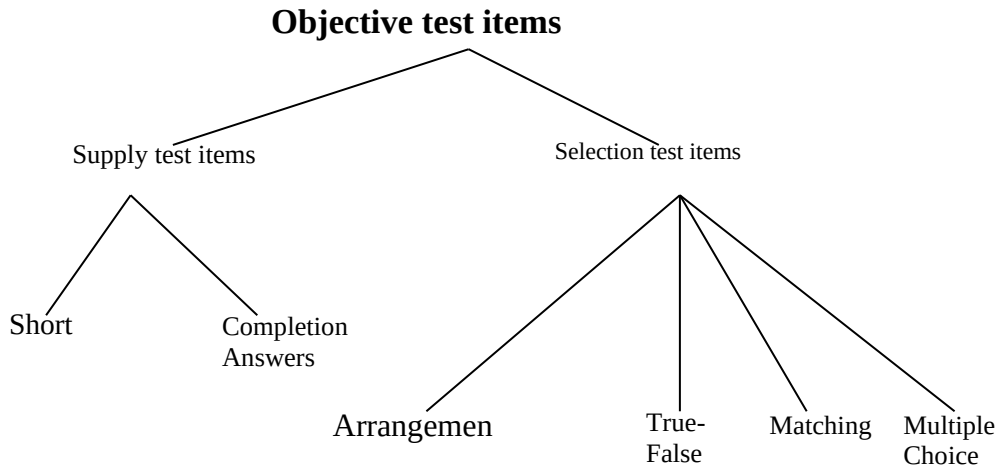


Fig.81. Types of Objectives tests Items

- Short answers
- Completion
- Arrangements
- True-false
- Matching and
- Multiple choice items.

Let us look at them briefly

3.2.2 Supply Test Items

This is the type of test item, which requires the testee to give very brief answers to the questions. These answers may be a word, a phrase, a number, a symbol or symbols etc. in order to write effective supply test items, you should make use of these tips; construct the questions so as to be as brief as possible. The questions should be carefully worded so as to require an exact answer. Make sure that the questions have only one answer each.

Supply test items can be in the form of:

a. short-answer test items: which requires that the testee should provide a short answer to the question asked. Example
   - Who is the senate president of the Nigerian senate?
   - What is the technical drawing instrument used for measuring angles?
   - The two instruments used for cutting screw threads are tap and
   - Which of the energies is used for motion?

b. Completion test items: which require the testee to provide or complete one or more missing words in a sentence. Examples

- A triangle with all sides and all angles equal is called
- The process of splitting a leg of wood into planks is called
- is a special chord passing through the center of circle.

3.2.3 Selection Test Items

This is the type where possible alternatives are provided for the testee to choose the most appropriate or the correct option.

Can you mention them? Let us take them one by one.

a. True-False: In this type a statement is presented to the testee and he is required to state whether it is true or false, Yes or No, Agree or disagree etc. there are only two options.
Examples: In the statement below, answer true or false if the statement is correct and false if the statement is wrong.

- The Tee Square is used for drawing horizontal lines.
- Wood can be hardened and tempered to improve its qualities.
- Any material which obstruct the flow of magnetic flux is called resistor.
- The use of flowing garment in a workshop can cause accident.

b. Multiple choice: this is very useful in measuring students achievement in schools. It is very popular and a widely used type of objective test items. It has two parts. The first is called the stem or premise, which is the question, while the second is the suggested answers called alternatives, options or choices.

The correct option is the answer or the key while other options are called distractors or distracters. Now develop an objective item of the multiple choice type and indicate the parts. Before we give examples of multiple choice items, let us examine the following tips on how to make good multiple choice items.

c. The stems must be simple and clear.

- All alternatives should be possible answers, related to the stems but with only one most correct or best answer to the question.

- The distractors should be as attractive as the key and very effective in distracting testees who are not sure of the key.
- To avoid guessing the correct answers should not be positioned in a particular way so as to form a patter. It should be randomly positioned.
- There should be a minimum of four options for each item. Five options are generally preferred.
- Avoid as much as possible the options 'none of the above', 'all of the above' or a combination of options like 'canded'
- You can now construct good multiple choice items using these hints. Let us see some examples.
- The two forms of oblique projections are:……
  a. axonometric and planometric b. cavalier and planometric c. cavaliers and cabinet
  d. cabinet and axonometric e. pictorial and perspective.
- At what angle should projections line be to an inclined surface to obtain a true shape of the surface?
  - a. 30^0 b. 45^0 c. 60^0 d. 90^0
  - e. 120^0
- Which of these is NOT an autonomous community in 1 litte
  - a. Umuihi b. Ihinna c. Amainyi
  - d. Avutu e. Amakohia
- Which tribe in Nigeri is associated with 'Amala and Ewedu'
  a. Yorubas b. Ndi igbo c. Hausa
  d. Efiks e. Jukuns

SELF-ASSESSMENT EXERCISE

i.. What are the two sub-divisions of the supply test items?

ii. (a) Subjectivity in scoring is a major limitation of essay test? True or false
(b) Essay questions cover the content of a course and the objectives as comprehensively as possible? True False
(c) Grading of essay questions is time consuming? True False
(d) Multiple choice questions should have only two options. True False

iii. Construct five multiple choice questions in any course of your choice.

SELF-ASSESSMENT EXERCISE

1. Short answers and completion items.
2 a. True b. False c. True
d. False
3. You may wish to give the answers from any course of your choice.

c. Arrangement: In this case a number of concepts are presented for the testee to arrange them in a particular order.
Example: Arrange the following numbers in ascending order of size. 58, 75, 14, 69, 50, 35, 64 , 48 and 41.

a. Arrange these in order of construction in building a house. Roof, foundation, floors, walls, floor finishes, painting, wiring.

d. Matching: In this case two parallel columns or lists are involved. One list contains the questions or the premises, while the other list contains the responses. The responses which are usually more than the premises are to be matched according to a given instruction or directions. These two lists constitute the matching set. Examples: Given two lists A and B. match each country from A to the correct player from B.

List A	List B	List C	List D
1. Nigeria	1. Michael Essien	1. Rip Saw	1. Priest
2. Ghana	2. Emmanuel Adebayo	2. Stethoscope	2. Broadcast
3. Egypt	3. Joseph Yobo	3. Bible	3. Physician
4. Ivory Coast	4. Barakat	4. Microphone	4. Carpenter
5. South Africa	5. Beeny Mckaty	5. Trowel	5. Architect
6. Togo	6. Didie Drogba	6. Drawing Set	6. Farmer
7. Senegal	7. Elhaji Diof	7. Wrench	7. Teacher
8. Congo D.R	8. Shovel		8. Lawyer
9. England	9. Yissef Haji		9. Plumber
	10. Lionel Messi		10. Builder
	11. Etoo Fills		11. Journalist
	12. Shevshenko		12. Politian
	13. Michael Owen		
	15. Allan Shearer		

Match the instruments in list C with the profession in B

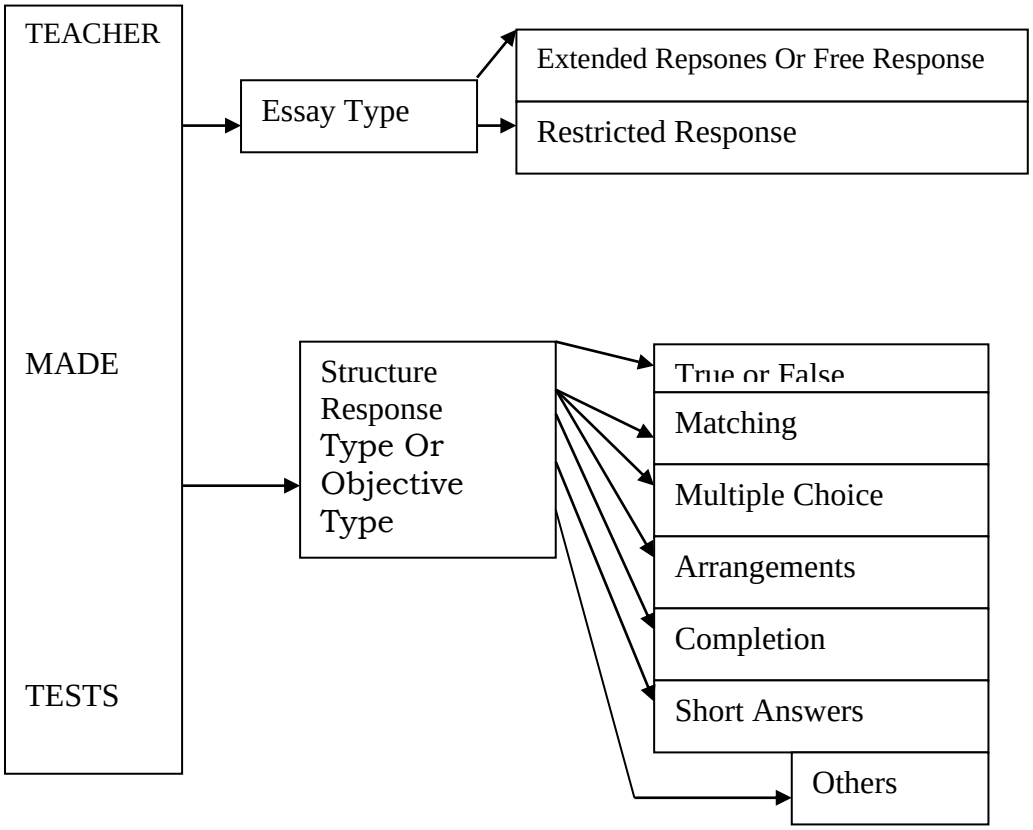
4.0 CONCLUSION

In this unit you have worked through the teacher made test. You have known that we have the essay type and the objective type. The major

difference between them is that whereas in objective tests the answers can be marked without the feelings, biases, knowledge or ignorance of the examiner affecting the scores obtained, in the essay marking may be affected by these factors. You have to follow the hints given in the corresponding sections to avoid any pitfalls.

5.0 SUMMARY

In this unit, we have gone through the teacher made tests used in the classroom. You have learnt how to construct them, the advantages and the disadvantages, how to make use of the advantages and how to avoid the disadvantages. The figure below illustrates the types of teacher made tests.



6.0 TUTOR-MARKED ASSIGNMENT

1. Set any two questions in each of the following types:

   a. Extended response type
   b. Restricted response type
   c. Multiple choice type
   d. True-False type and
   e Completion type.

7.0 REFERENCES/FURTHER READING

Obimba, F.U. (1989) Fundamental of Measurement and Evaluation in Education and Psychology. Owerri Totan Publishers Ltd.

Ohuche, R.O. and Akeju, S.A. (1977) Testing and Evaluation in Education**.** Lagos: African Educational Resources (AER).

Nwana, O.C. (1981) Educational Measurement for Teachers. Lagos: Nelson Africa.

MODULE 3

UNIT 1 TYPES OF TEST

CONTENTS

1.0 INTRODUCTION

In this unit you will be taken through the concept of test which you are already familiar with. Here, we should consider the concept of test as a measuring instrument and its limitations. You will also learn about types of test such as Intelligence Test, Aptitude Test and Achievement Test. While taking you through this unit, you will attempt few Self assessment exercise (SAEs). These are designed to enable you monitor your progress in the Unit. At the end of the unit, you will be requested to answer some Tutor Marked Assignments (TMAs).

1.0 OBJECTIVES

At the end of this unit, you should be able to:

- state the concept of test
- identify types of test
- classify achievement test
- explain types of achievement test.

3.0 MAIN CONTENT

3.1 Concept of Test

As you already know a test is an instrument or device used for finding out the presence or absence of a particular phenomenon or trait possessed by an individual or group of individuals. For instance, one can use an achievement test in integrated science to determine how well the testee learned what he was exposed to.

3.1.1 Concept of Test as Measuring Instrument

A test is the major and most commonly used instrument for the assessment of cognitive behaviours. In this context test simply means a set of questions which students are expected to answer? Their responses to the questions give a measure of their level of performance or achievement.

Usually, the test is based on learned content of subject specific area(s) and is directed to measure the learner's level of attainment of pre-specified objectives. You know that to measure an attribute, a standard instrument is needed. Therefore, unlike physical attributes, measurements are done by describing the characteristics associated with such constructs in behavioural terms.

The expected, behaviours (aptitude) such as ability to state, define, manipulate or perform experiment for instance in integrated science and similar activities are put down in form of test. The test score gives quantitative information about the existence of the construct (attribute) possessed by the testee. For this reason, the test items as measuring instrument must be valid, reliable and usable in order to give dependable result.

3.1.2 Limitations of Test as Measuring Instrument

The limitations of test as measuring instrument arise because a test measures attributes indirectly. Hence the accuracy of information obtained from test results depend on the representativeness and adequacy of the sample of test items with respect to the behaviour associated with the attribute. In other words, a test as a measuring instrument is supposed to have a representative sample of items which measure all and what it purports to measure. Moreover, unlike physical measuring instruments test scores are not absolute. The real value of score of 0 percent does not mean that the learner has zero aptitude and therefore has not learned anything.

However, we know that a learner who scored 60 percent in a given test has more aptitude than another learner who scored 30 percent. But we cannot really say by how much. Therefore, the scores are interpreted with caution.

As you know, no test is accepted universally as standard measure for a specific attribute on its own. A perceived representative samples developed based on some common objective and content areas of a given locality cannot represent all versions of likely attributes of interest to all the students outside that given locality. Thus the use of every test is often localized to specific class, school or area.

3.2 Types of Test

Test may be classified into two broad categories on the basis of nature of the measurement. These are: Measures of maximum performance and measures of typical performance. In measures of maximum performance you have those procedures used to determine a person's ability. They are concerned with how well an individual performs when motivated to obtain as high a score as possible and the result indicates what individuals can do when they put forth their best effort. Can you recall any test that should be included in this category? Examples are Aptitude test, Achievement tests and intelligence tests.

On the other hand measures of typical performance are those designed to reflect a person's typical behaviour. They fall into the general area of personality appraisal such as interests, attitudes and various aspects of personal social adjustment. Because testing instruments cannot adequately be used to measure these attributes self-report and observational techniques, such as interviews, questionnaires, anecdotal records, ratings are sometimes used. These techniques are used in relevant combinations to provide the desired results on which accurate judgment concerning learner's progress and change can be made.

3.2.1 Intelligence Test (or General Mental Ability Test)

You will recall that intelligence is the ability to reason and learn from experience. It is thought to depend both on inherited ability (nature) and on surroundings in which a person is brought up (nurture). The first intelligence tests were devised by Alfred Binet in 1905 to give an Intelligence Quotient (IQ). Intelligence test provides an indication of an individual's general mental capacity. An Intelligence test usually includes a wide variety of tests so as to sample several aspects of cognitive function. Some people believe that Intelligence can be expressed only in speech and writing and therefore cannot be tested.

3.2.2 Aptitude Tests (Separate Ability)

When we talk about aptitude, we refer to the natural talent or ability especially specified. Thus, aptitude tests measure specialized abilities and the potential to learn or perform new tasks that may be relevant to later learning or performance in a specific area. Hence they are future oriented. Can you mention any one of such tests that is familiar to you? An example is the Common Entrance Examination into Vocational Schools and even Secondary Schools.

3.2.3 Achievement Test

Achievement tests are designed to measure the effects of a specific programme of instruction or training which the learners attained usually by their effort. Generally, they represent a terminal evaluation of the learner's status on the completion of a course of study or training. That is, it is used to determine how much the learner has learned from specified content via systemic and controlled instructions. End of term examinations and classroom tests are mostly achievement tests**.**

SELF-ASSESSMENT EXERCISE 1

Now answer the following questions to assess your level of progress.

i. What do you understand by the term test?
ii. Name two examples each of measures of maximum performance and that of typical performance.
iii. Briefly identify the major difference between Intelligence test and Achievement test.

SELF-ASSESSMENT EXERCISE 1

A 1: Your answers should show that test is a measuring instrument that consists of a set of questions carefully designed to assess cognitive behaviours especially in the classroom.
A 2: Measures of maximum performance include: Aptitude tests, Achievement tests and Intelligence test. Measures of typical performance include self reports and observational techniques.
A 3: The major difference between Intelligence test and Achievement test is that while achievement test focuses on measuring the effects of a specific programme of instruction attained by the learners usually by their efforts Intelligence test focus eson inherited ability of the learner and on the surrounding on which the learner is brought up.

3.3 Classification of Achievement Test

Achievement test may be classified in the following ways:

By mode of Response.

- Oral test
- Written test
- Practical test

By Purpose of Testing

- Placement Test
- Formative Test
- Diagnostic Test
- Summative Test

By Desired Speed of Response

- Power test
- Speed test

By Degree of Rigour Employed in Preparation and Scope of Applicability

- Teacher Made Tests
- Standardized Tests

By Mode of Interpreting Results

- Norm –referenced Testing
- Criterion-referenced Testing
- Self-referenced Testing

By Format of Test Items

- Objective Test Items
- Essay Test Items

SELF-ASSESSMENT EXERCISE 1

Match items in set A with those in set B.

SET A	**SET B**
Written test	Purpose of test
Diagnostic test	Desired speed of response
Teacher made test	Rigour employment in Preparation
Power test	Mode of interpreting result
Objective test	Format of test items.

SELF-ASSESSMENT EXERCISE 1

SET A	**SET B**
Written test	Mode of response
Diagnostic test	Purpose of testing
Teacher made test	Rigour employed in preparation
Power test	Desired speed of response
Objective test	Format of test items

4.0 CONCLUSION

In this Unit, you learned about test as a measuring instrument and its limitations. You also learned about types of tests. Here you learned that test may be classified into two broad categories on the basis of nature of measurement. Furthermore, you learned more about classifications of Achievement test which is most commonly used in classroom.

5.0 SUMMARY

In this Unit, you learned that

- Test is an instrument or device used for for finding out the presence or absence of a particular phenomenon or trait possessed by an individual or group of individuals.
- A test is the major and most commonly used instrument for the assessment of cognitive behaviours.
- The limitations of test as a measuring instrument arise because a test measures attributes indirectly.
- Tests are classified into two broad categories namely measures of maximum performance and measures of typical performance.
- Examples of measures of maximum performance are Aptitude test, Achievement test and Intelligence tests

6.0 TUTOR-MARKED ASSIGNMENT

1. Briefly state the concept of test.
2. Why is test limited as a measuring instrument?
3. What is an achievement test?

7.0 REFERENCES/FURTHER READING

Iwuji, V.B.C. (1997) Measurement and Evaluation for Effective Teaching and Learning. CRC Publishing. Owerri:.

Onasanya, Kola (1991), Evaluation of Students Achievement .Pius Debo (Nigeria), Ijebu –Od:,

Ughamadu, K.A (1991). Understanding and Implementing Continuous Assessement. Benin City :World of Books Publishers, , Nigeria.

UNIT 2 ESSAY TEST

CONTENTS

1.0 INTRODUCTION

In the last Unit, you learned about types of Test. There you learned that Achievement

Test may be classified by Format of test items into Essay and Objective test items. In this Unit, you will learn about Essay test, it advantages and disadvantages. Further more you will learn about the classification and types of Essay test. Finally you will learn how to construct Essay test.

2.0 OBJECTIVE

At the end of this unit, you should be able to:

- explain the meaning of Essay test;
- state the advantages and disadvantages of Essay test;
- enumerate when to use essay test;
- classify types of essay item response; and
- construct the essay questions

3.0 MAIN CONTENT

3.1 Essay Test

Essay tests are tests consisting of questions (items) designed to elicit From the learners through freedom of response the extent to which they have acquired the behaviour called for in the course objectives. The answers to such questions which the learners are confronted vary in quality and degree of correctness. Most times, these answers are not complete and thorough.

They also have poor psychometric quality or measurement qualities although popular among classroom teachers especially those who are deficient in the skill required for item construction. For this reason, we should examine how to construct essay items in this unit and subsequently examine how to administer and score such items in order to improve their reliability and validity.

3.1.1 Advantages of Essay Test

- It measures complex learning outcomes that cannot be measured by other means.
  For instance, it has the ability to measure learner's communication skills. That is, the learner's ability to produce an answer, synthesize and organize ideas and present them readably in a logical and coherent form. This is the major advantage.
- It also enables the measurement of organizational and divergent thinking skills by laying emphasis on the integration and application of thinking and problem solving skills, creativity and originality.
- It is very applicable for measuring learning outcomes at the higher levels of educational objectives such as application, analysis, synthesis and evaluation of levels of the cognitive domain.
- It is easy and economical to administer. It can be easily and conveniently written on the chalkboard because of the few items involved. This saves materials and time for production.
- Essay item is easy to construct and does not take much time. This fact has to be guarded seriously to avoid constructing questions that can be very misleading by not asking for specific behaviours emphasized in a particular set of learning outcomes.
- It can be used to measure in-depth knowledge especially in a restricted subject matter area.
- It does not encourage guessing and cheating during testing.

3.1.2 Disadvantages of Essay Test

Despite the advantages already proffered for essay test, it does not satisfy the two most important qualities of a good measuring instrument. Its advantages include that:

- It is inadequate in sampling subject matter content and course objectives since it provides limited sampling. The provision of few questions results in the invalid and narrow coverage of subject matter and instructional objectives. Also as Nenty (1985), rightly pointed out, fewness of the number of questions often asked encourages permutation of some content areas and creaming of ideal responses to suspected questions. In this regard, essay questions discourage the development of good study habit.
- In addition to the invalidity of the measurement, evaluating the answers to carelessly developed questions tends to be confusing and time consuming task. This results in poor reliability in scoring. Studies have shown that answers to essay questions are scored differently by different teachers and that even same teachers score the answers differently at different times. A variation which range from near perfect scores to those representing dismal failure. This may be attributed to the inability of scorers to clearly identify the learning outcomes being measured. When the evaluation of answers is not guided by clearly defined outcomes, it tends to be based on less stable, initiative judgments.
- Sometimes an essay question implies many skills other than that which the item was intended to measure. The testee therefore perceives and reacts to the same questions differently. The differences in the perception of the questions encourage bluffing and hides differences in the knowledge of basic factual material and the learners ability to use and organize such facts.
- The essay test item does not readily lend itself to empirical study of item qualities like difficulty and discrimination based on which improvements on the item could be made.

