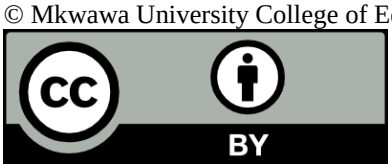
Mkwawa University College

Educational Measurement and Evaluation



Citation:
Mkwawa University College of Education (MUCE). (2013). Educational Measurement and Evaluation. Iringa: MUCE.

MUCE welcomes feedback on these materials and would like to hear from anybody who has used them as is or used and adapted them or who would be interested to work with MUCE more generally.

PRINCIPAL ADDRESSES

PRINCIPAL	Tel. 026-2702751
P.O. BOX 2513	Fax: 026-2702751
IRINGA	E-mail: principal.muce@udsm.ac.tz
DEPUTY PRINCIPAL (ACADEMIC)	Tel. 026-2701192
P.O. BOX 2513	Fax: 026-2702751
IRINGA	E-mail:dpacademic.muce@udsm.ac.tz
DEPUTY PRINCIPAL (ADMINISTRATION)	Tel. 026-2701191
P.O. BOX 2513	Fax: 026-2702751
IRINGA	E-mail:dpadministration.muce@udsm.ac.tz
DEAN FACULTY OF EDUCATION	Tel. 026-2700630
P.O. BOX 2513	Fax: 026-2702751
IRINGA	E-mail:dpacademic.muce@udsm.ac.tz
DEAN FACULTY OF HUMANITIES AND SOCIAL SCIENCES	Tel. 026-2700636
P.O. BOX 2513	Fax: 026-2702751
IRINGA	E-mail:dpacademic.muce@udsm.ac.tz
DEAN FACULTY OF SCIENCE	Tel. 026-2700632
P.O. BOX 2513	Fax: 026-2702751
IRINGA	E-mail:dpacademic.muce@udsm.ac.tz

Course: **Educational Measurement and Evaluation**

Introduction

This course aims at introducing you to basic concepts of educational measurement, monitoring, assessment and evaluation. Furthermore, it seeks to equip you with basic knowledge and skills that are important for developing tools for measurement, assessment, and monitoring of educational attainments and institutional performance.

Learning outcomes

By the end of this course, you should be able to:

- develop an appreciation for the need to measure and evaluate educational outcomes
- develop skills in the construction of different test items and measurement scales
- develop abilities to analyse test/examination results and other measures of pupil characteristics.
- develop an understanding of the use and misuse of examinations.
- develop a critical understanding of the practice of examinations in Tanzania.

Course Outline

Module 1: Basic Concepts in Measurement and evaluation

1.1 Measurement, Evaluation, Testing
1.2 Assessment and Monitoring
1.3 Purpose of Evaluation
1.4 Instructional objectives and evaluation
1.5 Taxonomies of Educational Objectives

Module 2: Principles of Test Construction

2.1 Consequences of Testing or not Testing
2.2 Purpose of Testing
2.3 Tables of specification of instructional objectives
2.4 Characteristics/qualities of a good Tests
2.5 Classification of tests
2.6 Constructions of test items

Module 3: Assembling, Administration and Analysis of Test Results

3.1 Assembling of classroom tests

3.2 Test administration and marking
3.3 Summarizing test results
3.4 Item analysis: level of difficult and discrimination
3.5 Reporting test performance

Module 4: Assessment of Non-Cognitive outcomes and IQ

4.1 Classroom observation techniques
4.2 Peer appraisal and self assessment
4.3 Measurement of attitude, interests, and personality traits
4.4 Intelligence and aptitude tests

Module 5: Examination System in Tanzania

5.1 Historical Perspectives
5.2 Types of Examination in Tanzania
5.3 Methods of establishing standards: score equivalence
5.4 Standards across countries and times periods

Course Evaluation :

- **Coursework 40%**
- **Final Examination 60%**

References

Core readings

1. Airasian, P.W. (2001). *Classroom assessment: Concepts and applications* (4th ed.). NY: McGraw Hill.
2. Cohen, R.J. & Swerdlik, M.E. (2005). *Psychological testing and assessment*. NY: McGraw Hill.
3.
4. Ingule, F.O., Rono, R.C., Ndambuki, P.W. (1996). *Introduction to educational psychology*. Nairobi: East African Educational Publishers
5. Ebel R. L. and Frisbie, (1991), D.A. ***Essentials of Education Measurement***. New York: Prentice Hall.
6. Gronlund, N.E. & Linn, R.L. (1990). ***Measurement and Evaluation in Teaching*** N.Y. Macmillan,.