3.1.3 When to use Essay Questions

- You should use essay questions in the measurement of complex achievement when its distinctive feature of freedom of response is required. Learners are free to select, relate and present ideas in their own words. This freedom enhances the value of essay questions as a measure of complex achievement but it introduces scoring difficulties that make them insufficient as measure of factual knowledge.

- Essay questions should also be used to measure those learning outcomes that cannot be measured by objective test items. The specific features of essay questions can be utilized most fully when their shortcomings are offset by the need for such measurement.
- They should be used when learning outcomes concerned with the abilities to select, organize, integrate, relate, and evaluate ideas require the freedom of response and the originality provided by essay questions .More so, when these outcomes are of such great educational significance that the expenditure of energy in the difficulty and time-consuming task of evaluating the answers can be easily justified.

SELF-ASSESSMENT EXERCISE 1

i. Briefly identify the two most outstanding weaknesses of essay test as a measuring instrument.
ii. When should essay test be used?

SELF-ASSESSMENT EXERCISE 1

A1: The two outstanding weaknesses of essay test as measuring instrument are the Validity and Reliability of most essay questions.
A2: Essay test should be used in the measurement of complex achievement when its distinctive feature and freedom of response is required.

3.2 Classification of Essay Test items

Essay questions are classified into two types namely the restricted response type and the extended response type. The classification is based on the degree of freedom of response associated with the question. For instance, an essay question may require just a few short sentences as answer as in the short-answers objective item where a sentence or two could be all that is required. Whereas, another essay question may give the examinees complete freedom in making their responses and their answers may require several pages.

However, there are variations in freedom of response that fall within these extreme conditions. But for convenience, essay questions are presently classified as restricted response type in which examinees are given almost complete freedom in making their responses and the restricted response type in which the nature, length or organization of the response is limited.

3.3 Types of Essay Item Response

As already discussed, there are two types of essay item response. These are the responses to restricted item questions and responses to extended response questions.

3.3.1 Extended Responses Essay Questions

These are responses to essay questions in which the examinee is only restricted by time as no bound is placed as regards the dept, breadth and the organization of the response. An example of question in this category is: "Open and Distance Learning is a viable option for the eradication of illiteracy in Nigeria. Discuss"

In response to such a question the examinee demonstrates his ability to select and recall the facts which he thinks are pertinent, organize and present his ideas in a logical and coherent form. This freedom to decide which facts the thinks is most pertinent to select his own method of organization and to write as much as seems necessary for a comprehensive answer tends to reveal the ability to evaluate ideas, relate them, coherently and to express them succinctly. In addition, they expose the individual differences in attitudes, values and creative ability.

This type of essay item is mostly useful in measuring learning outcomes at the higher cognitive levels of educational objectives such as analysis, synthesis and evaluation levels. Although, the extended response essay type are also limited by two weaknesses which are:

- They are insufficient for measuring knowledge of factual materials because they call for extensive details in selected content area at a time.
- Scoring such responses is usually difficult and unreliable since the examinees have free will in the array of factual information of varying degree of correctness, coherence and expression.

These limitations are minimized in the Restricted Response Type.

3.3.2 Restricted Response Essay Questions

In this type the examinee is limited to the nature, length or organization of response to be made. The items are directional questions and are aimed at the desired responses. This limits the examinee freedom to select, recall, and synthesize all that he knows and to present them logically as he may wish. This type of essay item is most useful in measuring learning outcomes at the lower cognitive levels of educational objectives, that is, knowledge, comprehension and

application levels. An example of restricted response essay question is "state two advantages and two disadvantages of essay questions".

The restricted nature of the expected response in this type of items makes it more efficient for measuring knowledge of factual material. It reduces to a reasonable extent the difficulty of scoring and encourages more reliability in scoring. However, the restriction makes it less effective as a measure of ability to select, organize and integrate ideas and present them in an original and coherent form which is one of the major advantages of essay test.

3.4 Constructing the Essay Questions

You are now aware of the handicaps of essay questions as a measuring instrument. Therefore, an essay test is a useful measurement instrument only to the extent that it is constructed, administered and scored to ensure a high level of objectivity. For this reason, essay test items should consist of items that will ensure the same understanding and elicit only the skill or ability one is interested in measuring from every examinee. Also, the responses are to be such to which two or more examiners would assign the same score and should attract consistent interpretation from everybody.

You know that this is difficult to achieve and needs a lot of effort. Hence the following points are suggested as guide for construction of good essay test item that call for the desired behaviour.

i. Restriction of the use of essay questions to only those learning outcomes that cannot be satisfactorily measured by objective items. That is, essay questions are to be used only when it's desirable and very adequate for measuring the learning outcomes for full realization of learner's achievement. In other words, they are to be used for questions that call for complex learning outcomes that pertain to the organization, integration and expression of ideas which would not have been possible without the use of essay test items.
ii. Formulation of questions that call forth the behaviour specified in the learning outcomes. Essay questions should be designed to elicit only the skill which the item was intended to measure. This can be achieved by expressing clearly and precisely the question in line with clearly defined instructional objective. In addition, an action verb like compare, contrast, illustrates, differentiates, criticized and so on could be used to give the test items more focus.

iii. Phrase each question to clearly indicate the examinees task. An essay question has to specify precisely what is required of the examinee. Ensure that the testee's task is clearly indicated by delimiting the area covered by the item, using descriptive words to give specific direction towards the desired response. Indicate the score allotted to the test. This suggestion easily lend itself to restricted response type and care should be taken not to narrow the questions when constructing the extended response type in order not to reduce it's effectiveness as a measuring of the ability to select, organize and integrate ideas. Also, adapt the length and complexity of the answer to the testees' level of maturity.

iv. Indication of approximate time limit for each question. It is necessary to indicate time allotted to each question to enable the testees to pace their writing on each question and to allay any anxiety that might arise. The timing should take care of slower testees writing speed so as not to put them at disadvantage for a satisfactory response.

v. Avoidance of the use of optional questions. The provision of optional questions although generally favoured by testees obviously means that they are taking different tests and therefore the common basis for evaluating their achievement is lost. Moreover, optional questions might also influence the validity of test results since some examinees may be favoured in their advanced preparation of selected areas of study. It is also not easy to construct essay questions of the same difficulty level. Hence, making valid comparisons of performance among them especially for norm reference setting will not be possible.

SELF-ASSESSMENT EXERCISE 2

i. How would you classify the essay test items?

ii. List the five major considerations in the construction of essay test items.

SELF-ASSESSMENT EXERCISE 2

A1: The essay test items are classified into two namely:
- i Extended response essay test items; and
- ii Restricted response essay test items.

A2: The five major considerations in the construction of essay test items are:

i. Restriction of the use of essay questions to only those learning outcomes that cannot be satisfactory measured by objective items.

ii Formulation of questions that call forth behaviour specified in the learning outcomes.
iii Phrasing each question to clearly indicate the examinees task..
iv Indication of approximate time limit for each question.
v. Avoidance of the use of optional question.

4.0 CONCLUSION

In this unit you learned about essay test. Specifically, you learned about its advantages and disadvantages and when to use essay test. Furthermore, you learned about the classification of essay test and the considerations for the construction of good essay test items.

5.0 SUMMARY

In this unit, you learned that:

- Essay tests are tests consisting of questions (items) designed to elicit from the learners, through freedom of response, the extent to which they have acquired the behaviour called for in the course objectives.
- The main advantage of the essay question is that it measures complex learning outcomes that cannot be measured by other means. The extended response question lays emphasis on the integration and application of thinking and problem-solving skills. These questions have a desirable influence on learners' study habits.
- It is inadequate for sampling subject matter content and course objectives because of the limited sample it provides. The scoring is also unreliable.
- Essay questions should be used in measurement of complex achievement when its distinctive feature of freedom of response is required.
- There are two types of essay questions. These are the extended response type and the restricted response type.
- Essay test should be constructed, administered and scored to ensure a high level of objectivity. It should consist of items that will ensure the same understanding and elicit only the skill or ability one is interested in measuring.

6.0 TUTOR-MARKED ASSIGNMENT

1. What do you understand by essay test?
2. What are the advantages of essay test?
3. List the five major considerations in the construction of essay test items.

7.0 REFERENCES/FURTHER READING

Gronlund, N.E (1985). Measurement and Evaluation in Teaching. New York.: Macmillan Publishing Company.

Nenty, H.J (1985), Fundamentals of Measurement and Evaluation in Education. Calabar: University of Calabar Press.

UNIT 3 OBJECTIVE TEST

CONTENTS

1.0 INTRODUCTION

In the last unit you learned about Essay Test. In this unit you will learn about the objective test. You will learn about the advantages and disadvantages of objective test. Furthermore, you will learn when to use objective test and types of objective test. Finally, you will learn how to construct the various types of objective test.

2.0 OBJECTIVE

At the end of this unit, you should be able to:

- explain the meaning of objective test,
- state the advantages and disadvantages of objective test;
- identify when to use objective test;
- enumerate the various types of objective test and their peculiar advantages and disadvantages;
- construct the various types of objective test.

3.0 MAIN CONTENT

3.1 Objective Test

Objective tests are those test items that are set in such a way that one and only one correct answer is available to a given item. In this case every scorer would arrive at the same score for each item for each examination even on repeated scoring occasions. This type of items sometimes calls on examinees to recall and write down or to supply a word or phrase as an answer (free – response type). It could also require the examinees to recognize and select from a given set of possible answers or options the one that is correct or most correct (fixed-response type). This implies that the objective test consist sof items measuring specific skills with specific correct response to each of the items irrespective of the scorer's personal opinion, bias, mood or health at the time of scoring.

3.1.1 Advantages of Objective Test

- Objective test enhances the assessment of learner's response to test items because the scoring is not influenced by the scorer's bias or disposition at the time of scoring but by the correctness of the answer. By overcoming the subjectivity of the essay test, the reliability of the test as measuring instrument is enhanced.
- Scoring of objective test is easy and takes little time. It is also scored by a machine and facilitates high efficiency in testing a large number of examinees.
- The result of objective test especially the multiple choice items can be used for diagnostic purposes since they provide clues for factual errors and misunderstanding that need remediation.
- It is adequate for sampling the subject matter and instructional objectives of the course because the relatively large number of items set enhances effective coverage of the content areas on which the test is based. The result provides a more valid and reliable ability of the examinees performance.
- It is efficient for measuring knowledge of facts. It can also be designed to measure understanding, thinking skills and other complex outcomes.
- Objective test items can be pre-test, refined through item analysis, standardized and reused a number of times if properly handled.
- It is fair to all examinees since it does not call on other skills outside the skill it is intended to measure. That is, its validity is not affected by good handwriting, bluffing or the verbiage.

3.1.2 Disadvantages of Objective Test

- It does not encourage the development of examinees originality in desirable skills such as the ability to select, organize or synthesize ideas and to present them correctly in a logical and coherent form. The complete structuring of task is not suitable for assessing learning abilities in this form.
- It tends to measure only factual knowledge. This disadvantage can be overcome by developing items for the objective items rigorously following the steps involved in item development process.
- Development of good objective test items requires training of test developers in the skills necessary for constructing effective, valid and reliable items.
- It needs time, commitment and adequate planning.
- Objective test items lend themselves to guessing especially when the test items are not skillfully developed. An examinee can guess correctly on few items and earn some undeserved points even in a well-constructed objective test. It is also easier to cheat in an objective test than in essay test if the test is poorly administered.

3.1.3 When to use objective test

- It is used when highly structured task are needed to limit the type of response the examinees can make and to obtain correct answers from learners by demonstrating the specific knowledge or skill called for in the item.
- It is used to appraise more effectively the achievement of any of the educational objectives of simple learning outcomes as well as the complex outcomes in the knowledge, understanding, and application and even in higher levels covering large content areas if skillfully constructed. It is possible to set as many as 120 objective tests spread over many lesson units and several cognitive levels of educational objective for one hour or two hours.
- It is used when objective, quick, easy and accurate scoring is desired especially when the number of examinees is large.
- It is used to measure understanding, thinking skills and other complex learning outcomes of the learners.
- It can also be used for diagnosis of learning deficiency and the result used for remediation process.

SELF-ASSESSMENT EXERCISE 1

1. Briefly explain the meaning of objective test.
2. What are the two major advantages of objective test items over the Essay test item?
3. What is the major feature of objective test that distinguishes it from essay test?

SELF-ASSESSMENT EXERCISE 1

A1: Objective test is that test which consists of items set in such as way that one and only one correct answer is available to a given item. That is, the objective test consists of items measuring specific skills with specific correct response to each of the items irrespective of the scorer's personal opinion, bias, mood or health at the time of scoring.

A2: The two major advantages of objective test items over the essay test items are that if skillfully constructed they are more reliable and valid as measuring instrument.

A3: The major feature of objective test that distinguishes it from essay test is that the objective test presents the examinees with a highly structured task that limits the type of response they can make. To obtain the correct answer the examinees must demonstrate the specific knowledge, understanding, or skill called for in the item. They are not free to redefine the problem or to organize and present the answer in their own words.

3.2 Types of Objective Test

The objective test can be classified into those that require the examinee to supply the answer to the test items (free-response type) and those that require the examinee to select the answer from a given number of alternatives (fixed response type). The free-response type consists of the short answer and completion items while the fixed response type is commonly further divided into true-false or alternative response matching items and multiple-choice items.

3.2.1 The Free Response Test Items

The free response type of objective test tends to represent a compromise between the essay and the objective items. The free response type namely short-answer item and the completion item both are supply-type test items consisting of direct questions which require a short answer (short-answer type) or an incomplete statement or question to which a

response must be supplied by an examinees (completion type). The answers to such questions could be a word, phrase, number or symbol. It is easy to develop and if well developed, the answers are definite and specific and can be scored quickly and accurately. An example of question in the class is:

Short Answer: Who is the first Vice Chancellor of the National Open University of Nigeria? (Professor Olugbemiro Jegede).

Completion: The name of the first Vice Chancellor of the National Open University of Nigeria is (Professor Olugbemiro Jegede).
The free – response type is very adaptable for item construction in mathematics, physical sciences and other areas where questions are computational problems requiring examinees to supply the solutions.

Uses

- It is suitable for measuring a wide variety of relatively simple learning outcomes such as recall of memorized information and problem solving outcomes measured in mathematics and sciences.
- It can be used to measure the ability to interprete diagrams, charts, graphs and pictorial data.
- It is used when it is most effective for measuring a specific learning outcome such as computational learning outcomes in mathematics and sciences.

Advantages

- It measures simple learning outcomes, which makes it easier to construct.
- It minimizes guessing because the examinees must supply the answer by either think and recall the information requested or make the necessary computations to solve the problem presented. It is unlike the selection item where partial knowledge might enable the examinee to choose the correct answer.

Disadvantages

- It is not suitable for measuring complex learning outcomes. It tends to measure only factual knowledge and not the ability to apply such knowledge and it encourages memorization if excessively used.
- It cannot be scored by a machine because the test item can, if not properly worded, elicit more than one correct answer. Hence the scorer must make decision about the corrections of various

responses. For example, a question such as "Where was Dr. Nnamdi Azikiwe born?" Could be answered by name of the town, state, country or even continent. Apart from the multiple correct answers to this question, there is also the possibility of spelling mistakes associated with free-response questions that the scorer has to contend with.

3.2.2 The Alternative Response Test Item

The alternative response test item commonly called the true-false test item because the true-false option is commonly used consists of item with declarative statement to which the examinee is asked to give either of two options concerning the item. The two options could be true or false, right or wrong, correct or incorrect, yes or no, fact or opinion, agree or disagree and so on.

Most times the alternative response item includes opinion statement and the examinee is also required to response to them as merely true or false. The opinion item is not desirable from the standpoint of testing, teaching and learning. If opinion statement is to be used, it has to be attributed to some source thereby making it possible to assign the option of true or false to the statement based on knowledge concerning the belief held by an individual or the values supported by an organization or institution. An example of alternative response item is as follows:

Read the following statement if the statement is true circle the **T** if it is false circle the **F.**

T. F Solar Energy is the energy radiated from the sun.

The correct answer to the example above is true and is always true.

Uses

- It is commonly used to measure the ability to identify the correctness of statements of fact, definitions of terms, statements of principles and other relatively simple learning outcomes to which a declarative statement might be used with any of the several methods of responding.
- It is also used to measure examinee ability to distinguish fact from opinion; superstition from scientific belief.
- It is used to measure the ability to recognize cause – and – effect relationships.
- It is best used in situations in which there are only two possible alternatives such as right or wrong, more or less, and so on.

Advantages

- It is easy to construct alternative response item but the validity and reliability of such item depend on the skill of the item constructor. To construct unambiguous alternative response item, which measures significant learning outcomes, requires much skill.
- A large number of alternative response items covering a wide area of sampled course material can be obtained and the examinees can respond to them in a short period of time.

Disadvantages

- It requires course material that can be phrased so that the statement is true or false without qualification or exception as in the Social Sciences.
- It is limited to learning outcomes in the knowledge area except for distinguishing between facts and opinion or identifying cause – and – effect relationships.
- It is susceptible to guessing with a fifty-fifty chance of the examinee selecting the correct answer on chance alone. The chance selection of correct answer has the following effects.

  i. It reduces the reliability of each item thereby making it necessary to include many items in order to obtain a reliable measure of achievement.
  ii. The diagnostic value of answers to guess test items is practically nil because analysis based on such response is meaningless.
  iii. The validity of examinees response is also questionable because of response set.

Response set is a consistent tendency to follow a certain pattern in responding to test items. For instance some examinees will consistently mark “true” those items they do not know while others will consistently mark them “false. Any given test will therefore favour one response set over another thereby introducing an element into the test score that is irrelevant to the purpose of the test.

3.2.3 The Matching Test Items

The matching test items usually consist of two parallel columns. One column contain a list of word, number, symbol or other stimuli (premises) to be matched to a word, sentence, phrase or other possible answer from the other column (responses) lists. The examinee is directed to match the responses to the appropriate premises. Usually,

the two lists have some sort of relationship. Although the basis for matching responses to premises is sometimes self-evident but more often it must be explained in the directions.

The examinees task then is to identify the pairs of items that are to be associated on the basis indicated. Sometimes the premises and responses list is an imperfect match with more list in either of the two columns and the direction indicating what to be done. For instance, the examinee may be required to use an item more than once or not at all, or once. This deliberate procedure is used to prevent examinees from matching the final pair of items on the basis of elimination. An example of matching item is given below.

Choose the <u>most</u> appropriate approach to Validity from the list in Column B that matches each of the Validity Evidence on the list in Column A.

Column A	Column B
Validity Evidence	Approaches to Test Validation
a) Content-Related Evidence measure of performance	i. Compare test scores with another obtained at a later date
b) Criterion-Related Evidence scores on the test by test and experimentally influence test performance.	ii. Establish the meaning of the controlling the development of the determining the factors that
c) Construct-Related Evidence test tasks represents the	iii. Establish how well the sample of domain of tasks to be measured.
	iv. Compare the test tasks to the test . specifications describing . the test domain under consideration.

Uses

- It is used whenever learning outcomes emphasize the ability to identify the relationship between things and a sufficient number of homogenous premises and responses can be obtained.
- Essentially used to relate two things that have some logical basis for association.
- It is adequate for measuring factual knowledge like testing the knowledge of terms, definitions, dates, events, references to maps and diagrams.

Advantages

- The major advantage of matching exercise is that one matching item consists of many problems. This compact form makes it possible to measure a large amount of related factual material in a relatively short time.
- It enables the sampling of larger content, which results in relatively higher content validity.
- The guess factor can be controlled by skillfully constructing the items such that the correct response for each premise must also serve as a plausible response for the other premises.
- The scoring is simple and objective and can be done by machine.

Disadvantages

- It is restricted to the measurement of factual information based on rote learning because the material tested lend themselves to the listing of a number of important and related concepts.
- Many topics are unique and cannot be conveniently grouped in homogenous matching clusters and it is some times difficult to get homogenous materials clusters of premises and responses that can sufficiently match even for contents that are adaptable for clustering.
- It requires extreme care during construction in order to avoid encouraging serial memorization rather than association and to avoid irrelevant clues to the correct answer.

3.2.4 The Multiple Choice Test Items

The multiple choice item consists of two parts – a problem and a list of suggested solutions. The problem generally referred to as the stem may be stated as a direct question or an incomplete statement while the suggested solutions generally referred to as the alternatives, choices or options may include words, numbers, symbols or phrases. In its standard form, one of the options of the multiple choice item is the correct or best answer and the others are intended to mislead, foil, or distract examinees from the correct option and are therefore called distracters, foils or decoys. These incorrect alternatives receive their name from their intended function – to distract the examinees who are in doubt about the correct answer. An example of multiple-choice item is given below.

Which one of the following factors contributed most to the selection of Abuja as the Federal Capital Territory of Nigeria?

(A) Central location
(B) Good climate
(C) Good highways
(D) Low population
(E) High population.

The best-answer form of Multiple Choice Item is usually more difficult than the correct answer form. This is because such items are used to measure more complex learning outcomes. It is especially useful for measuring learning outcomes that require the understanding, application or interpretation of factual information. An example is given below.

Which of these best describe the property of speed?

(A) It has magnitude
(B) It has direction
(C) It is a scalar quantity
(D) It is a vector quantity
(E) It has magnitude and direction.

Uses

- The multiple-choice item is the most widely used of the types of test available. It can be used to measure a variety of learning outcomes from simple to complex.
- It is adaptable to any subject matter content and educational objective at the knowledge and understanding levels.
- It can be used to measure knowledge outcomes concerned with vocabulary, facts, principles, method and procedures and also aspects of understanding relating to the application and interpretation of facts, principles and methods.
- Most commercially developed and standardized achievement and aptitude tests make use of multiple-choice items.

Advantages

- The main advantage of multiple-choice test is its wide applicability in the measurement of various phases of achievement.
- It is the desirable of all the test formats being free of many of the disadvantages of other forms of objective items. For instance, it present a more well-defined problem than the short-answer item, avoids the need for homogenous material necessary for the matching item, reduces the clues and susceptibility to guessing characteristics of the true-false item and is relatively free from response sets.

- It is useful in diagnosis and it enables fine discrimination among the examinees on the basis of the amount of what is being measured possessed by them.
- It can be scored with a machine.

Disadvantages

- It measures problem-solving behaviour at the verbal level only.
- It is inappropriate for measuring learning outcomes requiring the ability to recall, organize or present ideas because it requires selection of correct answer.
- It is very difficult and time consuming to construct.
- It requires more response time than any other type of objective item and may favour the test-wise examinees if not adequately and skillful constructed.

3.2 Constructing the Objective Test Items

You have seen that simple put a test item is a statement sometimes in question form that tries to elicit a testee's level of knowledge, ability or understanding of a specific subject matter. Therefore, writing a good test item is an art that requires some skill, time, perseverance, and creativity. The following are some general guidelines for the construction of any type of objective test item.

- The wording of the item should be clear and as explicit as possible.
- Avoid setting interrelated items
- Items should be designed to test important and not trivial facts or knowledge.
- Write an item to elicit discriminately the extent of examinees possession of only the desired behaviour as stipulated in the course instructional objectives answers.
- Ensure that there is one and only one correct or best answer to each item.
- Avoid unintentionally giving away the answer through providing irrelevant clues.
- Use language appropriate to the level of the examinees.
- Items in an achievement test should be constructed to elicit specific course content and not measure general intelligence.
- Have an independent reviewer to vet your test items.

(Nenty, 1985: 204)

SELF-ASSESSMENT EXERCISE 2

Which of the types of Objective test items would you recommend for school wide test and why?

SELF-ASSESSMENT EXERCISE 2

A: The Objective test item to be recommended is the Multiple Choice Type.

Reasons

i. It has all the advantages of the other types of Objective test formats.
ii. It is free of many of the limitations of other forms of Objective test items.

4.0 CONCLUSION

In this unit, you learned about the objective test items. Specifically you learned about the free-response items, the matching items, the alternative response items and the multiple-choice items. Furthermore, you learned about their advantages, disadvantages and when to use each of the objective test formats. Finally you learn the general guideline for the construction of objective test items.

5.0 SUMMARY

- Objective test are those test items that are set in such a way that one and only one correct answer is available to a given item.
- The various types of objective test discussed in this unit are:

  - The Free Response Test Item
  - The Alternative Response Test Item
  - The Matching Test Item
  - The Multiple Choice Item.

- The Multiple Choice Item is the best and most widely used of all the objective test items. It consists of two parts- a problem and a list of suggested solutions. The problem is referred to as the stem while the options are referred to as distracters.
- General guideline for the construction of valid and reliable objective test items is also listed.

6.0 TUTOR-MARKED ASSIGNMENT

1. What do you understand by Objective Test?
2. List the disadvantages of Objective test
3. When should Objective test be used?

7.0 REFERENCE/FURTHER READING

Gronlund, N.E. (1985). Measurement and Evaluation in Teaching. New York: Macmillan Publishing Company,

Nenty, H.J (1985). Fundamentals of Measurement and Evaluation in Education Calabar, University of Calabar.

UNIT 4 TEST DEVELOPMENT – PLANNING THE CLASSROOM TEST

CONTENTS

1.0 INTRODUCTION

In this unit, you will learn how to plan a classroom test. You will learn what to consider in the planning stage, how to carry out content survey and to scrutinize the instructional objectives as relevant factors in the development of table of specification/test blue print. Thereafter you will learn how to develop the test blue print, moderate items generated and prepare the items for use

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- identify the sequence of planning a classroom test,
- prepare a table of specification for classroom test in a given subject.
- carry out item moderation processes; and
- assemble moderated test items for use.

3.0 MAIN CONTENT

3.1 Test Development – Planning the Classroom Test

The development of valid, reliable and usable questions involves proper planning. The plan entails designing a framework that can guide the test developers in the items development process. This is necessary because classroom test is a key factor in the evaluation of learning outcomes. The validity, reliability and usability of such test depend on the care with which the test are planned and prepared. Planning helps to ensure that the test covers the pre-specified instructional objectives and the subject matter (content) under consideration. Hence, planning classroom test entails identifying the instructional objectives earlier stated, the subject matter (content) covered during the teaching/learning process. This leads to the preparation of table of specification (the test blue print) for the test while bearing in mind the type of test that would be relevant for the purpose of testing. An outline of the framework for planning the classroom test is as prescribed below.

3.1.1 Considerations in Planning a Classroom Test.

To plan a classroom test that will be both practical and effective in providing evidence of mastery of the instructional objectives and content covered requires relevant considerations. Hence the following serves as guide in planning a classroom test.

- Determine the purpose of the test;
- Describe the instructional objectives and content to be measured.
- Determine the relative emphasis to be given to each learning outcome;
- Select the most appropriate item formats (essay or objective);
- Develop the test blue print to guide the test construction;
- Prepare test items that is relevant to the learning outcomes specified in the test plan;
- Decide on the pattern of scoring and the interpretation of result;
- Decide on the length and duration of the test, and
- Assemble the items into a test, prepare direction and administer the test.

3.1.2 Scrutiny of the Instructional Objectives

The instructional objectives of the course are critically considered while developing the test items. This is because the instructional objectives are the intended behavioural changes or intended learning outcomes of instructional programmes which students are expected to possess at the

end of the course or programme of study. The instructional objectives usually stated for the assessment of behaviour in the cognitive domain of educational objectives are classified by Bloom (1956) in his taxonomy of educational objectives into knowledge, comprehension, application, analysis, synthesis and evaluation. The objectives are also given relative weight in respect to the level of importance and emphasis given to them. Educational objectives and the content of a course form the nucleus on which test development revolves.

3.1.3 Content Survey

This is an outline of the content (subject matter or topics) of a course of programme to be covered in the test. The test developer assigns relative weight to the outlined content – topics and subtopics to be covered in the test. This weighting depends on the importance and emphasis given to that content area. Content survey is necessary since it is the means by which the objectives are to be achieved and level of mastering determined.

3.1.4 Planning the table of specification/test blue print

The table of specification is a two dimensional table that specifies the level of objectives in relation to the content of the course. A well-planned table of specification enhances content validity of that test for which it is planned. The two dimensions (content and objectives) are put together in a table by listing the objectives across the top of the table (horizontally) and the content down the table (vertically) to provide the complete framework for the development of the test items. The table of specification is planned to take care of the coverage of content and objectives in the right proportion according to the degree of relevance and emphasis (weight) attached to them in the teaching learning process. A hypothetical table of specification is illustrated in table 3.1 below:

Table 3.1 A Hypothetical Test Blue Print/Table of Specification.

Content		Objectives						Total
Area	Weight	Knowledge 10%	Comprehension 15%	Application 15%	Analysis 30%	Synthesis 10%	Evaluation 20%	100%
Set A	15%	-	1	-	2	-	-	3
Set B	15%	-	1	-	2	-	-	3
Set C	25%	1	-	1	1	1	1	5
Set D	25%	1	-	1	1	1	1	5
Set E	20%	-	1	1	-	-	2	4
Total	100%	2	3	3	6	2	4	20

i. The first consideration in the development of Test Blue –Print is the weight to be assigned to higher order questions and the lower order questions (That is, to educational objectives at higher and at

lower cognitive levels). This is utilized in the allocation of numbers of questions to be developed in each cell under content and objective dimensions. In the hypothetical case under consideration, the level of difficulty for lower order questions (range: knowledge to application) is 40% while the higher order questions (range: analysis to evaluation) is 60%. This means that 40% of the total questions should be lower order questions while 60% of the questions are higher order questions. The learners in this case are assumed to be at the Senior Secondary Level of Education. Also, an attempt should be made as in the above to ensure that the questions are spread across all the levels of Bloom's (1956) Taxonomy of Educational Objectives.

ii. The blue-print is prepared by drawing 2-dimensional framework with the list of contents vertically (left column) and objectives horizontally (top row) as shown in table 3.1 above.

iii. Weights are assigned in percentages to both content and objectives dimensions as desired and as already stated earlier.

iv. Decisions on the number of items to be set and used are basis for determining items for each content area. For instance, in table 3-1, set A is weighted 15% and 20 items are to be generated in all. Therefore, total number of items for set A is obtained thus:

- Set A, weight: 15% of 20 items = 3 items
- Set B, weight: 15% of 20 items = 3 items
- Set C, weight: 25% of 20 items = 5 items
- Set D, weight: 25% of 20 items = 5 items
- Set E, weight: 20% of 20 items = 4 items.

The worked out values are then listed against each content area at the extreme right (Total column) to correspond with its particular content.

v. The same procedure is repeated for the objective dimension. Just like in the above.

- Knowledge: weight 10% of 20 items = 2 items
- Comprehension: weight 15% of 20 items = 3 items
- Application: weight 15% of 20 items = 3 items
- Analysis: weight 30% of 20 items = 6 items
- Synthesis: weight 10% of 20 items = 2 items
- Evaluation: weight 20% of 20 items = 4 items.

Here also the worked out values are listed against each objective at the last horizontal row, alongside the provision for total.

vi. Finally, the items for each content are distributed to the relevant objectives in the appropriate cells. This has also been indicated in the table 3.1 above. The Table of Specification now completed, serves as a guide for constructing the test items. It should be noted that in the table knowledge, comprehension and application levels have 2, 3, and 3 items respectively. That is, 2+3+3 = 8 items out of 20 items representing 40% of the total test items. While analysis, synthesis and evaluation have 6, 2 and 4 items respectively. That is, 6+2+4 = 12 items out of 20 items representing 60% of the total items.

vii. The development of table of specification is followed by item writing. Once the table of specification is adhered to in the item writing, the item would have appropriate *content validity* at the required level of difficulty. The table of specification is applicable both for writing essay items (subjective questions) and for writing objective items (multiple choice questions, matching sets items, completion items, true/false items).

3.2 Item Writing

The next task in planning the classroom test is to prepare the actual test items. The following is a guide for item writing:

i. Keep the test blueprint in mind and in view as you are writing the test items. The blueprint represents the master plan and should readily guide you in item writing and review.
ii. Generate more items than specified in the table of specification. This is to give room for item that would not survive the item analysis hurdles.
iii. Use unambiguous language so that the demands of the item would be clearly understood.

iv. Endeavor to generate the items at the appropriate levels of difficulty as specified in the table of specification. You may refer to Bloom (1956) taxonomy of educational objectives for appropriate action verb required for each level of objective.
v. Give enough time to allow an average student to complete the task.
vi. Build in a good scoring guide at the point of writing the test items.
vii. Have the test exercises examined and critiqued by one or more colleagues. Then subject the items to scrutiny by relevant experts.

The experts should include experts in measurement and evaluation and the specific subject specialist. Incorporate the critical comments of the experts in the modification of the items.

viii. Review the items and select the best according to the laid down table of specification/test blue print.

Also associated with test development is the statistical analysis – The Item analysis. This is used to appraise the effectiveness of the individual items.

Another important factor is reliability analysis. Both item analysis and reliability analysis would be treated in subsequent units. The item analysis and validity are determined by trail testing the developed items using a sample from the population for which the test is developed.

3.3 Moderation of Test Items

As earlier mentioned, one of the important stages of item development is to have test exercise examined and critiqued by one or more colleagues and experts. This process is known as "moderation" of items. After the item development phase, the test items are moderated by an expert or panel of experts before using them especially for school wide or class wide test such as end of term test. Before sending the items to external moderators (assessors), the items are to be given first to the subject head who should read through the items. Make intelligent input and some modifications on areas of need identified. The subject expert may also deem it necessary to engage others in the department who are knowledgeable in hat discipline to carry out similar exercise (subject experts' validation)before selecting the most appropriate one for external assessors (subject specialists and evaluation experts) to make final input before use.

The marking scheme and the marks allocated to various sections of the content covered should be sent along the test items to the external assessors. When this process is effectively carried out, the resulting items (items that survived the hurdles of moderation exercise) would have face, construct and content validity as test measuring instrument.

3.4 Assembling the Test items

You are to assemble the test items for use after the moderation process. The following are guides to enable you prepare and assemble both essay and objective test items for use.

3.4.1 Guide for Preparing the Objective Test Items for Use:

- Arrange the items on the test so that they are easy to read.
- Plan the layout of the test in such a way as to be convenient for recording answers and also scoring of the test items on separate answer sheets.
- Group items of the same format (true-false, multiple choice, matching items, completion items) together with their relevant directions on what need to be done by the testees.
- Group items dealing with the same content together within item types.
- Arrange the test items in progressive order of difficulty starting from simple to complex questions.
- Ensure that one item does not provide clues to the answer of another item or items in the same or another section of the test.
- Ensure that the correct responses form essentially a random pattern and in each of the possible response positions about the same percentage of the time for multiple choice items.

3.4.2 Guide for Preparing Essay Test for Use.

- The test items should not be too many or two lengthy for the testees to answer in the time available
- Ensure that a range of complexity and difficulty are in the test items especially when several essay items are given.
- It is preferable to give all testees same type of essay questions to answer in classroom test.
- Write a set of general directions for the test.
- Specify the point value for each question on the test.

Once you are through with the preparation of the test items, you now assemble the test items in the desired format for the testing exercise. The test items are now ready in the usable format for test administration.

SELF-ASSESSMENT EXERCISE

i. Identify five important considerations in planning a classroom test.
ii. Why is it necessary to prepare a table of specification or test blue print before writing test items?

SELF-ASSESSMENT EXERCISE

A1**:** Nine important considerations in planning a classroom test are:

- Determining the purpose of the test.
- Description of the Instructional Objectives and Content to be measured.
- Determination of the relative emphasis to be given to each learning outcome.
- Selection of the most appropriate test format(s)
- Development of test blue print to guide the test construction.
- Preparation of test items that are relevant to the learning outcomes specified in the test plan.
- Decision on the pattern of scoring and the interpretation of result.
- Decision on the length and direction of the test.
- Assembling the items for a test, preparation of direction and administration of the test.

A2: It is necessary to prepare test blue print before writing test items in order to ensure content validity of the test items.

4.0 CONCLUSION

In this unit, you learned the important considerations in planning the classroom test. You also learn about instructional objectives and content survey as relevant factors in test development. Furthermore, you learned how to plan the table of specification or test blue print. In addition, you learned about the guide for item writing and moderation of test items. Finally, you learned how to prepare objective and essay test items for use.

5.0 SUMMARY

- Planning classroom test entails designing a framework that can guide the test developers in the items development process.
- The outline of such a framework takes into account the content, the instructional objective and the test blue print.
- The content and instructional objectives are given relevant weight in planning the table of specification. This weighting depends on the importance and emphasis given to that content and the instructional objective

6.0 TUTOR-MARKED ASSIGNMENT

1. Enumerate what you may consider as guide in planning a classroom test.
2. Enumerate the guides for good item writing
3. Why is it necessary to prepare a table of specification or test blue print before writing test items?

7.0 REFERENCES/FURTHER READING

Bloom, B. (ed) (1956). Taxonomy of educational objectives, Handbook 1: Cognitive Domain. New York: Longmans, Green & Co.

Gronlund N.E (1985). Measurement and Evaluation in Teaching. New York: Macmillan Publishing Company.

Nenty, H.J (1985), Fundamentals of Measurement and Evaluation in Education. Calabar: University of Calabar, Press

Thorndike R.L and Hagen, E.P (1977). Measurement and Evaluation in Psychology and Education 4e. New York: John Wiley & Sons.

UNIT 5 THE ADMINISTRATION AND SCORING OF CLASSROOM TEST

CONTENTS

1.0 INTRODUCTION

In the last two units you learned about test. Here you will learn test administration and scoring of classroom test. You will learn how to ensure quality in test administration as well as credibility and civility in test administration. Further more you will learn how to score essay and objective test items using various methods.

2.0 OBJECTIVES

By the time you finish this unit you will be able to:

- Explain the meaning of test administration
- State the steps involved in test administration
- Identify the need for civility and credibility in test administration
- State the factors to be considered for credible and civil test administration
- Score both essay and objective test using various methods.

3.0 MAIN CONTENT

3.1 Test Administration

Test Administration as you know refers to the procedure of actually presenting the learning task that the examinees are required to perform in order to ascertain the degree of learning that has taken place during

the teaching-learning process. This procedure is as important as the process of preparing the test. This is because the validity and reliability of test scores can be greatly reduced when test is poorly administered. While administering test all examinees must be given fair chance to demonstrate their achievement of the learning outcomes being measured. This requires the provision of a physical and psychological environment which is conducive to their making their best efforts and the control of such factors such as malpractices and unnecessary threat from test administrators that may interfere with valid measurement.

3.1.1 Ensuring Quality in Test Administration

Quality and good control are necessary components of test administration. The following are guidelines and steps involved in test administration aimed at ensuring quality in test administration.

- Collection of the question papers in time from custodian to be able to start the test at the appropriate time stipulated.
- Ensure compliance with the stipulated sitting arrangements in the test to prevent collision between or among the testees.
- Ensure orderly and proper distribution of questions papers to the testees.
- Do not talk unnecessarily before the test. Testees' time should not be wasted at the beginning of the test with unnecessary remarks, instructions or threat that may develop test anxiety.
- It is necessary to remind the testees of the need to avoid malpractices before they start and make it clear that cheating will be penalized.
- Stick to the instructions regarding the conduct of the test and avoid giving hits to testees who ask about particular items. But make corrections or clarifications to the testees whenever necessary.
- Keep interruptions during the test to a minimum.

3.1.2 Credibility and Civility in Test Administration

Credibility and Civility are aspects of characteristics of assessment which have day to day relevance for developing educational communities. Credibility deals with the value the eventual recipients and users of the results of assessment place on the result with respect to the grades obtained, certificates issued or the issuing institution. While civility on the other hand enquires whether the persons being assessed are in such conditions as to give their best without hindrances and encumbrances in the attributes being assessed and whether the exercise is seen as integral to or as external to the learning process. Hence, in test administration, effort should be made to see that the testees are given a

fair and unaided chance to demonstrate what they have learnt with respect to:

i. Instructions: Test should contain a set of instructions which are usually of two types. One is the instruction to the test administrator while the other one is to the testee. The instruction to the test administrator should explain how the test is to be administered the arrangements to be made for proper administration of the test and the handling of the scripts and other materials. The instructions to the administrator should be clear for effective compliance. For the testees, the instruction should direct them on the amount of work to be done or of tasks to be accomplished. The instruction should explain how the test should be performed. Examples may be used for illustration and to clarify the instruction on what should be done by the testees. The language used for the instruction should be appropriate to the level of the testees. The necessary administrators should explain the testees instruction for proper understanding especially when the ability to understand and follow instructions is not part of the test.

ii. Duration of the Test: The time for accomplishing the test is technically important in test administration and should be clearly stated for both the test administrators and testees. Ample time should be provided for candidates to demonstrate what they know and what they can do. The duration of test should reflect the age and attention span of the testees and the purpose of the test.

iii. Venue and Sitting Arrangement: The test environment should be learner friendly with adequate physical conditions such as work space, good and comfortable writing desks, proper lighting, good ventilation, moderate temperature, conveniences within reasonable distance and serenity necessary for maximum concentration. It is important to provide enough and comfortable seats with adequate sitting arrangement for the testees' comfort and to reduce collaboration between them. Adequate lighting, good ventilation and moderate temperature reduce test anxiety and loss of concentration which invariably affects performance in the test. Noise is another undesirable factor that has to be adequately controlled both within and outside the test immediate environment since it affects concentration and test scores.

iv. Other necessary conditions: Other necessary conditions include the fact that the questions and questions paper should be friendly with bold characters, neat, decent, clear and appealing and not such that intimidates testee into mistakes. All relevant materials for carrying out the demands of the test should be provided in reasonable number, quality and on time.

All these are necessary to enhance the test administration and to make assessment civil in manifestation.

On the other hand for the credibility, effort should be made to moderate the test questions before administration based on laid down standard. It is also important to ensure that valid questions are constructed based on procedures for test construction which you already know as we have earlier discussed this in units 2 and 3 of this module. Secure custody should be provided for the questions from the point of drafting to constituting the final version of the test, to provision of security and safe custody of live scripts after the assessment, transmitting then to the graders and provision of secure custody for the grades arising from the assessment against loss, mutilation and alteration. The test administrators and the graders should be of proven moral integrity and should hold appropriate academic and professional qualifications. The test scripts are to be graded and marks awarded strictly by using itemized marking schemes. All these are necessary because an assessment situation in which credibility is seriously called to question cannot really claim to be valid.

3.2 Scoring the Test

In the evaluation of classroom learning outcomes marking schemes are prepared alongside the construction of the test items in order to score the test objectively. The marking scheme describes how marks are to be distributed amongst the questions and between the various parts of the question. This distribution is dependent on the objectives stated for the learning outcome during teaching and the weight assigned to the questions during test preparation and construction of the test item. The marking scheme takes into consideration the facts required to answer the questions and the extent to which the language used meets the requirement of the subject. The actual marking is done following the procedures for scoring essay questions (for essay questions) and for scoring objective items (for objective items)

3.2.1 Scoring Essay Test

As you are already aware the construction and scoring of essay questions are interrelated processes that require attention if a valid and reliable measure of achievement is to be obtained. In the essay test the examiner is an active part of the measurement instrument. Therefore, the viabilities within and between examiners affect the resulting score of examinee. This variability is a source of error, which affects the reliability of essay test if not adequately controlled. Hence, for the essay test result to serve useful purpose as valid measurement instrument conscious effort is made to score the test objectively by using

appropriate methods to minimize the effort of personal biases and idiosyncrasies on the resulting scores; and applying standards to ensure that only relevant factors indicated in the course objectives and called for during the test construction are considered during the scoring. There are two common methods of scoring essay questions. These are:

The Point or Analytic Method

In this method each answer is compared with already prepared ideal marking scheme (scoring key) and marks are assigned according to the adequacy of the answer. When used conscientiously, the analytic method provides a means for maintaining uniformity in scoring between scorers and between scripts thus improving the reliability of the scoring.

This method is generally used satisfactorily to score Restricted Response Questions. This is made possible by the limited number of characteristics elicited by a single answer, which thus defines the degree of quality precisely enough to assign point values to them. It is also possible to identify the particular weakness or strength of each examinee with analytic scoring. Nevertheless, it is desirable to rate each aspect of the item separately. This has the advantage of providing greater objectivity, which increases the diagnostic value of the result.