6. Sax, G. (1997). *Principles of educational and psychological measurement and evaluation*. London: Wadsworth Publishing Company
7. Omari, I.M. (1995). Conceptualizing quality in Primary Education**, *Papers in Education and Development, 16, 25 – 48.***
8. Gronlund N.E(1985). ***Stating Objectives*** New York: Macmillan.
9. Other relevant books in testing, measurement, assessment and evaluation.

Module 1: Basic Concepts in Measurement and Evaluation

Overview

Testing, measurement and evaluation are important activities in the teaching and learning process. Without them, it would have been very difficult for teachers to track and follow-up their students' achievement. Equally important, are the concepts monitoring and assessment. These two additional concepts supplement/complement the roles played by the other concepts.

In this module, the five concepts are defined and clarified so that their similarities and differences are addressed when it comes to their use in educational settings.

Learning outcomes

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

- define basic concepts: test, measurement assessment, monitoring and evaluation.
- distinguish various types of evaluation in education.
- explain the purposes of assessment and evaluation in teaching and learning.
- state basic principles of evaluation.
- use the taxonomies of educational objectives in stating instructional objectives and designing assessment tasks.

Learning Unit 1.1: Measurement, Evaluation and other related Concepts

Introduction

Different people have different understandings about the concepts testing, measurement, assessment, monitoring and evaluation. In most cases, most of these people have been using the concepts as though they mean the same thing. Although the concepts are related when it comes to their use in education or rather instructional settings, they are distinct. It is important for you to realise their differences and similarities so as to use them appropriately and hence avoid misconceptions. Therefore, in this learning unit, you are introduced to the concepts by highlighting their definitions, differences and similarities. Also the unit takes you through the interrelationships among the concepts. But before exposing you to the concepts, go through a Case Study 1. This case study reveals what different people understand about these important concepts.

Case study 1.1

One day, tutors at Mkanyageni Teachers College were arguing about which concept was more appropriate to use when addressing different areas pertaining to their students' achievement. One of them said that "If I want to determine how good my students are in my subject, then measurement would be appropriate to use". Another one argued that "For me, if I want to determine my students' achievement, I would use tests". The last one in this conversation claimed that "For me, assessment is more appropriate for that purpose". It was not easy for these tutors to reach an agreement on who was right in this issue.

Having gone through Case Study 1.1, attempt Activity 1.1. To get more support in attempting the activity, go to learning Resource 1.1.

Activity 1.1

Attempt the following question:
Among the three tutors in the Case Study 1, which one was using the appropriate concept that would help him/her get comprehensive data about his/her students' achievement? Why?

Resource 1.1

Definitions, Similarities and Differences

- Test, measurement and evaluation – distinct but closely related terms – sometimes all of them can be involved in a single process.
- For example, if we ask students to answer a series of questions concerning science/geography, obtain their scores by counting the number of correct answers, and conclude that the students are making good learning progress, we are concerned with all the three concepts.

What is a test?

- It is an instrument or systemic procedure for measuring a sample of behaviour.
- It helps to tell us "How well does the individual perform either in comparison with others or in comparison with a domain of performance tasks?"

What is measurement?

- This is a process of assigning numbers to individuals or their characteristics according to specified rules.
- It tries to answer the question "How much?"

What is evaluation?

- At classroom level, it can be defined as the systematic process of collecting, analysing and interpreting information to determine the extent to which pupils are achieving instructional objectives.
- It answers the question "How good?"

Relationship between testing, measurement and evaluation

- Evaluation is a more comprehensive & inclusive term than measurement.
- Testing is just one type of measurement.
- Measurement is limited to quantitative description of pupils, results are expressed in numbers, e.g. Musa solved 25 of the 50 mathematics questions, does not include qualitative description such as Musa's score was satisfactory.
- Evaluation may include quantitative descriptions (measurement) and qualitative disjunction (non-measurement) of pupils.
- Evaluation includes value judgment concerning the desirability of the results.
- Figure 1 shows the relationship diagrammatically.