The Global/Holistic of Rating Method

In this method the examiner first sorts the response into categories of varying quality based on his general or global impression on reading the response. The standard of quality helps to establish a relative scale, which forms the basis for ranking responses from those with the poorest quality response to those that have the highest quality response. Usually between five and ten categories are used with the rating method with each of the piles representing the degree of quality and determines the credit to be assigned. For example, where five categories are used, and the responses are awarded five letter grades: A, B, C, D and E. The responses are sorted into the five categories where A -quality responses, B – quality, C – quality D- quality and E-quality. There is usually the need to re-read the responses and to re-classify the misclassified ones.

This method is ideal for the extended response questions where relative judgments are made (no exact numerical scores) concerning the relevance of ideas, organization of the material and similar qualities evaluated in answers to extended response questions. Using this method requires a lot of skill and time in determining the standard response for each quality category. It is desirable to rate each characteristic separately. This provides for greater objectivity and

increases the diagnostic value of the results. The following are procedures for scoring essay questions objectively to enhance reliability.

i. Prepare the marking scheme or ideal answer or outline of expected answer immediately after constructing the test items and indicate how marks are to be awarded for each section of the expected response.
ii. Use the scoring method that is most appropriate for the test item. That is, use either the analytic or global method as appropriate to the requirements of the test item.
iii. Decide how to handle factors that are irrelevant to the learning outcomes being measured. These factors may include legibility of handwriting, spelling, sentence structure, punctuation and neatness. These factors should be controlled when judging the content of the answers. Also decide in advance how to handle the inclusion of irrelevant materials (uncalled for responses).
iv. Score only one item in all the scripts at a time. This helps to control the "halo" effect in scoring.
v. Evaluate the answers to responses anonymously without knowledge of the examinee whose script you are scoring. This helps in controlling bias in scoring the essay questions.
vi. Evaluate the marking scheme (scoring key) before actual scoring by scoring a random sample of examinees actual responses. This provides a general idea of the quality of the response to be expected and might call for a revision of the scoring key before commencing actual scoring.
vii. Make comments during the scoring of each essay item. These comments act as feedback to examinees and a source of remediation to both examinees and examiners.
viii. Obtain two or more independent ratings if important decisions are to be based on the results. The result of the different scorers should be compared and rating moderated to reflect the discrepancies for more reliable results.

3.2.2 Scoring Objective Test.

Objective test can be scored by various methods with ease unlike the essay test. Various techniques are used to speed up the scoring and the techniques to use sometimes depend on the type of objective test. Some of these techniques are as follows:

i. Manual Scoring

In this method of scoring the answer to test items are scored by direct comparison of the examinees answer with the marking key. If the answers are recorded on the test paper for instance, a scoring key can be

made by marking the correct answers on a blank copy of the test . Scoring is then done by simply comparing the columns of answers on the master copy with the columns of answers on each examinee's test paper. Alternatively, the correct answers are recorded on scripts of paper and this script key on which the column of answers are recorded are used as master for scoring the examinees test papers.

ii. Stencil Scoring

On the other hand, when separate sheet of answer sheets are used by examinees for recording their answers, it's most convenient to prepare and use a scoring stencil. A scoring stencil is prepared by pending holes on a blank answer sheet where the correct answers are supposed to appear. Scoring is then done by laying the stencil over each answer sheet and the number of answer checks appearing through the holes is counted. At the end of this scoring procedure, each test paper is scanned to eliminate possible errors due to examinees supplying more than one answer or an item having more than one correct answer.

iii. Machine Scoring

Usually for a large number of examinees, a specially prepared answer sheets are used to answer the questions. The answers are normally shaded at the appropriate places assigned to the various items. These special answer sheets are then machine scored with computers and other possible scoring devices using certified answer key prepared for the test items.

In scoring objective test, it is usually preferable to count each correct answer as one point. An examinee's score is simply the number of items answered correctly. Sometimes examiners may prefer to correct for guessing. To do this the following formula may be used.

Correction for Guessing

The most common correction – for – guessing formula although rarely used is:

$$\text{Score} = \text{Right} - \frac{\text{wrong}}{n - 1}$$

Where n is the number of alternatives for an item.

SELF-ASSESSMENT EXERCISE

1. What is test administration?
2. What are the steps involves in test administration?
3. Why is it necessary to ensure credibility and civility in test administration?

4. List the methods for scoring
   i. Essay test item
   ii. Objective test items.

SELF-ASSESSMENT EXERCISE

A1: Test administration refers to the procedure of actually presenting the learning tasks that the examiners are required to perform to them after the teaching learning process.

A2: The steps involved in test administration are as discussed in section 3.1 of this unit.

A3: Credibility and Civility in Test Administration are aspects of characteristics of assessment, which have day-to-day relevance for developing educational communities. Hence, it is necessary to ensure credibility and Civility in test administration because they contribute significantly in enhancing the reliability and validity of test scores.

A4: i. The methods of Scoring Essay test items are:
   - By using Analytic Method
   - By using Rating Method

   ii. The methods of Scoring Objective test items are:
   - By Manual Scoring
   - By Stencil Scoring
   - By Machine Scoring

4.0 CONCLUSION

In this unit, you learned about test administration and scoring of test items. You learned how to ensure Quality in test administration and the need for credibility and civility in test administration. Further more you learned about scoring the test. Specifically you learned how to score essay and objective test objectively.

5.0 SUMMARY

- Test Administration refers to the procedure of presenting the learning tasks that the examinees are required to perform in order to ascertain the degree of learning that has taken place during the teaching – learning process.
- Credibility and Civility in Test Administration are aspects of characteristics of assessment, which have day-to-day relevance for developing educational communities.

- Credibility deals with the value the eventual recipients and users of the result of assessment place on the result with respect to grades obtained.
- Civility enquires whether the person being assessed is in such conditions as to give their best without hindrances and encumbrances in the attributes being assessed.
- Scoring test involves the preparation of marking scheme which describes how marks are to be distributed amongst the questions and between the various parts of the question.
- There are two methods of scoring essay questions:

Analytic method in which each answer is compared with already prepared ideal marking scheme (scoring key) and marks are assigned according to the adequacy of the answer. When used conscientiously, it provides a means for maintaining uniformity in scoring between scorers and between scripts and thereby improving the reliability of the scoring.

Rating Method is where the examiner first sort the responses into categories of varying quality based on his general or global impression on reading the responses. The standard of quality helps to establish a relative scale which forms the basis for ranking response from those with the poorest quality to those that have the highest quality response.

Using this method requires a lot of skill and time in determining the standard response for each quality category

- Methods of Scoring Objective Test are by Manual Scoring, Stencil Scoring and Machine Scoring.
- The Correction formula for guessing in Objective test is given by:

$$\text{Score} = \text{Right} - \frac{\text{Wrong}}{n - 1}$$

Where n is the number of alternatives for an item.

6.0 TUTOR-MARKED ASSIGNMENT

1. Explain the meaning of test administration.
2. Why is test administration important in the teaching learning process?
3. What are the steps to be taken in test administration to ensure quality?
4. Define Civility and Credibility with respect to test administration and state why it is necessary to ensure credibility and civility in test administration?

7.0 REFERENCES/FURTHER READING

Gronlund, N.E (1985). Measurement and Evaluation in Teaching New York: Macmillan Publishing Company.

Nenty, H.J (1985), Fundamentals of Measurement and Evaluation in Education. Calabar: University of Calabar,

MODULE 4

UNIT 1 JUDGING THE QUALITY OF A CLASSROOM TEST

CONTENTS

1.0 INTRODUCTION

In this unit you will learn how to judge the quality of a classroom test. Specifically, you will learn about item analysis – purpose and uses. Furthermore you will learn the process of item analysis for Norm-referenced classroom test and the computations involved. In addition you will learn item analysis of Criterion-referenced mastery items. Finally, you will learn about building a test item file.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- define and differentiate distinctively between item difficulty, item discrimination and the distraction power of an option
- recognize the need for item analysis, its place and importance in test development
- conduct item analysis of a classroom test
- calculate the value of each item parameter for different types of items
- appraise an item based on the results of item analysis.

3.0 MAIN CONTENT

3.1 Judging the Quality of a Classroom Test

The administration and scoring of a classroom test is closely followed by the appraisal of the result of the test. This is done to obtain evidence concerning the quality of the test that was used such as identifying some of the defective items. This helps to better appreciate the careful planning and hard work that went into the preparation of the test. Moreover, the identified effective test items are used to build up a file of high quality items (usually called question bank) for future use.

3.1.1 Item Analysis

Item analysis is the process of "testing the item" to ascertain specifically whether the item is functioning properly in measuring what the entire test is measuring. As already mentioned, item analysis begins after the test has been administered and scored. It involves detailed and systematic examination of the testees' responses to each item to determine the difficulty level and discriminating power of the item.

This also includes determining the effectiveness of each option. The decision on the quality of an item depends on the purpose for which the test is designed. However, for an item to effectively measure what the entire test is measuring and provide valid and useful information, it should not be too easy or too difficult. Moreover, its options should discriminate validity between high and low performing learners in the class.

3.1.2 Purpose and Uses of Item Analysis

Item analysis is usually designed to help determine whether an item functions as intended with respect to discriminating between high and low achievers in a norm-referenced test, and measuring the effects of the instruction in a criterion referenced test items. It is also a means of determining items having desirable qualities of a measuring instrument, those that need revision for future use and even for identifying deficiencies in the teaching/learning process. In addition, item analysis has other useful benefits amongst which are providing data on which to base discussion of the test results, remediation of learning deficiencies and subsequent improvement of classroom instruction. Moreover, the item analysis procedures provide a basis for increase skill in test construction.

3.2 The Process of Item Analysis for Norm – Referenced Classroom Test

The method for analyzing the effectiveness of test items differs for Norm-referenced and Criterion–referenced test items. This is because they serve different functions. In Norm-referenced test, special emphasis is placed on item difficulty and item discriminating power. The process of item analysis begins after the test has been administered (or trial tested), scored and recorded. For most Norm – referenced classroom tests, a simplified form of item analysis is used.

The process of Item Analysis is carried out by using two contracting test groups composed from the upper and lower 25% or 27% of the testees on which the items are administered or trial tested. The upper and lower 25% is the optimum point at which balance is obtained between the sensitivity of the groups in making adequate differentiation and reliability of the results for a normal distribution. On the other hand, the upper and lower 27% when used are better estimate of the actual discrimination value. They are significantly different and the middle values do not discriminate sufficiently. In other to get the groups, the graded test papers are arranged from the highest score to the lowest score in a descending order. The best 25% or 27% are picked from the top and the poorest 25% or 27% from the bottom while the middle test papers are discarded.

To illustrate the method of item analysis using an example with a class of 40 learners taking a 10 item test that have been administered and scored, and using 25% test groups. The item analysis procedure might follow this basic step.

i. Arrange the 40 test papers by ranking them in order from the highest to the lowest score.
ii. Select the best 10 papers (upper 25% of 40 testees) with the highest total scores and the least 10 papers (lower 25% of 40 testees) with the lowest total scores.
iii. Drop the middle 20 papers (the remaining 50% of the 40 testees) because they will no longer be needed in the analysis.
iv. Draw a table as shown in table 3.1 in readiness for the tallying of responses for item analysis.
v. For each of the 10 test items, tabulate the number of testees in the upper and lower groups who got the answer right or who selected each alternative (for multiple choice items).
vi. Compute the difficulty of each item (percentage of testees who got the item right).
vii. Compute the discriminating power of each item (difference between the number of testees in the upper and lower groups who got the item right).
viii. Evaluate the effectiveness of the distracters in each item (attractiveness of the incorrect alternatives) for multiple choice test items.

Table 3.1 Format for Tallying Responses for Item Analysis

Item No.	Testees	Alternatives with Correct Option Starred A B C D E Omit	Total	P-Value	D-Value	Option Distracter Index A B C D E
1	Upper 25% Lower 25%	0 10* 0 0 0 0 0 2 4* 1 3 0	10 10	0·70	0·60	0·20 * 0·10 0·30 0·00
2	Upper 25% Lower 25%	1 1 0 7* 1 0 1 2 1 4* 1	10 10			
3	Upper 25% Lower	3 0 1 2 4* 0	10 10			

	25%	1 0 1 3 5* 0				
4	Upper 25% Lower 25%	0 0 10* 0 0 0 0 0 10* 0 0 0	10 10			
5	Upper 25% Lower 25%	2 3 3 1 1* 0 3 3 1 2 1* 0	10 10			
• • •						
10	Upper 25% Lower 25%	6* 1 1 1 0 1 3* 2 2 2 1 0	10 10			

3.2.1 Computing Item Difficulty

The difficulty index P for each of the items is obtained by using the formula:

Item Difficulty (P) $= \frac{\text{Number of testees who got item right (T)}}{\text{Total number of testees responding to item (N)}}$

i.e. $P = T/N$

Thus for item I in table 3.1,

$$P = \frac{14}{20} = 0.7$$

The item difficult indicates the percentage of testees who got the item right in the two groups used for the analysis. That is 0.7 x 100% = 70%.

3.2.2 Computing Item Discriminating Power (D)

Item discrimination power is an index which indicates how well an item is able to distinguish between the high achievers and low achievers given what the test is measuring. That is, it refers to the degree to which it discriminates between testees with high and low achievement. It is obtained from this formula:

$$\text{Item Discrimination Power (D)} = \frac{\text{Number of high scorers who got items right (H)} - \text{Number of low scorers who got item right (L)}}{\text{Total Number in each group (n)}}$$

That is,

$$D = \frac{H - L}{n}$$

Hence for item 1 in table 3.1, the item discriminating power D is obtained thus:

$$D = \frac{H - L}{n} = \frac{10-4}{10} = \frac{6}{10} = 0{\cdot}60$$

Item discrimination values range from – 1·00 to + 1·00. The higher the discriminating index, the better is an item in differentiating between high and low achievers.

Usually, if item discriminating power is a:

- positive value when a larger proportion of those in the high scoring group get the item right compared to those in the low scoring group.
- negative value when more testees in the lower group than in the upper group get the item right.
- zero value when an equal number of testees in both groups get the item right; and
- 1.00 when all testees in the upper group get the item right and all the testees in the lower group get the item wrong.

3.2.3 Evaluating the Effectiveness of Distracters

The distraction power of a distractor is its ability to differentiate between those who do not know and those who know what the item is measuring. That is, a good distracter attracts more testees from the lower group than the upper group. The distraction power or the effectiveness of each distractor (incorrect option) for each item could be obtained using the formula:

$$\text{Option Distractor Power (Do)} = \frac{\text{Number of low scorers who marked option (L)} - \text{Number of high scorers who marked option (H)}}{\text{Total Number in each group (n)}}$$

That is,

$$Do = \frac{L - H}{n}$$

For item 1 of table 3.1 effectiveness of the distracters are:

For option A: $Do = \frac{L - H}{n} = \frac{2 - 0}{10} = 0{\cdot}20$

B: The correct option starred (*)

C: $Do = \frac{L - H}{n} = \frac{1 - 0}{10} = 0{\cdot}10$

D: $Do = \frac{L - H}{n} = \frac{3 - 0}{10} = 0{\cdot}30$

E: $Do = \frac{L - H}{n} = \frac{0 - 0}{10} = 0{\cdot}00$

Incorrect options with positive distraction power are good distracters while one with negative distracter must be changed or revised and those with zero should be improved on because they are not good. Hence, they failed to distract the low achievers.

3.3 Item Analysis and Criterion – Referenced Mastery Tests

The item analysis procedures we used earlier for norm – referenced tests are not directly applicable to criterion – referenced mastery tests. In this case indexes of item difficulty and item discriminating power are less meaningful because criterion referenced tests are designed to describe learners in terms of the types of learning tasks they can perform unlike in the norm-referenced test where reliable ranking of testees is desired.

3.3.1 Item Difficulty

In the Criterion –Referenced Mastery Tests the desired level of item difficulty of each test item is determined by the learning outcome it is designed to measure and not as earlier stated on the items ability to discriminate between high and low achievers. However, the standard

formula for determining item difficulty can be applied here but the results are not usually used to select test items or to manipulate item difficulty. Rather, the result is used for diagnostic purposes. Also most items will have a larger difficulty index when the instruction is effective with large percentage of the testees passing the test.

3.3.2 Item Discriminating Power

As you know the ability of test items to discriminate between high and low achievers are not crucial to evaluating the effectiveness of criterion –referenced tests this is because some of the best items might have low or zero indexes of discrimination. This usually occurs when all testees answer a test item correctly at the end of the teaching learning process implying that both the teaching/learning process and the item are effective. Moreover, they provide useful information concerning the mastery of items by the testees unlike in the norm-referenced test where they would be eliminated for failing to eliminate between the high and the low achievers. Therefore, the traditional indexes of discriminating power are of little value for judging the test items quality since the purpose and emphasis of criterion –referenced test is to describe what learners can do rather than to discriminate among them.

3.3.3 Analysis of Criterion - Referenced Mastery Items

Ideally, a Criterion – referenced mastery test is analyzed to determine extent to which the test items measure the effects of the instruction. In other to provide such evidence, the same test items is given before instruction (pretest) and after instruction (posttest) and the results of the same test pre-and-post administered are compared. The analysis is done by the use of item response chart. The item response chart is prepared by listing the numbers of items across the top of the chart and the testees names / identification numbers down the side of the chart and the record correct (+) and incorrect (-) responses for each testee on the pretest (B) and the posttest (A). This is illustrated in Table 3.2 for an arbitrary 10 testees.

Table 3.2: An item – response chart showing correct (+) and incorrect (-) responses for pretest and post test given before (B) and after (A) instructions (Teaching / learning process) respectively.

Item	Testee Identification Number							Remark
	001	002	003	004	005	…	010	
Pretest (B) Posttest (A)	- +	- +	- +	- +	- +	… …	- +	Ideal

Pretest (B) Posttest (A)	+ +	+ +	+ +	+ +	+ +	… …	+ +	Too easy
Pretest (B) Posttest (A)	- -	- -	- -	- -	- -	… …	- -	Too difficult
Pretest (B) Posttest (A)	+ -	+ -	+ -	+ -	+ -	… …	+ -	Defective
Pretest (B) Posttest (A)	- +	+ +	- +	- +	+ +	… …	- -	Effective

An index of item effectiveness for each item is obtained by using the formula for a measure of Sensitivity to Instructional Effects (S) given by

$$S = \frac{RA - RB}{T}$$

Where

R_A = Number of testees who got the item right after the teaching / learning process.

R_B = Number of testees who got the item right before the teaching / learning process.

T = Total number of testees who tried the item both times.

For example, item 1 of table 3.2, the index of sensitivity to instructional effect (S) is

$$S = \frac{RA - RB}{T} = \frac{10-0}{10} = 1{\cdot}00$$

Usually for a criterion-referenced mastery test with respect to the index of sensitivity to instructional effect:

- an ideal item yields a value of 1.00.
- effective items fall between 0.00 and 1.00, the higher the positive value, the more sensitive the item to instructional effects; and
- items with zero and negative values do not reflect the intended effects of instruction.

3.3.4 Effectiveness of Distracters

In a criterion-referenced test, it is important to note how well each alternative function in a multiple – choice item. Ideally, testees should

choose one of the incorrect alternatives if they have not achieved the objective that the test item measures. This is done by checking the frequency with which those failing an item select each distracter. This type of analysis is best done on the pretest in which a relatively large proportion of pupils can be expected to fail the items. However, items containing distracters that are not selected at all or rarely selected need to be revised

3.4 Building a Test Item File (Item Bank)

This entails a gradual collection and compilation of items administered, analyzed and selected based on their effectiveness and psychometric characteristics identified through the procedure of item analysis over time. This file of effective items can be built and maintained easily by recording them on item card, adding item analysis information indicating both objective and content area the item measures and can be maintained on both content and objective categories. This makes it possible to select items in accordance with any table of specifications in the particular area covered by the file. Building item file is a gradual process that progresses over time. At first it seems to be additional work without immediate usefulness. But with time its usefulness becomes obvious when it becomes possible to start using some of the items in the file and supplementing them with other newly constructed ones. As the file grows into item bank most of the items can then be selected from the bank without frequent repetition. Some of the advantages of item bank are that:

- Parallel test can be generated from the bank which would allow learners who were ill for a test or due to some other reasons were unavoidable absent to take up the test later;
- They are cost effective since new questions do not have to be generated at the same rate from year to year;
- The quality of items gradually improves with modification of the existing ones with time; and
- The burden of test preparation is considerably lightened when enough high quality items have been assembled in the item bank.

SELF-ASSESSMENT EXERCISE

i. What is the difference between index of discriminating power (D) and index of sensitivity to instructional effects (S)?
ii. Compute (i) the item difficulty; (ii) the discriminating power and (iii) effectiveness of distracters for item 2 of table 3.1

SELF-ASSESSMENT EXERCISE

A1. The index of discriminating power (D) indicates the degree to which an item discriminates between high and low achiever on a single administration of the test while the index of sensitivity to instructional effects (S) indicates the degree to which an item reflects the intended effects of instruction determined based on pretest and posttest results. Although the same values are used to represent both discriminating power (D) and effective criterion – referenced test item (·00 to 1·00).

A2: (i) The item difficulty (P) for item 2 of table 3.1
Formula:
Item difficulty (P) = Number of testees who got item right (T)

Total number of testees responding to item (N)

$$P = \frac{T}{N} = \frac{11}{20} = 0{\cdot}55$$

(ii) The discriminating power (D) for item 2 of table 3.1
Formula:

Item Discrimination Power (D) = Number of high scorers who got items right (H) - Number of low scorers who got item right (L) / Total Number in each group (n)

$$D = \frac{H - L}{n} = \frac{7 - 4}{10} = \frac{3}{10} = 0{\cdot}30$$

(iii) The effectiveness of distracters for item 2 of table 3.1
Formula:

Option Distractor Power (Do) = Number of low scorers who marked option (L) - Number of high scorers who marked option (H) / Total Number in each group (n)

Option A: $Do = \frac{L - H}{N} = \frac{1 - 1}{10} = \frac{0}{10} = 0{\cdot}00$

$$B: Do = \frac{L - H}{n} = \frac{2 - 1}{10} = \frac{1}{10} = 0{\cdot}10$$

$$C: Do = \frac{L - H}{n} = \frac{1 - 0}{n} = \frac{1}{10} = 0{\cdot}10$$

D: The correct option starred.

$$E: Do = \frac{L - H}{n} = \frac{1 - 1}{n} = \frac{0}{10} = 0{\cdot}00$$

4.0 CONCLUSION

In this unit you learned how to judge the quality of classroom test by carrying out item analysis. You learn the purpose and uses of item analysis, the process of item Analysis for Norm-referenced classroom test and how to compute item difficulty, item discriminating power and evaluating the effectiveness of distracters. Furthermore, you learned item analysis and criterion-referenced mastery tests with respect to item difficulty, item discriminating power, analysis of criterion-referenced mastery items and effectiveness of distracters. Finally you learned how to build a Test Item File/Item Bank.