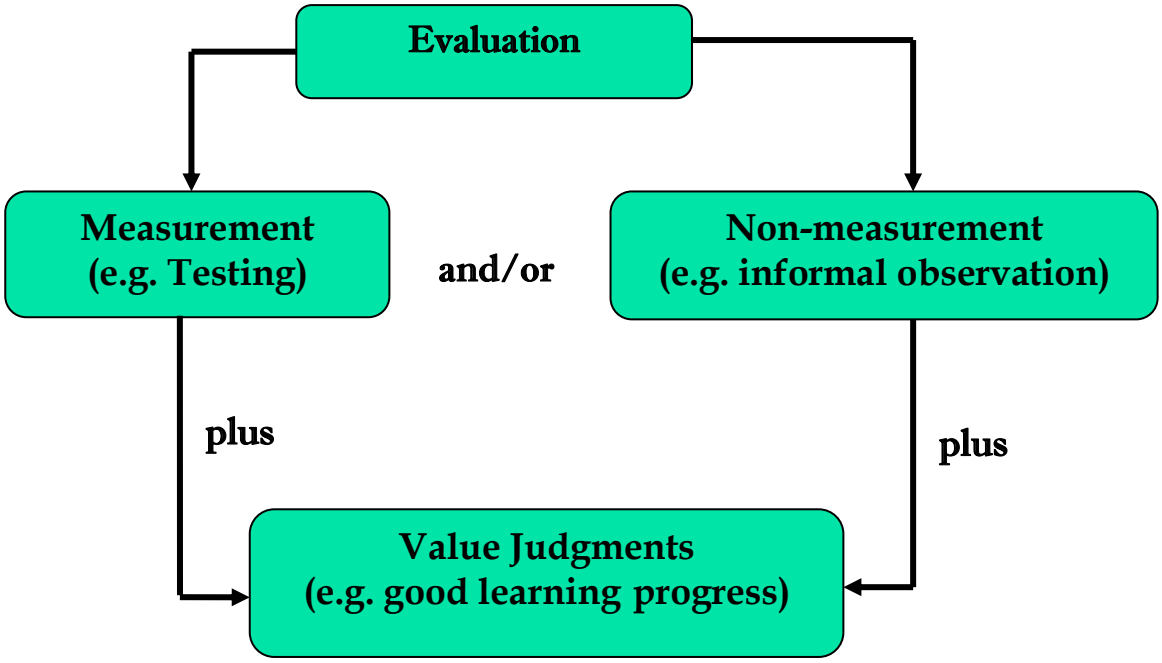


Figure 1.1: Relationship between testing, measurement and evaluation diagrammatically

Assessment versus Evaluation

- The terms assessment and evaluation are related and often used interchangeably – yet they differ when used in an educational or training context.
- Assessment is the systematic, continuous process of monitoring the various pieces of learning to evaluate student achievement and instructional effectiveness – includes tests, homework assignments, class projects, class presentations, class participation and teacher observation.
- In short, assessment means those activities that are designed to measure learner achievement brought about as result of an instructional programme of some sort. Figure 1.2 shows the characteristics of effective assessment.
- Evaluation refers to a broader process that involves examining many components of a whole and making instructional decisions.
- The evaluation process focuses upon determining the attainment of previously-established priorities and goals.
- Evaluation helps document the effectiveness of a course or programme, identifies weaknesses and strengths, and spots areas in need of revision.
- In short, evaluation refers to a series of activities that are designed to measure the effectiveness of the instructional system or a component thereof.
- The two processes are closely related – the results of student assessment constitute one of the most important sets of data in the evaluation of any course or curriculum.

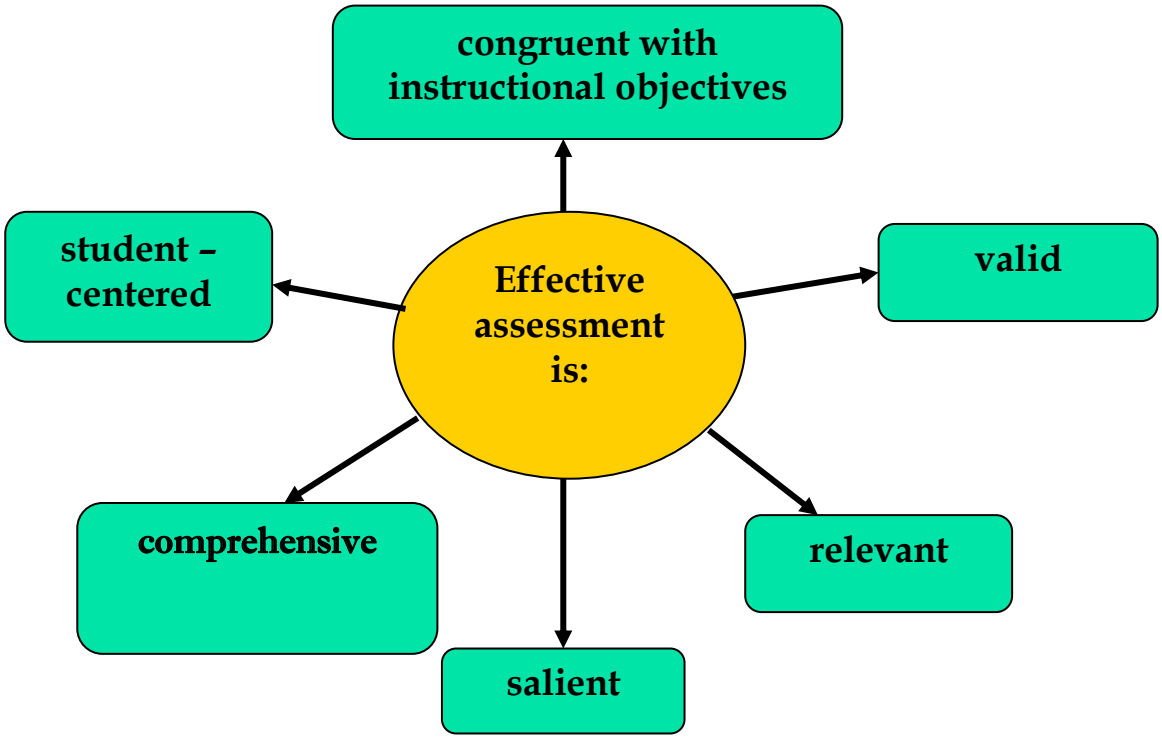


Figure 1.2: Characteristics of effective assessment

General Principles of Evaluation

1. Clearly specifying what is to be evaluated has priority in the evaluation process.
2. An evaluation technique should be selected in terms of its relevance to the characteristics or performance to be measured.
3. Comprehensive evaluation requires a variety of evaluation techniques.
4. Proper use of evaluation techniques requires an awareness of their limitations.
5. Evaluations is a means to an end, not an end in itself.

Activity 1.2

Answer the following questions in not more than one page and then send your answers through the following email address… … …

1. ***Why is it important for you to be able to identify the differences and similarities among the following concepts: test, measurement, assessment, monitoring and evaluation?***
2. ***Critically elaborate this statement "Evaluations is a means to an end, not an end in itself".***

Learning Unit 1.2: Purposes and Types of Evaluation

Introduction

Evaluation as an umbrella term among the five concepts, serves different purposes. Firstly, evaluation can be used to determine the effectiveness of courses and educational programmes. Secondly, evaluation helps to provide a basis for improving courses or programmes. In order to gain some insights about the concept evaluation as it is used in education, go through a Case Study 2.

Case Study 1.2

Mrs. Masasi is a head of one of the private secondary schools in one of the regions in the Southern part of Tanzania. One day, she was talking to one of the parents about different types of exams that are conducted at her school. She told the parent that before the school admits Form 1 pupils, the applicants have to do entry exams. Those who qualify are admitted to join Form 1. During the teaching and learning process, she adds, the pupils are subjected to tests, assignments, projects, presentations and home assignments. She emphasised that these activities are very important in monitoring pupils' progress. She also added that the pupils, who are experiencing persistent learning difficulties, are subjected to remedial programmes. The aim of this programme is to try to minimise the difficulties.