5.0 SUMMARY

- Item analysis is the process of "testing the item" to ascertain specifically whether the item is functioning properly in measuring what the entire test is measuring.
- The method for analyzing the effectiveness of test items differs for norm-referenced and criterion-referenced test items because they serve different functions.
- Item difficulty – The item difficulty (P) indicates the percentage of testees who get the item right.
- Item discriminating power (D) - This is an index which indicates how well an item is able to distinguish between the high achievers and low achievers given what the test is measuring.
- The Effectiveness of Distracters - The distraction power of a distracter is its ability to differentiate between those who do not know and those who know what the item is measuring.
- In the criterion – referenced mastery test the desired level of item difficulty of each test item is determined by the learning outcome it is designed to measure and not on the items ability to discriminate between high and low achievers.
- A criterion-referenced mastery test is analyzed to determine the extent to which the test items measure the effects of the

instruction. In doing this the same test is given before (pretest) and after instruction (posttest) and the results are compared using the item response chart.

6.0 TUTOR-MARKED ASSIGNMENT

1. What is Item Analysis?
2. Why is it necessary to carry out item analysis?
3. What is the difference between index of discriminating power (D) and index of sensitivity to instructional effects (S)?
4. List the basic steps in item analysis procedure.

7.0 REFERENCES /FURTHER READING

Gronlund, N.E (1985), Measurement and Evaluation in Teaching. New York. Macmillan Publishing Company.

Nenty, H.J (1985), Fundamentals of Measurement and Evaluation in Education. Calabar University of Calabar.

UNIT 2 INTERPRETING CLASSROOM TEST SCORES

CONTENTS

1.0 INTRODUCTION

In the previous unit you learned how to judge the quality of classroom test. In this unit, you will learn how to interpret test scores. You will start by learning the methods of interpreting test scores such as the criterion-referenced interpretation and norm-referenced interpretation. Furthermore, you will learn about Norms and the most common types of test norms which are grade norms, age norms and percentile norms. In addition you will learn about the standard scores which comprise the standard deviation and the normal curve, the Z- score, the T – score and the stanines. Finally, you will learn how to assign stanine to raw scores and how to compare scores.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- interpret classroom test scores by criterion-referenced or norm-referenced.

- use the common types of test norms such as grade norms, age norms, percentile norms in interpreting classroom test scores;
- convert raw scores to z-scores, T-scores and stanines scores.
- convert from one standard score to another and
- use the standard score in interpreting test scores.

3.0 MAIN CONTENT

3.1 Methods of Interpreting Test Scores

Test interpretation is a process of assigning meaning and usefulness to the scores obtained from classroom test. This is necessary because the raw score obtained from a test standing on itself rarely has meaning.

For instance, a score of 50% in one mathematics test cannot be said to be better than a score of 40% obtained by same testee in another mathematics test. The test scores on their own lack a true zero point and equal units. Moreover, they are not based on the same standard of measurement and as such meaning cannot be read into the scores on the basis of which academic and psychological decisions may be taken. To compensate for these missing properties and to make test scores more readily interpretable various methods of expressing test scores have been devised to give meaning to a raw score. Generally, a score is given meaning by either converting it into a description of the specific tasks that the learner can perform or by converting it into some type of derived score that indicates the learner's relative position in a clearly defined reference group. The former method of interpretation is referred to as criterion – referenced interpretation while the later is referred to as Norm – referenced interpretation.

3.1.1 Criterion – Referenced Interpretation

Criterion - referenced interpretation is the interpretation of test raw score based on the conversion of the raw score into a description of the specific tasks that the learner can perform. That is, a score is given meaning by comparing it with the standard of performance that is set before the test is given. It permits the description of a learner's test performance without referring to the performance of others. This is essentially done in terms of some universally understood measure of proficiency like speed, precision or the percentage correct score in some clearly defined domain of learning tasks. Examples of criterion-referenced interpretation are:

- Types 60 words per minute without error.
- Measures the room temperature within $\pm$ 0·1 degree of accuracy (precision).

- Defines 75% of the elementary concepts of electricity items correctly (percentage-correct score).

Such interpretation is appropriate for tests that are focused on a single objective and for which standards of performance can be either empirically or logically derived. The percentage-correct score is widely used in criterion-referenced interpretation. This type of interpretation is primarily useful in mastery testing where a clearly defined and delimited domain of learning tasks can be most readily obtained.

For Criterion-referenced test to be meaningful, the test has to be specifically designed to measure a set of clearly stated learning tasks. Therefore, in order to be able to describe test performance in terms of a learner's mastery or non mastery of the predefined, delimited and clearly specified task, enough items are used for each interpretation to enable dependable and informed decisions concerning the types of tasks a learner can perform.

3.1.2 Norm-Referenced Interpretation

Norm – referenced interpretation is the interpretation of raw score based on the conversion of the raw score into some type of derived score that indicates the learner's relative position in a clearly defined referenced group. This type of interpretation reveals how a learner compares with other learners who have taken the same test.

Norm – referenced interpretation is usually used in the classroom test interpretation by ranking the testees raw scores from highest to lowest scores. It is then interpreted by noting the position of an individual's score relative to that of other testees in the classroom test. The interpretation such as third position from highest position or about average position in the class provides a meaningful report for the teacher and the testees on which to base decision. In this type of test score interpretation, what is important is a sufficient spread of test scores to provide reliable ranking. The percentage score or the relative easy / difficult nature of the test is not necessarily important in the interpretation of test scores in terms of relative performance.

3.2 Norms – Most Common Types of Test Norm

Norms are reference frames on which interpretation of test scores are based. They represent the typical performance of the testees in the reference frame on which the test raw scores were standardized by the administration of the test to representative sample of the testees for whom the test was designed. The resulting test norm merely represents the typical performance of the participants and as such is not to be seen

as desired goals or standards. Nevertheless, the comparisons of the test scores with these reference frames make it possible to predict a learner's probable success in various areas like the diagnosis of strength and weakness, measuring educational growth and the use of the test results for other instructional guidance purposes. These functions of test scores derived from norm-referenced interpretation would have been limited without test norms.

The most common types of these reference frames referred to as norms are grade norms, age norms, percentile norms and standard score norms. These are informed by the general ways in which we may relate a person's score to a more general framework. One of these ways is to compare the individual learner with a graded series of groups and see which one he matches. Usually each group in this series represents a particular school grade or a particular chronological age. The other way is to determine where the individual learner falls in terms of the percentage of the group he surpasses or in terms of the groups mean and standard deviation. Hence they form main patterns for interpreting the score of an individual. A summary of these most common types of test norms is presented schematically in table 3.1

Table 3.1: Most Common Types of Test Norms for Education and Psychological Tests.

Type of Test Norm	Name of Derived Score	Type of Comparison	Type of Group	Meaning in Terms of Test Performance
Grade Norms	Grade equivalents	Individual matched to group whose performance he equals	Successive grade groups	Grade group in which the testee's raw score is average
Age Norms	Age equivalents	Individual matched to group whose performance he equals	Successive age group	Age group in which the testee's raw score is average
Percentile Norms	Percentile ranks (or percentile scores)	Percentage of group surpassed by individual	Single age or grade group to which individual belongs	Percentage of testee's in the reference group who fall below the testee's raw score
Standard Score Norms	Standard Scores	Number of standard deviations	Single age or grade group to	Distance of testee's raw score above or

		individual falls above or below average of group	which individual belongs	below the mean of the reference group in terms of standard deviation units.

3.2.1 Grade Norms

Grade norms are reference framework for interpreting the academic achievement of learners in the elementary schools. They represent the typical (average) performances of specific groups. Grades are like classes in our school system. Grades norms are prepared for traits that show a progressive and relatively uniform increase from one school grade (class) to the next higher grade. The norm for any grade is then the average score obtained by individuals in that grade. The process of establishing grade norms involves giving the test to a representative sample of pupils in each of a number of consecutive grades to evaluate the (mean) performance of individuals in their respective specific grades (classes) of the school system. This is achieved by limiting the content and objectives of the test to the content and objectives relevant to the class. The mean score obtained becomes the typical score of the grade (class) against which the performance of the members of the grade can be compared and are interpreted in terms of grade equivalents.

Grade equivalents are expressed in range of 10 consisting of two numbers. The first indicates the year and the second the month. For example, if a test is given in October to grade (class) 2 pupils in a school for a calendar year that runs September to June, the average (mean) of the raw scores obtained by the representative sample of these grades will be assigned a grade equivalent of 2.2 This is because the class is 2 and is obtained in the second month of the school 10 month year. The grade in this case ranges from 2·0 to 2·9

If the same test is given to grades (classes) 3, 4, 5 and 6 respectively at same period, the grade equivalents will be 3·2, 4·2, 5·2 and 6·2 respectively. Assuming the average scores of these grades are respectively 20, 23, 26, 28, 30. These means scores then become the typical scores of the grades (classes) respectively against which the performance of members of the grades (classes) can be compared and interpreted. The table of grade equivalent in this case will look like this:

Grade	Typical Score	Grade Equivalent
2	20	2.2
3	23	3.2
4	26	4.2

5	28	5.2
6	30	6.2

The grade equivalents and the typical scores indicate the average performance of pupils at the various grade levels. The in between scores are obtained by interpretation. If the test is given to learners in the same grade at any other time, their scores would be compared with the table and the grade equivalent would be read from the table on the basis of which result would be interpreted

You will however note that for any particular grade equivalent typical score 50% of the samples used are above while 50% are below since the typical score is an average of their performances. Hence, grade Norms have these limitations.

- The units are not equal on different parts of the scale from one test to another.
- The norm depends on
  - the ability of the pupils used in the preparation;
  - the extent to which the learning outcomes measured by test reflect the curriculum emphasis (content and objective).
  - the educational facilities at the users disposal.

Based on these limitations grade norm is not to be considered as a standard of excellence to be achieved by others. However, they are widely used at the elementary school level largely because of the apparent ease with which they can be interpreted.

3.2.2 Age Norms

Age Norms like the grade norms are based on the average scores earned by pupils at different ages and are interpreted in terms of age equivalents. A test may be prepared with a certain age or range of ages of testees in mind. Performance typical for specific ages in these tests and determined and the performance of an individual of the age group is compared against the typical performance. In this case, also the typical performance is determined by giving the test to a very large sample of the population for whom the test is meant and then after scoring, find the mean score of the large sample. This mean score then becomes the typical performance for the age. The performance of a pupil of this age in the test could then be interpreted to be higher or lower than or the same as the performance for the age. Table of age norms presents parallel columns of typical scores and their corresponding age equivalents.

Age Norms have essentially same characteristic and limitations as do the grade norms. The major differences between them are that

- Test performance is expressed in forms of age level rather than grade level;
- Age equivalent divide the calendar year into twelve parts rather than ten.

That is for example, the age equivalent for 12 years old ranges from 12·0 to 12·12 unlike in grade norm where it ranges from 12·0 to 12·9.

Like grade norms, age norms present test performance in units that are characteristically unequal although apparently easy to understand and are at such subject to misinterpretation especially at high school level.

Age norms are mostly used in elementary schools in areas of mental ability test, personality test, reading test and interest inventories where growth pattern tend to be consistent.

3.2.3 Percentile Norms

Percentile Norms are test norms that deal with percentile ranks or scores. They are used for comparison of percentage of group surpassed by individual in the single age or grade group to which individual belongs. Percentile norms are very widely adaptable and applicable and can be used whenever an appropriate normative group can be obtained to serve as a yardstick. They are appropriate for young learners in educational situations.

A percentile rank (or percentile score) describes a learner's performance in terms of the percentage of learners in some clearly defined group that earn a lower score. This might be a grade or age group or any other group that provides a meaningful comparison. That is, a percentile rank indicates a learner's relative position in a group in terms of the percentage of learners scoring lower. For instance, tables of norms with a learner's raw score of 29 equaling a percentile rank of 70 means that 70 percent of the learners in the reference group obtained a raw score lower than 29. In other words, this learner's performance surpasses that of 70 percent of the group. To surpass 90 percent of a reference comparison group signifies a comparable degree of excellence no matter the function being measured.

The limitation of percentile rank is that the units are typical and symmetrically unequal. This means that equal percentile differences do not in general represent equal differences in amount. Therefore, any interpretation of percentile ranks must take into account the fact that

such a scale has been pulled out at both ends and squeezed in the middle. Hence, they must be interpreted in terms of the norm group on which they are based. The limitation of unequal units can be offset by careful interpretation since the inequality of units follows a predictable pattern.

3.3 Standard Score

Standard score is a method of indicating a testee's relative position in a group by showing how far the raw score is above or below average. Standard scores express test performance in terms of standard deviation units from the mean. The mean (M) as you know is the arithmetic average obtained by adding all the scores and dividing by the number of scores while the standard deviation (SD) is a measure of the spread of scores in a group. The meaning of standard deviation and the standard scores based on it is explained in terms of the normal curve.

3.3.1 The Normal Curve and the Standard Deviation Unit.

The normal curve is a symmetrical bell-shaped curve that has many useful mathematical properties. When it is divided into standard deviation units, each portion under the curve contains a fixed percentage of cases under consideration. This division is very useful and is utilized in test interpretation. Usually, 34 percent of the cases under consideration fall between the mean and +1SD, 14 percent between +1SD and +2SD and 2 percent between +2SD and 3SD. These same proportions apply to the standard deviation intervals below the mean. About 0.13 percent of the cases fall below -3SD or above +3SD and are usually neglected in practice. That is, for all practical purposes, a normal distribution of scores falls between – 3 and +3 standard deviations from the mean. This is illustrated in figure 1 below.

0·13%	2%	14%	34%	34%	14%	2%	0·13%
Mean							

-4SD -3SD -2SD -1SD 0 +1SD +2 SD +3SD + 4SD

Figure 1: Normal curve indicating the approximate percentage of cases falling within each standard deviation interval.

Standard deviation enables the conversion of raw scores to a common scale that has equal units and that can be interpreted in terms of the normal curve.

The following characteristics possessed by normal curve makes it useful in test interpretation. These are:

- It provides a handy bench mark for interpreting scores and the standard error of measurement as both are based on standard deviation units.
- The fixed percentage in each interval makes it possible to convert standard deviation units to percentile ranks. For instance, -2SD equals a percentile rank of 2. This means that 2 percent of the cases under consideration fall below that point. Like wise, each point on the base line of the curve can be equated to percentile ranks starting from the left of the figure as:

  -2SD = 2%
  -1SD = 16% (2+14)
  0 (M) = 50% (16+34)
  +1SD = 84% (50 +34)
  +2SD = 98% (84+14)

The relationship for a normal curve between standard deviation units and percentile ranks makes it possible to interpret standard scores in simple and familiar terms.

3.3.2 The Z – Scores

The Z – score is the simple standard score which expresses test performance simply and directly as the number of standard deviation units a raw score is above or below the mean. The Z-score is computed by using the formula.

$$\text{Z – Score} = \frac{X - M}{SD}$$

Where
X = any raw score
M = arithmetic mean of the raw scores
SD = standard deviation

When the raw score is smaller than the mean the Z –score results in a negative (-) value which can cause a serious problem if not well

noted in test interpretation. Hence Z-scores are transformed into a standard score system that utilizes only positive values.

3.3.3 The T - Score

This refers to any set of normally distributed standard scores that has a mean score of 50 and a standard deviation of 10. The T – score is obtained by multiplying the Z-score by 10 and adding the product to 50. That is, T – Score = 50 + 10(z).

Example

A test has a mean score of 40 and a standard deviation of 4. What are the T – scores of two testees who obtained raw scores of 30 and 45 respectively in the test?

Solution

The first step in finding the T-scores is to obtain the z-scores for the testees. The z-scores would then be converted to the T – scores. In the example above, the z – scores are:

i. For the testee with raw score of 30, the Z – score is:
Z – Score = $\frac{X - M}{SD}$, where the symbols retain their usual meanings.
X = 30, M = 40, SD = 4. Substituting gives
Z – Score = $\frac{30 - 40}{4} = \frac{-10}{4} = -2{\cdot}5$
The T - Score is then obtained by converting the Z – Score (-2·5) to T – score thus:
T – Score = 50 + 10 (z)
= 50 + 10 (-2·5)
= 50 – 25
= 25

ii. For the testee with raw score of 45, the z – score is:
Z – Score = $\frac{X - M}{SD}$, where the symbols retain their usual meanings.
X = 45, M = 40, SD = 4. Substituting gives
Z – Score = $\frac{45 - 40}{4} = \frac{-15}{4} = -3{\cdot}75$
The T-Score conversion is:
T – Score = 50 + 10 (z)
= 50 + 10 (3·75)

$$= \quad 50 + 37{\cdot}5$$
$$= \quad 87{\cdot}5$$

The resulting T-Score are easily interpreted since T – Score always have a mean of 50 and a standard deviation of 10.

3.3.3 Stanine Norm

This is a single – digit standard scores which assumes that human beings are normally distributed in their possession of most psychological characteristics. This system of score is so named because the distribution of raw scores is divided into nine parts (standard nine). The stanine, which is a kind of standard score which divides a population according to some fixed proportions into nine parts, numbered 1 to 9 has a mean score of 5 and a standard deviation of 2. Each stanine corresponds to a score or a range of scores. Each individual's score falls within a stanine and such score can be described by reference to the stanine within which if falls. Stanines are widely used for local norms because of the ease with which they can be computed and interpreted. The strength of stanine norms include:

i. The use of nine point scale in which 9 is high, 1 is low and 5 is average or 1 is high, 9 is low and 5 is average. This later usage is employed in the Ordinary Level School Certificate Examination Result. This is illustrated below:

Stanine	Letter Grade	Remark
1	A1	Excellent
2	A2	Very Good
3	A3	Good
4	C4	Credit
5	C5	Credit
6	C6	Credit
7	P7	Pass
8	P8	Pass
9	F9	Fail

ii. Stanines are normalized standard scores. It is therefore possible to compare an individual's performance on different tests especially when the tests are based on a common group. An example is the comparison of West African Senior School Certificate Results with the National Examination Council Senior School Certificate results for a group of examinees. A difference of 2 stanine represents a significant difference in test performance between tests.

iii. The system makes it possible to readily combine diverse types of data.
iv. It uses a single-digit system which makes it easy to record.

i. It also takes up less space than other scores.

The limitation of stanine is that it plays down small differences in scores and expresses performance in broad categories so that attention tends to be focused in differences that are big enough to make a difference.

3.3.4 Assigning Stanine to Raw Scores

To assign stanines to raw scores, scores are ranked from high to low and frequency distribution table showing the cumulative frequency column is used in the construction of the staine table. The following guide is then employed to assign stanines to the raw score.

Guide for Assigning Stanines to Raw Scores

- Top 4% of the raw score are assigned a stanine score of 9
- Next 7% of the raw score are assigned a stanine score of 8
- Next 12% of the raw score are assigned a stanine score of 7
- Next 17% of the raw score are assigned a stanine score of 6
- Next 20% of the raw score are assigned a stanine score of 5
- Next 17% of the raw score are assigned a stanine score of 4
- Next 12% of the raw score are assigned a stanine score of 3
- Next 7% of the raw score are assigned a stanine score of 2
- Next 4% of the raw score are assigned a stanine score of 1

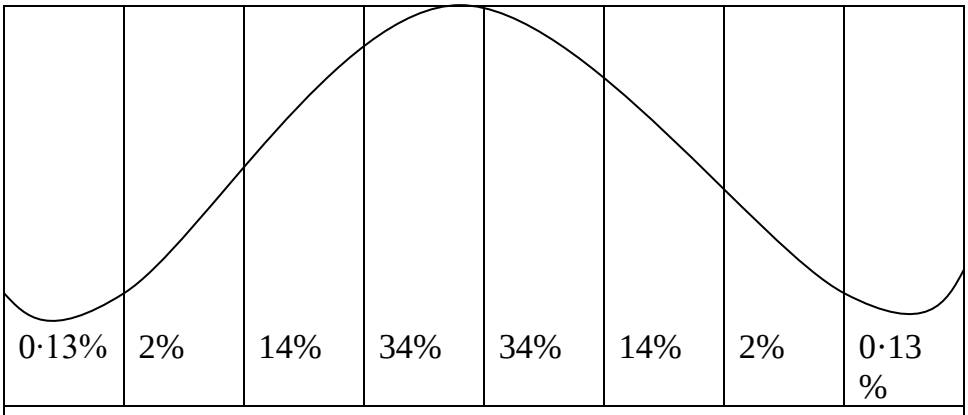
The number of examinees who should receive each stanine score is determined by multiplying the number of cases in the stanine level and rounding off the results. Usually, the distribution of test scores contains a number of examinees with the same raw score. Consequently, there are ties in rank that prevent obtaining of a perfect match with the theoretical distribution. Thus all examinees with the same raw score must be assigned the same stanine score. Hence, the actual grouping is approximated as closely as possible to the theoretical grouping. The relationship of stanine units to percentiles is illustrated below.

Stanine	Percentile Range
9	96 – 99
8	89 – 95
7	77 – 88
6	60 – 76
5	41 – 59
4	24 – 40

3	12 – 23
2	5 - 11
1	1 - 4

3.4 Comparison of the Score Systems

A normal distribution of scores makes it possible to convert back and forth between standard scores and percentiles thereby utilizing the special advantages of each. Standard scores can be used to draw on the benefits of equal units and it is also possible to convert to percentile equivalents when interpreting test performances to the testees, their parents and others who may need the information from the test results. Figure 2 illustrates the equivalence of scores in various standard score systems and their relation to percentiles and the normal curve.