The head of the school said further that, in completing their four years of secondary education, the pupils of this school also sit for national examinations popularly known as the Certificate of Secondary Education Examination (CSEE).

Having read the Case Study 2, try to attempt Activity 1.3.

Activity 1.3

From the Case Study 2, attempt to classify the types of tests that are conducted at the mentioned secondary school.

Resource 1.2

Purposes of evaluation

In broad terms, measurement and evaluation can serve a variety of purposes. The central purposes are:

- to determine the effectiveness of courses and educational programme.
- to provide a basis for improving courses/ programme.

Tests and other evaluation procedures can be classified in terms of their functional role in classroom instruction. The classification of evaluation is as follows:

1. Placement Evaluation

Placement evaluation is concerned with the student's entry behaviour before the beginning of instruction. It seeks to provide answers to the following important questions:

- Does the student have needed knowledge and skills to begin instruction?
- Has the pupil already mastered the objectives of the particular lesson or unit?
- Does the student's ability, learning style, attitude or interest indicate the child would benefit from a particular lesson or unit?

To answer these questions, use variety of techniques: readiness test, aptitude tests, pre-test on course objectives & observational procedures. The goal of this evaluation is to place students in the proper position in the instruction sequence and provide most beneficial method of instruction.

2. Formative evaluation

This evaluation is concerned with the monitoring of learning progress during instruction. It provides continuous feedback to both pupils and teachers concerning learning successes and failures:

- Feedback to pupils provides reinforcement of successful learning and identifies learning errors that are in need of correction.
- Feedback to the teachers provides information for modifying instruction and for prescribing group or remedial work.

Tests used for formative evaluation are most frequently teacher made. Observational techniques are also useful in monitoring pupils' progress and identify successes and challenges in learning**.**

3. Diagnostic Evaluation

Diagnostic evaluation is concerned with the persistent or recurring learning difficulties that are left unresolved by the standard corrective prescriptions of formative evaluation. The main aim of this evaluation is to determine the causes of learning problems and to formulate a plan for remedial action.

4. Summative Evaluation

Summative evaluation comes at the end of course (or unit) of instruction. It aims at determining how students have attained instructional objectives and providing information to grade students and/or to evaluate teacher effectiveness. Its general purpose is grading students or certification of student achievement or providing information for judging the appropriateness of the course objectives and effectiveness of instruction.

Purposes of Assessment

Assessment shapes what students learn and how they learn. It may serve the following functions:

- Determine whether students are learning what we are expecting them to learn.
- Motivate and help students structure their academic efforts.
- Help the teacher understand how successfully he is presenting the material.
- Reinforce learning by showing students what topics or skills they have not yet mastered and should concentrate **on.**

Learning Unit 1.3: Instructional Objectives and Evaluation

Introduction

Instructional objectives are definitely important in teaching. Without them, teaching is reduced to an endeavor with no definite goal, structure, or purpose. Instructional objectives serve as goals that teachers have set in the achievement of a greater goal. They also tell students what is expected of them. Instructional objectives make definite the direction in which teaching leads and become the focus of instruction, not only for the teachers, but also for the students. Without instructional objectives teaching is comparable to a fallen leaf whose destination is dependent on the will of the wind. Without instructional objectives, teachers will have nothing to follow in order to achieve what it should achieve. Therefore, in this unit you will learn in detail the issues related to instructional objectives. To start with, take a look at the Case Study 1.3 and then through it, attempt Activity 1.4.

Case Study 1.3

Mr. Chakubanga, a mathematics teacher, writes the following in his lesson plan:

Objectives

1. Understands the multiplication of fractions.
2. Fold papers.
3. Derive rule for multiplication.
4. Have students solve problems on page 25.

"What is this?" a friend asks, pointing to the lesson plan about folding papers in the book.
"Oh, I have them fold papers into thirds, we write the fractions on the board, and then I have them take a half of the third, write it again, and they see that one-half of one-third is one-sixth and then fold the papers into some other fractions. They see a pattern and we derive the rule for multiplying fractions. Then I have them practice, and on their tests they have to solve problems like those."

Activity 1.4

Look at Mr. Chakubanga's objective # 1. Is the objective acceptably written?