0·13%	2%	14%	34%	34%	14%	2%	0·13 %	
-4	-3	-2	-1	0	+1	+2	+3	+4
Standard Deviation								
-4·0	-3·0	-2·0	-1·0	0	+1·0	+2·0	+3·0	+4· 0
Z-Scores								
0·1	2	16	50	84	98	99·9		
Percentiles								
20	30	40	50	60	70	80		
T-Scores								
1	2	3	4 Stan	5 ines	6	7	8	9
4% Percent	7% in	12 %	17% Stan	20% ines	17%	12 %	7%	4%

Figure 2: A normal distribution curve showing corresponding standard scores and percentiles.

SELF-ASSESSMENT EXERCISE

i. What is test interpretation and why is it necessary to interpret classroom test?

i. Highlight the major difference between criterion-reference interpretation and Norm-referenced interpretation.
iii. What are the main advantages of stanine scores over grade norms?

SELF-ASSESSMENT EXERCISE

A1: Test interpretation is a process of assigning meaning and usefulness to the scores obtained from classroom test. It is necessary to interpret classroom test because the raw score obtained from a test standing on itself rarely has meaning. The scores on their own lack a true zero point and equal units. Hence it is not possible to take reliable academic and psychological decisions on the basis of the raw scores.

A2: The major difference between criterion referenced interpretation and norm referenced interpretation is that Criterion-referenced interpretation is the interpretation given to raw score by converting it into a description of the specific tasks that the testee can perform; while Norm-referenced interpretation is the interpretation given to raw score by converting it into some type of derived score that indicates the testee relative position in a clearly defined referenced group.

A3: The main advantage of stanines scores over grade norms is that stanine scores as standard scores provides equal units that can be treated arithmetically and any given raw score in the reference group can be described by reference to the stanine within which it falls. The ease with which stanines are converted to other standard scores makes its interpretation easy and it is also possible to compare an individual's performance based on a common group. Whereas the grade norms merely describe test performance in terms of the particular grade or age group in which the testee's raw score is just average. Moreover depicting test performance in terms of grade equivalent can lead to unsound decisions because of the inequality of the units and the invalid assumptions on which they are based.

4.0 CONCLUSION

In this unit you learned how to interpret test scores. This includes criterion-referenced and norm-referenced interpretation such as the age and the grade norms. You also learn test interpretation based on standard scores norms utilizing the normal curve and the standard deviation units such as the z-scores, the T – scores and stanine norms. In addition you

learned about the percentile norms and the comparison between the score systems.

5.0 SUMMARY

- Test interpretation is a process of assigning meaning and usefulness to the scores obtained from classroom test. This is because the test scores on their own lack a true zero point and equal units.
- Criterion-referenced interpretation is the interpretation of test raw score based on the conversion of the raw score into a description of the specific tasks that the learner can perform.
- Norm-referenced interpretation is the interpretation of raw score based on the conversion of the raw score into some type of derived score that indicates the learner's relative position in a clearly defined reference group.
- Norms are referenced frames on which interpretation of test scores are based. They represent the typical performance of the testees in the reference frame on which the test raw scores were standardized.
- Grade norms are reference frame work for interpreting the academic achievement of learner's in the elementary schools. They represent the typical performance of specific groups such as a class.
- Age norms like the grade norms are based on the average scores earned by pupils of different ages and are interpreted in terms of age equivalents.
- Percentile norms are test norms that deal with percentile ranks or scores. They are use for comparison of percentage of group to which individual belongs.
- A percentile rank (or score) describes a learner's performance in terms of the percentage of learners in some clearly defined group that earn lower scores.
- Standard score is a method of indicating a testee's relative position in a group by showing how far the raw score is above or below average. Standard scores express test performance in terms of standard deviation units from the means.
- The normal curve is a symmetrical bell-shaped curve that has many useful mathematical properties utilized in test interpretation using standard scores.
- The Z-score is a simple standard score which expresses test performance simply and directly as the number of standard deviation units a raw score is above or below the mean.

- The T – score refers to any set of normally distributed standard scores that has a means score of 50 and a standard deviation of 10.
- The stanine is a kind of standard score which divides a population according to some proportions into nine parts numbered 1 to 9. It has a mean score of 5 and a standard deviation of 2 . Each stanine corresponds to a score or a range of scores. Each individual's score falls within a stanine and such score can be described by reference to the stanine within which it falls. Stanines are widely used for local norms because of the ease with which they can be computed and interpreted.

6.0 TUTOR-MARKED ASSIGNMENT

1. With the help of a table present the most common types of test norms for education and psychological tests. For each type of test name the derived score, type of comparison undertaken, type of group used and the meaning in terms of test performance.
2. What is T-score?
3. A test has a mean score of 40 and a standard deviation of 4. What are the T – scores of two testees who obtained raw scores of 30 and 45 respectively in the test?
4. What are the limitations of the Percentile rank? How were these limitations overcome in Stanine scores?

7.0 REFERENCES/FURTHER READINGS

Gronlund, N.E (1985), Measurement and Evaluation in Teaching. New York:Macmillan Publishing Company.

Kerlinger, F.N (1973), Foundations of Behavioural Research. Holt, Rinehart and Winston Inc. New York:

Thorndike, R.L and Hagen, E.P (1977). Measurement and Evaluation in Psychology and Education. New York: John Wiley and Sons; New York.

UNIT 3 RELIABILITY OF A TEST

CONTENTS

1.0 INTRODUCTION

In this unit you will learn about test reliability and the methods of estimating reliability. Specifically, you will learn about test retest method, equivalent form method, split half method and Kuder Richardson Method.

Furthermore, you will learn about the factors influencing reliability measures such as length of test, spread of scores, difficulty of test and objectivity. Finally you will learn the methods of estimating reliability.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- Define reliability of a test;
- State the various forms of reliability;
- Identify and explain the factors that influence reliability measures;
- Compare and contrast the different forms of estimating reliability.

3.0 MAIN CONTENT

3.1 Test Reliability

Reliability of a test may be defined as the degree to which a test is consistent, stable, dependable or trustworthy in measuring what it is measuring. This definition implies that the reliability of a test tries to answer questions like: How can we rely on the results from a test? How dependable are scores from the test? How well are the items in the test consistent in measuring whatever it is measuring? In general, reliability of a test seeks to find if the ability of a set of testees are determined based on testing them two different times using the same test, or using two parallel forms of the same test, or using scores on the same test marked by two different examiners, will the relative standing of the testees on each of the pair of scores remain the same? Or, to what extent will their relative standing remain the same for each of the pair of scores? Therefore, reliability refers to the accuracy, consistency and stability with which a test measures whatever it is measuring. The more the pair of scores observed for the same testee varies from each other, the less reliable the measure is. The variation between this pair of scores is caused by numerous factors other than influence test scores.

Such extraneous factors introduce a certain amount of error into all test scores. Thus, methods of determine reliability are essential means of determining how much error is present under different conditions. Hence, the more consistent our test results are from one measurement to another, the less error there will be and consequently, the greater the reliability.

3.1 Types of Reliability Measures

There are different types of reliability measures. These measures are estimated by different methods. The chief methods of estimating reliability measures are illustrated in table 1 below.

Table 1: Methods of Estimating Reliability

	Types of Reliability Measure	Procedure
Test-retest method	Measure of stability	Give the same test twice to the same group with any time interval between tests
Equivalent-forms methods	Measure of equivalence	Give two forms of the test to the same group in close succession
Split – half method	Measure of internal consistency	Give test once. Score two equivalent halves say odd and even number items, correct reliability coefficient to fit whole test by Spearman-Brown formula
Kuder-Richardson methods	Measure of internal consistency	Give test once. Score total test and apply kuder-Richardson formula

3.2.1 Test Retest Method – Measure of Stability

Estimating reliability by means of test-retest method requires the same test to be administered twice to the same group of learners with a given time interval between the two administrations. The resulting test scores are correlated and the correlation coefficient provides a measure of stability. How long the time interval should be between tests is determined largely by the use to be made of the results. If the results of both administrations of the test are highly stable, the testees whose scores are high on one administration of the test will tend to score high on other administration of the test while the other testees will tend to stay in the same relative positions on both administration of the test. Such stability would be indicated by a large correlation coefficient.

An important factor in interpreting measures of stability is the time interval between tests. A short time interval such as a day or two inflates the consistency of the result since the testees will remember some of their answers from the first test to the second. On the other hand, if the time interval between tests is long about a year, the results will be influenced by the instability of the testing procedure and by the actual changes in the learners over a period of time. Generally, the longer the time interval between test and retest, the more the results will

be influenced by changes in the learners' characteristics being measured and the smaller the reliability coefficient will be.

3.2.2 Equivalent Forms Method - Measure of Equivalence

To estimate reliability by means of equivalent or parallel form method involves the use of two different but equivalent forms of the test. The two forms of the test are administered to the same group of learners in close succession and the resulting test scores are correlated. The resulted correlation coefficient provides a measure of equivalence. That is, correlation coefficient indicates the degree to which both forms of the test are measuring the same aspects of behaviour. This method reflects the extent to which the test represents an adequate sample of the characteristics being measured rather than the stability of the testee. It eliminates the problem of selecting a proper time interval between tests as in test retest method but has the need for two equivalent forms of the test. The need for equivalent forms of the test restricts its use almost entirely to standardized testing where it is widely used.

3.2.3 Split-Half Method – Measure of Internal Consistency.

This is a method of estimating the reliability of test scores by the means of single administration of a single form of a test. The test is administered to a group of testees and then is divided into two halves that are equivalent usually odd and even number items for scoring purposes. The two split half produces two scores for each testee which when correlated, provides a measure of internal consistency. The coefficient indicates the degree to which equivalent results are obtained from the two halves of the test. The reliability of the full test is usually obtained by applying the Spearman-Brown formula.

That is, Reliability on full test = $\frac{2 \times \text{Reliability on } ½ \text{ test}}{1 + \text{Reliability on } ½ \text{ test}}$

The split-half method, like the equivalent forms method indicates the extent to which the sample of test items is a dependable sample of the content being measured. In this case, a high correlation between scores on the two halves of a test denotes the equivalence of the two halves and consequently the adequacy of the sampling. Also like the equivalent-forms method, it tells nothing about changes in the individual from one time to another.

3.2.4 Kuder-Richardson Method – Measure of Internal Consistency.

This is a method of estimating the reliability of test scores from a single administration of a single form of a test by means of formulas such as those developed by Kuder and Richardson. Like the spilt-half method, these formulas provide a measure of internal consistency. However, it does not require splitting the test in half for scoring purposes. Kuder-Richardson formula 20 which is one of the formulas for estimating internal consistency is based on the proportion of persons passing each item and the standard deviation of the total scores. The result of the analysis using this formula is equal to all possible split-half coefficients for the group tested. However, it is rarely used because the computation is rather cumbersome unless information is already available concerning the proportion of each item. Kuder-Richardson formula 21, a less accurate but simpler formula to compute can be applied to the results of any test that has been scored on the basis of the number of correct answers. A modified version of the formula is:

$$\text{Reliability estimate (kR21)} = \frac{k}{K-1}\left(1 - \frac{m(k-m)}{ks^2}\right)$$

Where

K	=	the number of items in the test
M	=	the mean of the test scores
S	=	standard deviation of the test scores.

The result of this formula approximates that of Kuder-Richardson formula 20. It has a smaller reliability estimate in most cases.

This method of reliability estimate test whether the items in the test are homogenous. In other words, it seeks to know whether each test item measures the same quality or characteristics as every other. If this is established, then the reliability estimate will be similar to that provided by the split-half method. On the other hand, if the test lacks homogeneity an estimate smaller than split-half reliability will result. The Kuder-Richardson method and the Split-half method are widely used in determining reliability because they are simple to apply. Nevertheless, the following limitations restrict their value. The limitations are:

- They are not appropriate for speed test in which test retest or equivalent form methods are better estimates.
- They, like the equivalent form method, do not indicate the constancy of a testee response from day to day. It is only the test-retest procedures that indicate the extent to which test results are generalizable over different periods of time.

They are adequate for teacher-made tests because these are usually power tests.

3.3 Factors Influencing Reliability Measures

The reliability of classroom tests is affected by some factors. These factors can be controlled through adequate care during test construction. Therefore the knowledge of the factors are necessary to classroom teachers to enable them control them through adequate care during test construction in order to build in more reliability in norm referenced classroom tests.

3.3.1 Length of Test

The reliability of a test is affected by the length. The longer the length of a test is, the higher its reliability will be. This is because longer test will provide a more adequate sample of the behaviour being measured, and the scores are apt to be less distorted by chance factors such as guessing. If the quality of the test items and the nature of the testees can be assumed to remain the same, then the relationship of reliability to length can be expressed by the simple formula

$$r_{nn} = \frac{nr_{11}}{1 + (n - 1)\, r_{11}}$$

Where

r_{nn} is the reliability of a test n times as long as the original test;
r_{ii} is the reliability of the original test; and
n is as indicated, the factors by which the length of the test is increased.

Increase in length of a test brings test scores to depend closer upon the characteristics of the person being measured and more accurate appraisal of the person is obtained. However, we all know that lengthen a test is limited by a number of practical considerations. The considerations are the amount of time available for testing, factors of fatigue and boredom on part of the testees, inability of classroom teachers to constructs more equally good test items. Nevertheless, reliability can be increased as needed by lengthening the test within these limits.

3.3.2 Spread of Scores

The reliability coefficients of a test are directly influenced by the spread of scores in the group tested. The larger the spread of scores is, the higher the estimate of reliability will be if all other factors are kept constant. Larger reliability coefficients result when individuals tend to

stay in same relative position in a group from one testing to another. It therefore follows that anything that reduces the possibility of shifting positions in the group also contributes to larger reliability coefficient. This means that greater differences between the scores of individuals reduce the possibility of shifting positions. Hence, errors of measurement have less influence on the relative position of individuals when the differences among group members are large when there is a wide spread of scores.

3.3.3 Difficulty of Test

When norm-referenced test are too easy or too difficult for the group members taking it, it tends to produce scores of low reliability. This is so since both easy and difficult tests will result in a restricted spread of scores. In the case of easy test, the scores are closed together at the top of the scale while for the difficult test; the scores are grouped together at the bottom end of the scale. Thus for both easy and difficult tests, the differences among individuals are small and tend to be unreliable. Therefore, a norm-referenced test of ideal difficulty is desired to enable the scores to spread out over the full range of the scale. This implies that classroom achievement tests are to be designed to measure differences among testees. This can be achieved by constructing test items with at least average scores of 50 percent and with the scores ranging from zero to near perfect scores. Constructing tests that match this level of difficulty permits the full range of possible scores to be used in measuring differences among individuals. This is because the bigger the spread of scores, the greater the likelihood of its measured differences to be reliable.

3.3.4 Objectivity

This refers to the degree to which equally competent scorers obtain the same results in scoring a test. Objective tests easily lend themselves to objectivity because they are usually constructed so that they can be accurately scored by trained individuals and by the use of machines. For such test constructed using highly objective procedures, the reliability of the test results is not affected by the scoring procedures. Therefore the teacher made classroom test calls for objectivity. This is necessary in obtaining reliable measure of achievement. This is more obvious in essay testing and various observational procedures where the results of testing depend to a large extent on the person doing the scoring. Sometimes even the same scorer may get different results at different times. This inconsistency in scoring has an adverse effect on the reliability of the measures obtained. The resulting test scores reflect the opinions and biases of the scorer and the differences among testees in the characteristics being measured.

Objectivity can be controlled by ensuring that evaluation procedures selected for the evaluation of behaviour required in a test is both appropriate and as objective as possible. In the case of essay test, objectivity can be increased by careful framing of the questions and by establishing a standard set of rules for scoring. Objectivity increased in this manner will increase reliability without undermining validity.

3.4 Methods of Estimating Reliability

Different methods of estimating reliability of a test yield different values of reliability estimates even for the same test. This is because of the differences in the way each of the procedures defines measurement error. In other words, the size of the reliability coefficient is related to the method of estimating reliability. This is clarified below:

- Test-retest method: Reliability Coefficient based on test retest method is always lower than reliability obtained from split-half method but higher than those obtained through equivalent forms method. This is because test-retest method is affected by time to time fluctuation.
- Equivalent forms method: This method has the least value of reliability which real reliability could possibly take. Equivalent forms method is affected by both time to time and form to form fluctuations.
- Spilt-half method: Estimation of reliability from the spilt-half method gives the largest value which the real reliability can possibly take. This is because any reliability estimate that is based on the result of a single testing will result in an overestimation of the reliability index.
- Kuder-Richardson methods: This method like the spilt-half method is based on single testing and so the reliability index is over estimation. Its value is however lower than the value obtained for the spilt-half method.

It is clear from the above illustration that the size of the reliability coefficient resulting from the method of estimating reliability is directly attributable to the type of consistency included in each method. Thus, the more rigorous methods of estimating reliability yield smaller reliability coefficient than the less rigorous methods. It is therefore essential that when estimating the reliability of a measurement instrument, the method used, the time lapse between repeated administration and the intervening experience must be noted as well as the assumptions and limitations of the method used for a clearly understanding of the resulting reliability estimate.

SELF-ASSESSMENT EXERCISE

i. Define the reliability of a test.
ii. Mention the four chief methods of estimating reliability and the type of reliability measure associated with each of them.
iii. What are the factors that influence reliability measures?

SELF ASSESSMENT EXERCISE

A1: The reliability of a test may be defined as the degree to which a test is consistent, stable, dependable or trustworthy in measuring what it is measuring.
A2: The four chief methods of estimating reliability and the type of reliability measures associated with each of them are:

Method of estimating reliability		Type of reliability measure associated
i.	Test retest method	Measure of stability
ii.	Equivalent form method	Measure of equivalence
iii.	Split-half method	Measure of internal consistency
iv.	Kuder-Richardson method	Measure of internal consistency

A3: The factors that influence reliability measures are length of tests, spread of scores, difficulty of test and objectivity

4.0 CONCLUSION

In this unit you learned about test reliability. Furthermore you learned types of reliability measures and methods of estimating them. Specifically you learned about measure of equivalence, measure of stability and measure of internal consistency. In addition you learned about the test retest method of estimating reliability, the equivalent forms methods and the Kuder-Richardson method. Finally, you learned the factors such as length of test, spread of scores, and difficulty of test and objectivity that influencesreliability measures.

5.0 SUMMARY

- Reliability of a test is defined as the degree to which a test is consistent, stable, dependable or trustworthy in measuring what it is measuring.
- Measure of stability is estimated by test-retest method.
- Measure of equivalence is estimated by equivalent form method.

- Measure of internal consistency is measured by split-half method and Kuder-Richardson method.
- Factors affecting reliability measures are length of test, spread of scores, difficulty of test and objectivity.

6.0 TUTOR-MARKED ASSIGNMENTS

1. Define reliability of a test.
2. State the four common methods of estimating reliability, the type of reliability measure associated with each and the procedure for estimating each reliability measure.
3. Identify and explain the factors that influence reliability measures.

7.0 REFERENCE/FURTHER READING

Gronlund, N.E (1985), Measurement and Evaluation in Teaching. New York: Macmillan Publishing Company.

Nenty, H.J (1985), Fundamentals of Measurement and Evaluation in Education. Calabar: University of Calabar.

Thorndike, R.L and Hagen, E.P (1977). Measurement & Evaluation in Psychology and Education. 4e New York: John Wiley and Sans.

UNIT 4 VALIDITY OF CLASSROOM TEST

CONTENTS

1.0 INTRODUCTION

In this unit, you will learn about validity, which is the single most important criteria for judging the adequacy of a measurement instrument. You will learn types of validity namely content, criterion and construct validity. Furthermore, you will learn about content validation, criterion validation and construct validation. Finally you will learn validity of criterion-referenced mastery tests and factors influencing validity.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- define validity as well as content, criterion and construct validity
- differentiate between content, criterion and construct validity
- describe how each of the three types of validity are determined
- interpret different validity estimates
- identify the different factors that affect validity
- assess any test based on its validity.

3.0 MAIN CONTENT

3.1 Validity

Validity is the most important quality you have to consider when constructing or selecting a test. It refers to the meaningfulness or appropriateness of the interpretations to be made from test scores and other evaluation results. Validity is therefore, a measure or the degree to which a test measures what it is intended to measure. For example, a classroom essay test designed to measure students' understanding of the role of FIFA in World Cup, besides measuring the ability of the testees understanding of the role of FIFA also calls for ability to read and use language of expression as well as the technical terms appropriate in expressing oneself including grammar and spelling abilities.

Although these aspects were not originally intended, they are consciously or unconsciously considered when scoring the responses to the question. But the original design of the question was not to measure these skills. If students' skill in these other areas were of interest, appropriate tests should be designed to measure each of them.

With the present test, their influence impairs the validity of the resulting scores as accurate indicators of the students' understanding of the role of FIFA in World Cup. In general validity refers to the appropriateness of the interpretations made from test scores and other evaluation results, with regard to a particular use. Thus validity is always concerned with the specific use of the results and the soundness of our proposed interpretations. Hence, to the extent that a test score is decided by factors or abilities other than that which the test was designed or used to measure, its validity is impaired.

3.2 Types of Validity

The concern of validity is basically three, which are:

- Determining the extent to which performance on a test represents level of knowledge of the subject matter content which the test was designed to measure – content validity of the test;
- Determining the extent to which performance on a test represents the amount of what was being measured possessed by the examinee – construct validity of the test;
- Determining the extent to which performance on a test represents an examinee's probable task – criterion (concurrent and prediction) validity of a test.

These concerns of validity are related and in each case the determination is based on a knowledge of the interrelationship between scores on the test and the performance on other task or test that accurately represents the actual behaviour. Validity may therefore be viewed as a unitary concept based on various kinds of evidence. The three approaches to test validation are briefly discussed in figure 1 below.

	Procedure	Meaning
Content – Related Evidence	Compare the test tasks to the test specifications describing the task domain under consideration.	How well the sample of test tasks represents the domain of tasks to be measured
Criterion – Related Evidence	Compare test scores with another measure of performance obtained at a later date (for prediction) or with another measure of performance obtained concurrently (for estimating present status)	How well test performance predicts future performance or estimates current performance on some valued measures other than the test itself (a criterion)
Construct – Related Evidence	Establish the meaning of the scores on the test by controlling (or examining) the development of the test and experimentally determining what factors influence test performance	How well test performance can be interpreted as a meaning of some characteristics or quality

Figure 1: Approaches to Test Validation

3.3 Content Validation

This is the process of determining the extent to which a set of test tasks provides a relevant and representative sample of the domain of tasks under consideration. Content validation procedure is especially important in achievement testing and is of primary concern in

constructing classroom tests. The procedure involves writing test specifications that define a domain of instructionally relevant tasks and then constructing or selecting test items that provide a representative and relevant measure of the tasks. In classroom testing, the domains of achievement tasks are determined by the instruction, and test development involves:

- Clearly specifying the domain of instructionally relevant tasks to be measured; and
- Constructing or selecting a representative set of test tasks.

Therefore, to obtain a valid measure of learning outcomes, we start from what has been taught to what is to be measured and finally to a representative sample of the relevant tasks. That is, for a validity measure, you have to consider the instruction, the achievement domain before the test itself.

3.3.1 Content Validation and Test Development

You know that the most relevant type of validity associated with teacher-made classroom or achievement test is content validity. You also know that content validation typically takes place during test development. Test development is primarily a matter of preparing detailed test specifications and then constructing a test that meets these specifications. In the table of specification the content of a course may be defined to include both subject matter content and instructional objectives. While the subject matter topics is concerned with the topics to be learned, the instructional objectives is concerned with the type of performance the learners are expected to demonstrate. This is done so that the test we construct using this specification will aid in constructing test that will produce results that represents both the content areas and the objectives we wish to measure.

You will recall that in Module 3 Unit 4 you learned about the use of table of specification (table 3.1) in test development. The primary function of that table is for content validation. The percentages in the table indicate the relative degree of emphasis that each content area and each instructional objective is to be given in the test. Thus, to ensure content validity using the table, the specifications describing the achievement domain to be measured should reflect what was taught and the test items must function as intended if valid results are to be obtained.

3.2 Content Validation and Face Validity

Face Validation is a quick method of establishing the content validity of a test after its preparation. This is done by presenting the test to subject experts in the field for their experts' opinion as to how well the test "looks like" it measures what it was supposed to measure. This process is refereed to as face validity. It is a subjective evaluation based on a superficial examination of the items, of the extent to which a test measures what it was intended to measure.

Face Validity also applies to some extent to construct validity. It goes beyond the superficial examination of the items to determine how the test looked to the extent to which appropriate phrasing is used in the construction of an item to appear more relevant to the test taker. The validity of the test would be determined by how well it sampled the domain of task using appropriate phrases relevant to the test taker's level and /or field of study by using the appropriate subject language.

3.3 Criterion Validation

Criterion Validation is the process of determining the extent to which test performance is related to some other valued measure of performance. It tries to determine the level to which one can confidently tell, based on an individual's score on a given test, his performance on a current or future related task. The current or future related task is called the criterion. This may involve studies of how well test scores predict future performance (predictive validation study) or estimate some current performance (concurrent validation study). It can be a test on the job which requires skills and knowledge similar to those called for by the test whose criterion validity is being estimated. These validity studies are typically reported by means of correlation coefficient called Validity Coefficient.

Predictive Validity tells us the relationship between measures over an extended period of time. While concurrent validity is interested in estimating present status and this tells us the relationship between the measures obtained concurrently. In other words, while concurrent validity indicates the extent to which the test can be used to estimate an examinee's present standing on a criterion, predictive validity involves the extent to which scores from the test can be used to estimate or predict level of future performance on a criterion. That is, in both cases the intention is to predict present or future performance on a criterion based on observed performance on a given test. A typical example is how well trainees' performance on a written test relates to their performance on-the-job practical test which calls on the same skills and knowledge involved in the written test is an indication of concurrent

validity. But how well students' performance on Mock SSCE relates to their performance in the actual SSCE is an indication of the predictive validity of the Mock SSCE.

The major difference between Concurrent Validity and Predictive Validity is that while for a Concurrent Validity both the test information and the criterion data are gathered at the same time, in the Predictive Validity test information is to be obtained first and then at some later time, performance on the criterion would be obtained. Then the correlation between the two sets of scores would be determined. The usual procedure is to correlate statistically the two sets of scores and to report the degree of relationship between them by means of a correlation coefficient which enables validity to be presented in precise and universally understood terms.

3.4.1 Methods of Computing Correlation Coefficient

A correlation coefficient expresses the degree of relationship between two sets of scores by numbers ranging from +1·00 from – 1.00 to + 1.00 to -1·00. A perfect positive correlation is indicated by a coefficient of +1·00 and a perfect negative correlation by a coefficient of -1·00. A correlation coefficient of ·00 lies midway between these extremes and indicates no relationship between the two sets of scores. The larger the coefficient (positive or negative), the higher the degree of relationship expressed. There are two common methods of computing correlation coefficient. These are:

- Spearman Rank-Difference Correlation
- Pearson Product-Moment Correlation.

You will note that correlation indicates the degree of relationship between two scores but not the causation of the relationship. Usually, further study is needed to determine the cause of any particular relationship. The following are illustrations of the computation of correlation coefficients using both methods.

Spearman Rank-Difference Correlation: This method is satisfactory when the number of scores to be correlated is small (less than 30). It is easier to compute with a small number of cases than the Pearson Product-Moment Correlation. It is a simple practical technique for most classroom purposes. To use the Spearman Rank-Difference Method, the following steps listed in table 3.1 should be taken.

Table 3.1: Computing Guide for the Spearman Rank-Difference Correlation

Step	Procedure	Result in Table 3.2
1	2	3	4	5	6
1	Arrange pairs of scores for each examinee in columns	Columns 1 and 2
2	Rank examinees from 1 to N (number in group) for each set of scores	Columns 3 and 4
3	Rank the difference (D) in ranks by subtracting the rank in the right hand column (4) from the rank in the left-hand column (column 3)	Column 5
4	Square each difference in rank (column 5) to obtain difference squared (D^2)	Column 6
5	Sum the squared differences in column 6 to obtain ΣD^2	Bottom of Column 6
6	Apply the following formula ρ (rho) = 1 - $\frac{6 \times \Sigma D^2}{N(N^2 - 1)}$ Where Σ = Sum of D = Difference in rank N = Number in group	$\rho = 1 - \frac{6 \times 514}{20(20^2 - 1)}$ $= 1 - \frac{3084}{7980}$ $= 1 - \cdot 39$ $= \cdot 61$

Table 3.2:

Student Number	**Mathematics Score**	**Physics Score**	**Mathematics Rank**	**Physics Rank**	**D**	**D^2**
1	98	76	1	2	-1	1
2	97	75	2	3	-1	1
3	95	72	3	4	-1	1
4	94	70	4	5	-1	1
5	93	68	5	6	-1	1
6	91	66	6	7	-1	1
7	90	64	7	8	-1	1
8	89	60	8	10	-2	4
9	88	58	9	11	-2	4
10	87	57	10	12	-2	4
11	86	56	11	13	-2	4
12	84	54	12	14	-2	4
13	83	52	13	15	-2	4
14	81	50	14	16	-2	4
15	80	48	15	17	-2	4
16	79	46	16	18	-2	4
17	77	44	17	20	-3	9
18	76	45	18	19	-1	1
19	75	62	19	9	10	100
20	74	77	20	1	19	361
Rank Difference						$\Sigma D^2 = 514$

Correlation for a Pair of Hypothetical Data

Pearson Product-Moment Correlation: This is the most widely used method and thecoefficient is denoted by the symbol r. This method is favoured when the number of scores is large and it's also easier to apply to large group. The computation is easier with ungrouped test scores and would be illustrated here. The computation with grouped data appears more complicated and can be obtained from standard statistics test book. The following steps listed in table 3.3 will serve as guide for computing a product-moment correlation coefficient (r) from ungrouped data.

Table 3.3: Computing Guide for the Pearson Product-Moment Correlation Coefficient (r) from Ungrouped Data

Step	Procedure	Results in Table 3.4
1	Begin by writing the pairs of score to be studied in two columns. Make certain that the pair of scores for each examinee is in the same row. Call one Column X and the other Y	Columns 1 and 2
2	Square each of the entries in the X column and enter the result in the X^2 column	Column 3
3	Square each of the entries in the Y column and enter the result in the Y^2 column	Column 4
4	In each row, multiply the entry in the X column by the entry in the Y column, and enter the result in the XY column	Column 5
5	Add the entries in each column to find the sum of (Σ) each column. Note the number (N) of pairs of scores From Table 3.4, then $\Sigma X = 1717 \quad \Sigma X^2 = 148487$ $\Sigma Y = 1200 \quad \Sigma Y^2 = 74184$ $\Sigma XY = 103984$	
6	Substitute the obtained values in the formula: $r = \frac{\Sigma XY - \left(\frac{\Sigma X}{N}\right)\left(\frac{\Sigma Y}{N}\right)N}{N\sqrt{\left(\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2\right)\left(\frac{\Sigma Y^2}{N} - \left(\frac{\Sigma Y}{N}\right)^2\right)}}$	

OR

7 Divide the sum of each column by N before putting the data into the formula.

Thus for data from Table 3.4,

$\frac{\Sigma X}{N} = \frac{1717}{20} = 85{\cdot}85$ $\frac{\Sigma Y^2}{N} = \frac{74184}{20} = 3709{\cdot}20$

$\frac{\Sigma Y}{N} = \frac{1200}{20} = 60{\cdot}00$ $\frac{\Sigma XY}{N} = \frac{103984}{20} = 5199{\cdot}20$

$\frac{\Sigma X^2}{N} = \frac{148487}{20} = 7424{\cdot}35$

Then, substituting in the formula (no.6)

$$r = \frac{5199{\cdot}20 - (85{\cdot}85)(60{\cdot}00)}{\sqrt{7424{\cdot}35 - (85{\cdot}85)^2}\sqrt{3709{\cdot}20 - (60{\cdot}00)^2}}$$

$$r = \frac{5199{\cdot}20 - 5151{\cdot}00}{\sqrt{7424{\cdot}35 - 7370{\cdot}22}\sqrt{3709{\cdot}20 - 3600{\cdot}00}}$$

$$r = \frac{48{\cdot}20}{\sqrt{54.13}\sqrt{109{\cdot}20}} = \frac{48{\cdot}20}{7{\cdot}36 \times 10{\cdot}45} = \frac{48{\cdot}20}{76{\cdot}91}$$

$r = {\cdot}63$

9 Note that the computations involve finding the mean and standard deviation of each set of scores (X and Y). Hence the formula can also be written thus:

$$r = \frac{\frac{\Sigma XY}{N} - (Mx)(My)}{(SDx)(SDy)}$$

where

Mx = mean of scores in X column

My = mean of scores in Y column

SDx = standard deviation of scores in X column

SDy = standard deviation of scores in Y column

Thus for the same data

$$r = \frac{5199{\cdot}20 - (85{\cdot}85)(60{\cdot}00)}{7{\cdot}36 \times 10{\cdot}45}$$

$= {\cdot}63$

Table 3.4: Computing Pearson Product-Moment Correlation for a pair of the Hypothetical Ungrouped Data

	1	2	3	4	5
Student Number	Mathematics Score (X)	Physics Score (Y)	X^2	Y^2	XY
1	98	76	9604	5776	7448
2	97	75	9409	5625	7275
3	95	72	9025	5184	6840
4	94	70	8836	4900	6580
5	93	68	8649	4624	6324
6	91	66	8281	4356	6006
7	90	64	8100	4096	5760
8	89	60	7921	3600	5340
9	88	58	7744	3364	5104
10	87	57	7569	3249	4959
11	86	56	7396	3136	4816
12	84	54	7056	2916	4536
13	83	52	6889	2704	4316
14	81	50	6561	2500	4050
15	80	48	6400	2304	3840
16	79	46	6241	2116	3634
17	77	44	5929	1936	3388
18	76	45	5776	2025	3420
19	75	62	5625	3844	4650
20	74	77	5476	5929	5698
N=20	$(\Sigma X)=1717$	$(\Sigma Y)=1200$	$(\Sigma X^2)=148487$	$(\Sigma Y^2)=74184$	$(\Sigma XY)=103984$

4.2 Interpretation of Correlation Coefficient (r) Values

You know that correlation coefficient indicates the degree of relationship between two sets of scores by numbers ranging from +1·00 to -1·00. You also know that a perfect correlation is indicated by a coefficient of +1·00 and a perfect negative correlation by a coefficient of -1·00. Thus, a correlation coefficient of ·00 lies midway between these extremes and indicates no relationship between the two sets of scores. In addition to the direction of relationship which is indicated by a positive sign (+) for a positive (direct relation) or a negative sign (-) for a negative (inverse relation), correlation also has a size, a number which indicates the level or degree of relationship. The larger this number the more closely or highly two set of scores relate. We have said that correlation coefficient or index take on values between +1·00 and -1·00. That is, the size of relationship between two set of scores is never more

than +1·00 and never less than -1·00. The two values have the same degree or level of relationship but while the first indicates a direct relation the second indicates an inverse relation. A guide for interpreting correlation coefficient (r) values obtained by correlating any two set of test scores is presented in figure 2 below:

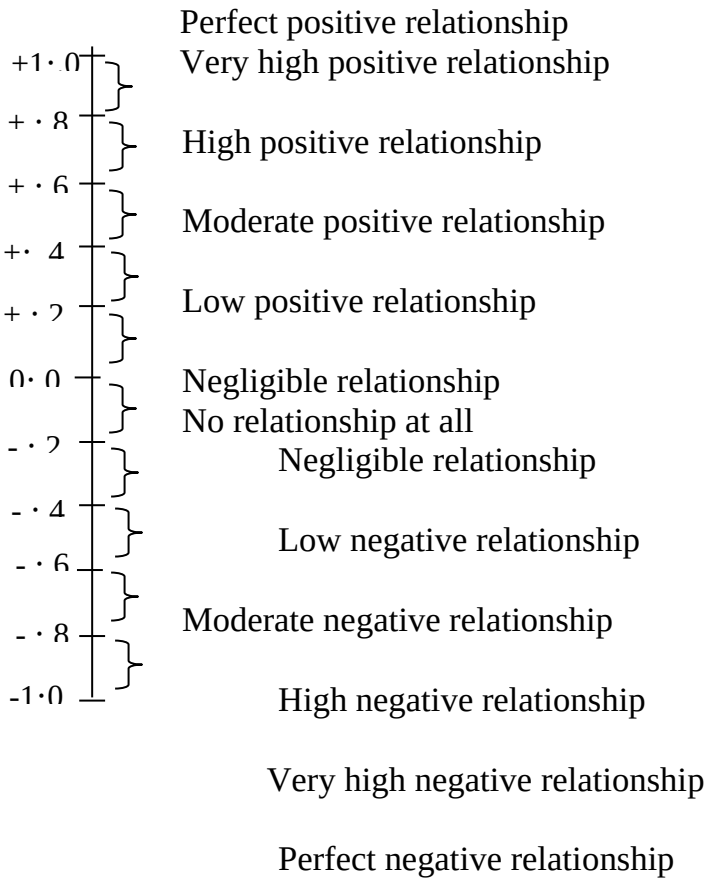


Figure 2: Interpretation of r-values

The value of criterion validity index is affected by:

- the validity of the criterion variable itself (validity of the instrument used in obtaining criterion scores).
- the time interval between the administration of the two tests (predictor and criterion measures; and
- the selection of criterion group based on performance on the predictor test scores which results in a homogeneous group that is no more a good representative of the original population to which inference based on the test validity is to be made.

3.5 Construct Validation

A Construct is a psychological quality that we assume exists in order to explain some aspect of behaviour. Examples of construct include

reasoning ability, intelligence, creativity, honesty and anxiety. They are called construct because they are theoretical constructions that are used to explain behaviour. When we interpret test scores as a measure of a particular construct, we are implying that there is such a construct, that it differs from other constructs, and that the test scores provide a measure of the construct that is little influenced by extraneous factors. Instruments through which attempts are made to define or give meaning to the construct are developed and construct validity is the degree to which such an instrument accurately defines or gives meaning to the construct.

The advantage of being able to interpret test performance in terms of the psychological construct is that each construct has an underlying theory that can be brought to bear in describing and predicting a person's behaviour. For instance, when we say that a person is highly intelligent, has good reading comprehension or is sociable, it gives idea to what type of behaviour might be expected in various specific situations. Hence construct validation is defined as the process of determining the extent to which test performance can be interpreted in terms of one or more psychological constructs. Construct Validation has implications for the practical use of test results. Usually, when a test is to be interpreted as measure of a particular construct, the various types of evidence useful for construct validation should be considered during its development and selection. This includes evidence from content and criterion referenced validation. The process of Construct Validation usually involves:

- identifying and describing by means of a theoretical framework, the meaning of the construct to be measured;
- deriving hypotheses, regarding test performance from the theory
- underlying the construct; and
- verifying the hypotheses by logical and empirical means.

Primarily, Construct Validation takes place during the development and try-out of a test and is based on an accumulation of evidence from many different sources. Thus, the development and selection of test items on the basis of construct validation focuses on the types of evidence that it seems reasonable to obtain. Moreover, special attention is given to interpretation to be made.

3.5.1 Methods Used in Construct Validation

You have seen that construct is a complex variable and so, sometimes simpler variables or factors are often postulated to explain them. This involves both logical analysis and various comparative and correlation studies accumulation of which is endless. Nevertheless, in practical

situations we focus on the types of evidence that it seems reasonable to obtain and that are most relevant to the types of interpretations to be made.

The following methods are commonly used.

i. Determining the domain of tasks to be measured: State the test blue print clearly to bring out the meaning of the construct and the extent to which the test provides a relevant and representative measure of the task domain.
ii. Analyzing the mental process required by the test items: Determine the mental process called forth by the test by examining the test items themselves and by administering the test to individual learners and having them "think aloud" as they answer.
iii. Comparing the Scores of known groups: This entails checking a prediction of differences for a particular test between groups that are known to differ and the results are to be used as partial support for construct validation.
iv. Comparing results before and after some particular treatment: This involves verification of test results for the theory underlying the trait being measured. This is done by administering the test before and after treatment to determine the effect of the treatment.
v. Correlating the scores with other tests: The scores of any particular test is expected to correlate substantially with the scores of other tests that presumably measure the same ability or treat and vice versa. Therefore, the construct validity estimate of a new test could be got by correlating scores from the test with scores from each of two or more existing valid tests that measure the same construct. If the resulting correlation coefficients are high then the new test would be said to measure the same construct with the older test. This is based on the fact that a test that is a valid measure of one construct should correlate highly with other construct that are known to relate to the construct it is measuring. The construct validity of a given test can also be estimated by correlating scores from it with scores from two or more valid tests that measure each of the related construct. In this case, the resulting coefficients might not be as high as those in the first case but moderately high. The resulting correlation coefficients are indications of the construct validity of the new test. This correlation coefficient must always be qualified by indicating the criterion variable, the population used and information regarding the type of inference or decision that could be made based on the test scores.

The construct validity of a test based on correlation method could be determined with the use of multi trait – multi method and the factor analysis.

3.6 Validity of Criterion-Referenced Mastery Tests

You know that Criterion-Referenced Mastery tests are not designed to discriminate among individuals; hence, statistical validation procedures play a less prominent role here. However, all three types of validity evidence are still important in constructing and selecting this type of test. Content-related evidence is of primary concern when criterion-referenced tests are used for instructional purposes. This process also involves specifying the performance domain to be measured and constructing or selecting a set of tasks that is both relevant and representative. The procedure depends on logical analysis and judgment; hence score variability is not crucial. For other uses of this test both criterion – related and construct related evidence are likely to receive increased emphasis.

3.7 Factors Influencing Validity

Many factors tend to influence the validity of test interpretation. These factors include factors in the test itself. The following are the list of factors in the test itself that can prevent the test items from functioning as intended and thereby lower the validity of the interpretations from the test scores.

They are:

- Unclear directions on how examinees should respond to test items;
- Too difficult reading vocabulary and sentence structure;
- Inappropriate level of difficulty of the test items;
- Poorly structured test items;
- Ambiguity leading to misinterpretation of test items;
- Inappropriate test items for the outcomes being measured;
- Test too short to provide adequate sample;
- Improper arrangement of items in the test; and
- Identifiable pattern of answer that leads to guessing.

Other factors that influence validity include:

- Functioning content and teaching procedures;
- Factors in the test administration and scoring;
- Factors in examinees' response; and
- Nature of the group and the criterion.

Thus, in order to ensure validity conscious effort should be made during construction, selection and use of test and other evaluation instruments to control those factors that have adverse effect on validity and interpretation of results.

SELF-ASSESSMENT EXERCISE

i. Define the following terms
   i. Content Validation
   ii. Criterion related Validation
   iii. Construct Validation
ii. What are the three main concerns of validity?
iii. What are the factors that affect the value of criterion validity index?

SELF ASSESSMENT EXERCISE

A1 Definitions of

i. **Content Validation:** This is the process of determining the extent to which a set of test tasks provided a relevant and representative sample of the domain of tasks under consideration.
ii. **Criterion related Validation:** It is defined as the process of determining the extent to which test performance is related to some other valued measure of performance.
iii. **Construct Validation:** It is defined as the process of determining the extent to which test performance can be interpreted in terms of one or more psychological constructs.

A2 The three main concerns of validity are:

i. determining the extent to which performance on a test represents level of knowledge of the subject matter content which the test was designed to measure – content validity of the test.
ii. determining the extent to which performance on a test represents the amount of what was being measured possessed by the examinee – construct validity of the test.
iii. determining the extent to which performance of a test represents an examinee's probable performance on some other related test or task-criterion (concurrent and predictive) validity of a test.

A3 The factors that affect the value of criterion validity index are:

i. the validity of the criterion variable itself (validity of the instrument used in obtaining criterion scores)
ii. the time interval between the administration of the two test (predictor and criterion measures)
iii. the selection of criterion group based on performance on the predictor test scores which results in a homogeneous group that is no more a group which inference based on test validity is to be made.

4.0 CONCLUSION

In this unit you learned about validity and types of validity. Furthermore you learned about content validation, criterion validation and construct validation. In addition you learned validity of criterion reference mastery tests. Finally you learned about factors that influence validity.