Resource 1.3

What are instructional objectives?
Instructional objectives are statements of desired learning outcomes. They describe ***skills***, ***behaviours***, and ***attitudes*** that students should be able to demonstrate after instruction.

Distinction between goals, objectives and outcomes

- Goals broad, general, long-range statements of educational purpose. They are primarily used in policy making and general programme planning.
- Objectives are brief statements that describe desired learning results or intended outcomes**.** They can be stated in two ways**:**
    - General objectives –these are intended outcomes of instruction that have been stated in general enough terms to encompass set of specific learning outcomes.
    - Specific objectives – these are intended outcomes of instruction that have been stated in terms specific and observable student/pupil performance.
- Outcomes – these achieved results. They include behaviours and products generated by learners.

How to state instructional objectives

- Instructional objectives are stated in terms of what we expect students to be able to do by the end of instruction.
- There are four components of an objective:
    - the action verb,
    - conditions,
    - standard,
    - the intended audience (always the student).

The action verb is the most important element of an objective and can never be omitted. The action verb states precisely what the student will do following instruction.

Example
After teaching the topic on the elements of weather we might expect students to be able to:

- list the elements of weather.
- identify the elements of weather.
- distinguish among the elements of weather.

Remember
Don't state objectives in terms of:

- teacher performance (e.g. teach pupils elements of weather)
- learning process (e.g. pupil learns elements of weather)
- course content (e.g. student studies elements of weather).
- two objects (student lists and explains elements of weather)

Learning Unit 1.4: Taxonomy of Educational Objectives

Introduction

Dear student, in Learning Unit 1.3 you learned about instructional objectives and their importance in the teaching and learning process. In the present unit, you are going to learn about what is called Taxonomy of Educational Objectives (TEO). This is a useful guide for developing comprehensive list of instructional objectives. It attempts to identify and classify all possible educational outcomes. The taxonomy is classified into three major domains:

- Cognitive domain (Bloom et al., 1956)
- Affective domain (Krathwohl, 1964)
- Psychomotor domain (Simpson, 1972).

Resource 1.4

Major Categories in the Cognitive Domain

Cognitive domain which is popularly known as Bloom's (1956) Taxonomy of Educational Objectives, is the most renowned descriptions of the levels of cognitive performance. The levels of this taxonomy are considered to be hierarchical. That is, learners must master lower level of objectives first before they can build on them to reach higher level objectives. The levels of the Taxonomy and examples of verbs and instructional objectives are given hereunder.

1. **Knowledge:** This is defined as remembering of previously learned material. Represents the lowest level of learning outcomes.
   - Sample verbs: write, list, label , name, state, define.
   - Example: List down 6 levels of Bloom's taxonomy of the cognitive domain.

2. **Comprehension:** This level refers to the ability to grasp the meaning of the material.
   - Sample verbs: explain, summarise, paraphrase, describe, illustrate.
   - Example: The student will summarise the uses of Bloom's taxonomy of cognitive domain.

3. **Application***:* This refers to the ability to use knowledge or principles in new or real life situations. The learner at this level solves practical problems by applying information comprehended at the previous level.
   - Sample verbs: use, compute, solve, demonstrate, apply, construct.
   - Example: The student will apply Newton's first law of motion to…….

4. **Analysis:** This level refers to the ability to break down material into its component parts so that its organisational structures are known. Also it may refer to breaking down complex information into simpler parts.
   - Sample verbs: analyse, categorise, compare, distinguish, contrast, separate.
   - Example: Students will compare and contrast the knowledge and comprehension levels of cognitive domain.

5. **Synthesis** – refers to the ability to put parts together to form a new whole. In other words, it consists of creating something that did not exist before.
   - Sample verbs: Create, design, hypothesise, invent, develop
   - Example: The student will design a classification scheme for writing educational objectives that combines the cognitive, affective, and psychomotor domains

6. **Evaluation:** This is the highest level of Bloom's hierarchy. It is concerned with the ability to judge the value of material for a give purpose.
   - Sample verbs: judge, recommend, critique.
   - Example: Students will determine the importance of Bloom's taxonomy in the teaching and learning process.