5.0 SUMMARY

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

The three concerns of validity are:

- determining the extent to which performance on a test represents level of knowledge of the subject matter content which the test was designed to measure (content validity of the test);
- determining the extent to which performance on a test represents the amount of what was being measured possessed by the examinee (construct validity of the test); and
- determining the extent to which performance on a test represents an examinee's probable performance on some other related test or task (criterion validity of a test).
- Content Validation is the process of determining the extent to which a set of test tasks provided a relevant and representative sample of the domain of tasks under consideration.
- Criterion Validation is the process of determining the extent to which test performance is related to some other valued measure of performance.
- Correlation coefficient expresses the degree of relationship between two sets of scores by numbers ranging from +1.00 to -1.00.

- Two common methods of computing correlation coefficients are:
  - Spearman Rank-Difference Correlation
  - Pearson Product-Moment Correlation
- Construct Validation is the process of determining the extent to which test performance can be interpreted in terms of one or more psychological construct.
- A construct is a psychological quality that are assumed exists in order to explain some aspect of behaviour
- Criterion – referenced mastery tests are not designed to discriminate among individuals. Therefore statistical validation procedures play a less prominent role.
- Many factors influence the validity of a test. These factors include factors in the test itself, factors in the test administration, factors in the examinee's response, functioning content and teaching procedures as well as nature of the group and the criterion.

6.0 TUTOR-MARKED ASSIGNMENT

1. What is Validity of a test?
2. In a class of 20 students the following scores were obtained in the first continuous assessment test in Economics and Geography:

Economics	98	97	95	94	93	91	90	89	88	87	86	84	83	81	80	79	77	76	75	74
Geography	76	75	72	70	68	66	64	60	58	57	56	54	52	50	48	46	44	45	62	77

Using Spearman rank-difference method, calculate the correlation coefficient between the two set of scores obtained by the students of this class.

3. Explain with the aid of a diagram how to interpret correlation coefficient (r) Values

7.0 REFERENCES/FURTHER READING

Gronlund, N.E. (1985). Measurement and Evaluation in Teaching, New York: Macmillan Publishing Company.

Nenty, H. J. (1985). Fundamentals of Measurement and Evaluation in Education. Calabar. University of Calabar.

UNIT 5 PROBLEM OF MARKING TEST AND QUALITY CONTROL IN MARKING SYSTEM

CONTENTS

1.0 INTRODUCTION

From Module 3 Unit 1 of this course we have been discussing test and test related issues. In this Unit we shall round-up our discussion with the problem of marking tests and quality control in marking system. Specifically, you will learn about marks and marking, types of marking systems namely the traditional marking system, the pass-fail system and checklist of objectives. Furthermore, you will learn quality control: current marking systems – multiple marking systems and guidelines for developing a multiple marking system. Finally, you will learn about assigning letter grades which involves determining what to include in a grade, combining data in assigning grades, selecting the proper frame of reference for grading and determining the distribution of grades.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain the meaning of mark and marking systems;
- identify the various types of marking systems and their peculiar characteristics,
- state the importance of the multiple marking system;
- enumerate the guidelines for developing a multiple marking system;
- assign letter grades to marks using the range and the stanine weighting systems; and
- select the proper frame of reference for grading

3.0 MAIN CONTENT

3.1 Marks and Marking

Marks and marking have been very deeply rooted in our educational system. This is to the extent that they have become the basis in whole or in part for a wide range of actions and decisions within a given educational institution, between levels in the educational structure, and in the relations of the educational system in the outside world. For instance, eligibility for admission to certain programmes, for scholarship and, for continuing in school …are determined by academic standing. For these reasons, marking system have been so desirable, durable and so resistant to change.

Marks with all these technical limitations remain one of the best predictors of later achievement and so are important in conveying information about likelihood of success in college generally or in specific institutions or programmes. Marks have a significant role to play in informing the individual as well as the institution to which he may later apply of his prospects for academic success. Marking practices on the other hand are expressions of individual and group value systems as much as they are impartial report of student behaviour. Nevertheless the marking practices are faced with some challenges such as:

- Should students be penalized for handling in paper late or get extra credit for doing optional work?
- Should students with the same level of scholastic aptitude get on the average the same grades irrespective of the subject or department involved?

These are questions of value that rarely are examined in detail and upon which members of a faculty seldom come to genuine agreement. As

long as disagreement exists within the local educational culture on questions of value such as these, the technical efforts of the psychometrician to introduce a consistent rational into grading practices will be ineffectual.

3.2 Types of Marking Systems

Any effective system of marking has to satisfy two principles which are:

- providing the type of information needed, and
- presenting it in an understandable form.

3.2.1 Traditional Marking Systems

In this system a single letter grade such as A, B, C, D, F or a single number such as 5, 4, 3, 2, 1 is assigned to represent a learner's achievement in each subject. This system although brief but comprehensive in expression and convenient has several defects which are:

- the meaning of such marks is often not easy to understand because they are a conglomerate of diverse factors like achievement, effort and good behaviour;
- the interpretation of such marks is difficult even when it is possible to limit the mark to achievement only ; and
- letter grades as typically used have resulted in an undesirable emphasis on marks as ends in themselves.

Moreover, the lack of information to clarify the single letter grade may likely contribute to this misuse. As a result of these, letter grades are viewed as goals to be achieved rather than as means for understanding and improving the student learning.

3.2.2 Pass-Fail System

The Pass-Fail system is an alternative to the traditional letter grade. In this case, students are permitted to take some elective courses under pass-fail option since the pass-fail grade is not included in their grade point average. Thus, the system encourages students to explore new areas of study without the fear of lowering their grade – point average. The pass-fail system offers less information than the traditional system. It can also encourage students to reduce their study efforts in these courses by shifting study time from these courses to those in which letter grades are to be assigned. However, these limitations can be minimized if the number of courses taken under the pass-fail option is restricted and made compulsory. The pass-fail grade marking system is typical of

courses taught under a pure mastery learning approach. In this case learners are expected to demonstrate mastery of all course objectives before receiving credit for a course and a simple pass is all that is needed to indicate mastery. This pass-fail no grade system when used, nothing is recorded on a learner's school record until mastery of the course is demonstrated.

3.2.3 Checklists of Objectives

This is a means of supplementing the traditional marking system with a list of objectives to be checked or rated. This provides more informative progress report. Some schools supplement the traditional marking system with a list of objectives to be checked or rated. These reports typically include rating of progress towards the major objectives in each subject matter area. There are variations in the symbols used for the rating from school to school. While some school retain the traditional A, B, C, D, F lettering system other more commonly shift to fewer symbols such as O (outstanding), S (Satisfactory) and N (Needs improvement). This form of reporting has the advantage of presenting a detailed analysis of learners' strengths and weaknesses. Moreover, they are used to highlight the objectives. But, keeping the list of objectives down to a workable number and stating them in terms that are understood by all users is a problem.

3.3 Quality Control: Current Marking Systems

You know that quality issues and quality assurance cannot be overlooked in any measurement procedure. Hence, the rigorous processes of validity and reliability in the process of test development and administration. These quality issues are necessarily imbedded in every step and every segment of the current marking systems to overcome the problems of the traditional marking system. Traditional marking system has always played a major role in marking and reporting learner's progress. The letter grades are favoured more than any other system. This is evidenced by its frequent use by teachers perhaps because of the ease with which such marks can be assigned may have contributed to their widespread use. Despite the widespread use and for Quality Control, some teachers use more than one method. This is because a multiple marking and reporting system tends to combine the advantages of each marking and reporting system and to overcome some of the inherent limitations of using any single marking and reporting method.

3.3.1 Multiple Marking System

The traditional letter grades (A, B, C, D, F) have been in use for several decades, despite efforts to replace them with a more meaningful system. Their continued patronage indicates that they are serving some useful functions in the school. Moreover, they are a simple and convenient means of maintaining permanent school records. Hence the emphasis has moved from replacing them to trying to improve on them by supplementing with more detailed and meaningful reports of learners progress. This has led to the use of multiple marking systems. The multiple marking systems retain the use of traditional marking (letter and number grades) and supplement the marks with checklists of objectives. In some cases, two marks are assigned to each subject, one for achievement and the other for effort, improvement or growth.

3.3.2 Guidelines for Developing a Multiple Marking System

There is no marking system that can be found equally satisfactory in all schools. Therefore each school has to develop methods that fit it own particular needs and peculiarities. In general, the following principles for devising a multiple marking and reporting system are guidelines for this purpose. The development of the marking and reporting system should be:

i. Guided by the functions to be served: This typically requires a study of the functions for which the resulting marks are to be used. A satisfactory compromise is then made to accommodate them as much as possible. The letter grade should be retained as a pure measure of achievement while a separate report could supplement course objectives and other desired outcomes.
ii. Developed corporately by parents, pupils and school personnel: This is done by a committee consisting of representative of these various groups. This creates avenue for the groups to make useful contribution towards evolving adequate marking system and understanding of the marking system while in use.
iii. Based on a clear statement of educational objectives: The same objectives that have guided instruction and evaluation should serve as a base for marking.
iv. Based on adequate evaluation: The need for objective evaluation of performance makes it necessary to take into account the types of evaluation data needed at the planning stage of a marking system. The data to be collected and used should contain valid and reliable information.
v. Detailed enough to be diagnostic and yet compact enough to be practical: This can be achieved by supplementing the letter grade system with more detailed report on other aspects of pupil

development. The detailed report will bring out the picture of their strengths and weaknesses as much as possible.

vi. Provide for parent–teacher conferences as needed. These serve as supplement to the marking system in use and are typically arranged as needed.

3.4 Assigning Letter Grades

Effective use of A, B, C, D, F marking system calls for answers to questions such as:

- What should be included in a letter grade?
- How should achievement data be combined in assigning letter grades?
- How should the distribution of letter grades be determined?
- What frame of reference should be used in grading?

Attempts are made below to discuss answers to the questions.

3.4.1 Determining What to Include in a Grade

Letter grades have been said to be most meaningful and useful when they represent achievement alone. When they are combined with various aspects of learner's development, they lose their meaningfulness as a measure of achievement.

In addition, they suppress aspects of development. The description of student learning and development can be enhanced if the letter grades are made as pure a measure of achievement as possible while reporting on other aspects of learning separately. Moreover, if letter grades are to serve as valid indicators of achievement, they must be based on valid measure of achievement.

This involves the process of defining the course objectives as intended learning outcomes and developing or selecting tests and other evaluation devices that measures these outcomes most directly. How much emphasis should be given tests, rating and other measures of achievement in the letter grades is determined by the nature of the course and the objective being stressed.

Also, the type of evaluation data to be included in a course grade and the relative emphasis to be given to each type of evidence are determined primarily by examining the instructional objectives. The more important the objective is, the greater the weight it should receive in the course grade. Thus, letter grades should reflect the extent to which learners

have achieved the learning outcomes specified in the course objectives, and these should be weighted according to their relative importance.

3.4.2 Combining Data in Assigning Grades

After determining what to include in a letter grade and the relative emphasis to be given to each aspect, the next step is to combine the various elements so that each element receives its intended weight. For instance, if we decide that the final examination should count 40 percent, the midterm 30 percent, laboratory performance 20 percent and written reports, 10 percent, then our course grades will be designed to reflect this emphasis. A typical procedure is to combine the elements into a composite score by assigning appropriate weights to each element and then use these composite scores as a basis for grading.

To weight composite in a composite score, the variability of the scores must be taken into account. This is illustrated by the simple example that follows. In this example we assume that we are to combine scores on a final examination and a term report and we want them to be given equal weight. If our range of scores on the two measures are as follows:

i. The range weighting system

	Range of Scores
Final Examination	60 to 100
Term Report	20 to 40

The range of scores provides a measure of scores variability or spread and this is used to equate the two sets of scores. The final examination and the term report is given equal weight in the composite score by using a multiplier that makes the two ranges equal. In the above example, the final examination scores have a range of 40 (100-60) and the term report scores a range of 20 (40 – 20). Thus, we need to multiply each term report by 2 to obtain the desired equal weight. On the other hand, if we wanted our final examination to count twice as much as the term report, then we need not multiply each of the term report by 2 because the final examination range is already twice that of the term report. Also, if our desire is for the term report to count twice as much as the final examination, then we would have to multiply each term report by 4. The range system is satisfactory for most classroom purposes.

ii. A more refined weighting system can be obtained by using the standard deviation as the measure of variability.

iii. The stanine weighting system: The components in a composite score can also be weighted properly by converting all sets of

scores to stanines (standard scores, 1 through 9) as we did in unit 2 of this module. When all scores have been converted to the same stanine system, the scores in each set will have the same variability. They are then weighted by simply multiplying each stanine score by the desired weight. Using this system, a learner composite score is determined thus:

Desired Weight	Learner's Stanine	Weighted Score
Examination	2	8
16		
Laboratory Work	2	9
18		
Written Reports	1	6
6		

These composite scores can be used to rank learners according to an overall weighted measure of achievement in order to assign letter grades.

3.4.3 Selecting the Proper Frame of Reference for Grading

Letter grades are typically assigned on the basis of norm referenced frame of reference or both norm and criterion referenced frames of reference. You know that assigning grades on a norm-referenced basis involves comparing a learner's performance with that of a reference group, typically one's classmates. Therefore in this system, the learner's relative ranking in the total group determines the grade rather than by some absolute standard of achievement. This grading, because it's based on relative performance, the grade is influenced by both the learner's performance and the performance of the group. Thus, a learner will fare much better grade wise in a low-achieving group than in a high achieving group. The disadvantage of norm-referenced grading is that it has shifting frame of reference (where grades depend on the group's ability). Despite this disadvantage, it is widely used in the schools because much of classroom testing is norm-referenced. The tests having been designed typically to rank learners in order of achievement rather than to describe achievement in absolute terms.

You also know that assigning grades on a criterion-referenced basis involves comparing a learner's performance to pre-specified standards set by the teacher. These standards are usually concerned with the degree of mastery to be achieved by the learners and may be specified as:

- tasks to be performed (type 60 words per minute without error); or

- the percentage of correct answers to be obtained on a test designed to measure a clearly defined set of learning tasks.

Hence, in this system letter grades are assigned on the basis of an absolute standard of performance rather than on a relative standard. If all learners demonstrate a high level of mastery, all will receive high grades. The use of absolute level of achievements as a basis for grading as implied in criterion-referenced system requires that:

- the domain of learning tasks be clearly defined;
- the standard of performance be clearly specified and justified; and
- the measures of learner's achievement be criterion referenced.

These conditions are difficult to meet except in a mastery-learning situation. Usually, when mastering learning is the goal, the learning tasks tend to be more limited and easily defined. Moreover, percentage –correct scores, which are widely used in setting absolute standards, are most meaningful in mastery learning since they indicate how far a learner is from complete mastery.

3.4.4 Determining the distribution of Grades

There are two ways of assigning letter grades to measure the level of learner achievement. These are: the norm-referenced system based on relative level of achievement and the criterion-referenced system based on absolute level of achievement.

Norm-Referenced Grading: Essentially, the ranking of norm-referenced grades is a matter of ranking the learners in order of overall achievement and assigning letter grades on the basis of each learner's rank in the group. The ranking might be limited to a single classroom group or on combined distributions of several classrooms groups taking the course. The proportion of As, Bs, Cs, Ds and Fs to be used are pre-determined before letter grades can be assigned. Grades can be assigned on the basis of the normal curve. Grading on the normal curve results in an equal percentage of As and Fs, and Bs and Ds regardless of the group's level of ability. That is, the proportion of high grades is balanced by an equal proportion of low grades. The limitations of assigning grades on the basis of the normal curve are that:

- the groups are usually small to yield a normal distribution;
- classroom evaluation instruments are usually not designed to yield normally distributed scores; and
- the learner population becomes more select as it moves through the grades and the less-able learners fail or drop out of school.

Thus, grading is seldom defensible. Grading on normal curve can be defended only when a course or combined courses have a largely and unselected group of learners. Nevertheless, the credibility of basing the decision concerning the distribution of grades on a statistical model (normal curve) rather than on a more rational basis is still questionable.

A credible approach is to have the school staff set general guidelines for the approximate distributions of marks for the letter grades. This might involve separate distributions for introductory and advanced courses, for gifted and slow learning classes and so on. That is, the distributions should be flexible enough to allow for variation in the caliber of learners from one course to another and from one time to another in the same course. This entails indicating ranges rather than fixed percentages of learners to receive each letter grade. There is no simple or scientific means of determining the ranges for a given situation. The decision is to be made by the school staff by taking into account the school's philosophy, the learner population and the purposes of the grades. A hypothetical suggestion is presented for your understanding and assimilation of the assigning grades.

A = 10 to 20 percent of learners
B = 20 to 30 percent of learners
C = 30 to 50 percent of learners
D = 10 to 20 percent of learners
F = 0 to 10 percent of learners

You should note that the distribution should provide for no failing grades.

This is because whether learners pass or fail a course should be based on their absolute level of learning rather than, their relative position in some group. That is, even when grading is done on a relative basis, the pass-fail decision must be based on an absolute standard of achievement if it is to be educationally sound.

Criterion-Referenced Grading: This system of grading is most useful when a mastery learning approach is used because mastery learning provides the necessary conditions for grading on an absolute basis. The process includes:

- delimiting the domain of learning tasks to be achieved;
- defining the instructional objectives in performance terms;
- specifying the standards of performance to be attained; and
- measuring the intended outcomes with criterion-referenced instruments.

In the criterion-referenced grading system, if the course's objectives have been clearly specified and the standards for mastery appropriately set, the letter grades may be defined as the degree to which the objectives have been attained as indicated below:

A = Outstanding. Learner has mastered all of the course's major and minor instructional objectives.
B = Very Good. Learner has mastered all of the course's major instructional objectives and most of the minor objectives.
C = Satisfactory. Learner has mastered all the course's major instructional objectives but just a few of the minor objectives.
D = Very Weak. Learner has mastered just a few of the course's major and minor instructional objectives and barely has the essentials needed for the next highest level of instruction. Remedial work would be desirable.
F = Unsatisfactory. Learner has not mastered any of the course's major instructional objectives and lacks the essentials needed for the next highest level of instruction. Remedial work is needed.

Furthermore, if the test and other evaluation instruments have been designed to yield scores in terms of the percentage of correct answers; criterion-referenced grading then might be defined as follows:

A = 95 to 100 percent correct
B = 85 to 94 percent correct
C = 75 to 84 percent correct
D = 65 to 74 percent correct

Here also, you should note that defining letter grades in this manner is defensible only if the necessary conditions of criterion –referenced system have been met. In general, the distribution of grades in criterion-referenced grading systems is not predetermined. If all learners demonstrate a high level of mastery, all will receive high grades. Also, if some learners demonstrate a low level of performance, they will receive low grades. Hence, the distribution of grades is determined by each learner's absolute level of performance, and not by the learner's relative position in the group.

The two procedures for determining letter grades in a criterion-referenced system are:

i. **The one-short system.** This provides a single opportunity to achieve the pre-specified standards. In this system the learner is assigned whatever grade is earned on the first attempt. This results in some failing grades.

ii. **The Repeated-Attempts System.** This procedure permits the learner to make repeated attempts to achieve the pre-specified standards.The learner is given corrective help and enough additional learning time as he progresses in the learning process to achieve a satisfactory level of mastery. This system systematically eliminates failure. Typically, only the letter grades A, B, C are used and learners are permitted to repeat examinations until a satisfactory level of performance is achieved.

In the criterion-referenced system, the letter grades are supplemented with comprehensive report which consists of a checklist of objectives to inform learners and parent the progress made by the end of the marking period.

SELF-ASSESSMENT EXERCISE

i. Identify the various types of marking systems.
ii. What factors should a teacher consider while considering what to include in a letter grade?
iii. Differentiate between the norm-referenced grading and the criterion referenced grading.

SELF-ASSESSMENT EXERCISE

A1: The various types of marking systems are:
- the traditional marking system;
- the pass-fail system
- the checklist of objectives

A2: The factors to be considered by a teacher while considering what to include in a letter grade are determining:
- What to include in a letter grade;
- How to combine the various achievement data into a composite;
- What frame of reference to be used, and
- What distribution of letter grades to use.

A3: To differentiate between the norm-referenced grading and the criterion-referenced grading.
Norm-referenced grading essentially involves ranking the learners in order of overall achievement and assigning letter grades on the basis of each learner's rank in the group. On the other hand, the distribution of grades in the criterion-referenced grading system is not predetermined; rather the distribution of grades is determined by each learner's absolute level of performance and not by the learner's relative position in the group.

4.0 CONCLUSION

In this unit you learned about marks and marking systems, types of marking systems namely the traditional marking systems, pass-fail system and checklists of objectives. Furthermore, you learned quality control: Current marking systems, multiple marking system and guidelines for developing a multiple marking system. In addition you learned about assigning letter grades to marks. Finally you learned how to determine what to include in a grade, combining data in assigning grades, selecting the proper frame of reference for grading and determining the distribution of grades.

5.0 SUMMARY

- Marks with all these technical limitations remain one of the best predictors of later achievement and so are important in conveying information about the learners progress.
- Marking practices are expressions of individual and group value systems.
- Types of marking systems include:
  - the traditional marking system;
  - pass-fail system; and
  - checklist of objectives
- The multiple marking system retains the use of traditional marking system and supplemented the marks with checklists of objectives.
- To assign letter grades the following issues have to be addressed:
  - What should be included in a letter grade?
  - How should achievement data be combined in assigning letter grades?
  - What frame of reference should be used in grading?
  - How should the distribution of letter grade be determined?

6.0 TUTOR-MARKED ASSIGNMENT

1. Identify the various types of marking systems.
2. What factors should a teacher consider while considering what to include in a letter grade?
3. Differentiate between the norm-referenced grading and the criterion referenced grading.
4. What are the principles that must be satisfied by any effective system of marking?
5. Enumerate the general principles for developing a multiple marking and reporting System.

7.0 REFERENCES/FURTHER READING

Gronlund, N.E. (1985). Measurement and Evaluation in teaching. New York Macmillan Publishing Company.

Thorndike, R.L and Hagen, E.P (1977).New York Measurement and Evaluation in Psychology. 4e John Wiley and Sons.