Major categories in the affective domain

The affective domain (Krathwohl, Bloom, & Masia, 1973) includes aspects such as feelings, values, appreciation, enthusiasm, motivation, and attitudes. The five major categories are listed from the simplest behaviour to the most complex.

1. **Receiving**: This refers to being aware or attending to something in the environment or students willingness to attend to particular stimuli or phenomena. Represents lowest level of learning outcomes in the affective domain.
   - Sample verbs: asks, chooses, follows, gives, replies, uses.
   - Example: Listen to others with respect.

2. **Responding:** This level is concerned with showing some new behaviours as a result of experience. It emphasises active participation on the part of student. Not only attends to a particular phenomenon but reacts to it in some way(i.e.willingness to respond, e.g. voluntarily reads beyond assignment; satisfaction in responding,e.g. reads for pleasure or enjoyment) .
   - Sample verbs: answers, assists, complies, conforms, practices, etc.
   - Example: Participates in games and sport.

3. **Valuing**: Concerned with the worth or value a student attaches to a particular object, phenomenon or behaviour. This ranges from simple acceptance to the more complex state of commitment. It is based on the internalisation of a set of specified values, while clues to these values are expressed in the learner's over behaviour and are often identifiable.
   - Sample verbs: completes, differentiates, initiates, invites, joins, proposes, shares, demonstrates, etc.
   - Example: Demonstrates belief in the democratic process.

4. **Organisation**: Concerned with bringing together values, resolving conflicts between them and beginning the building of an internally consistent value. The emphasis here is on comparing, relating and synthesising values.
   - Sample verbs: adheres, alters, arranges, compares, defends, integrates, modifies, etc.
   - Example: Recognises the need for the balance between freedom and responsibility in a democracy.

5. **Characterisation by value**: This focuses on value system that controls behavior. It emphasises acting consistently with the new value.
   - Instructional objectives are concerned with the student's general patterns of adjustment (personal, social, emotional).
   - Sample verbs: acts, discriminates, displays, modifies, verifies, practices, etc.
   - Example: Maintains good health habits.

Major categories in the psychomotor domain

The domain includes physical movement, co-ordination and use of motor skills. Development of the skills requires practice and is measured in terms of speed, precision, distance, procedures, or techniques in execution. It contains seven categories listed in order from simplest behaviour to the most complex.

1. **Perception**: Ability to use sensory cues to guide physical activities. This ranges from sensory stimulation through cue selection to translation.
   - Sample verbs: chooses, describes, detects, differentiates, isolates, relates.
   - Example: Relates music to a particular dance step.

2. **Set**: Readiness to act. It includes mental, physical and mental sets. The three sets are dispositions that predetermine a person's response to different situations (sometimes called mindsets). It requires the learner to demonstrate an awareness or knowledge of the behaviours needed to carry out skill.
   - Sample verbs: demonstrate, show, displays, explains, reacts, shows, moves, etc.
   - Example: Shows desire to type efficiently.

3. **Guided response**: The early stage of learning a complex skill, includes imitation and trial and error. Can complete steps involved in the skill as directed. Adequacy of performance is achieved by practicing.
   - Sample verbs: assembles, builds, constructs, dismantles, displays, dissects, fastens, manipulates, measures, etc.
   - Example: Performs a mathematical equation as demonstrated.

4. **Mechanism:** This stage focuses on the ability to perform a complex motor skill. It is the intermediate stage of learning a complex skill. At this level, learned responses have become habitual and the movements can be performed with some confidence and proficiency.
   - Sample verbs: constructs, fixes, cooks, constructs, manipulates, mends, mixes.
   - Example: Rides a bicycle.

5. **Complex over response**: This level involves the ability to perform the complete psychomotor skill correctly. In this case, proficiency is indicated by a quick, accurate and highly coordinated performance, requiring a minimum of energy. This category includes performing without hesitation and automatic performance.
   - Sample verbs: carry out, operate, perform, assembles, builds, manipulates, measures, mends, etc.
   - Example: Operates a machine quickly and accurately.

6. **Adaptation**: In this case skills are well developed and a student can modify motor skills to fit new situation or spcial requirements.
   - Sample verbs: adapt, changes, modify, rearranges, reorganises, revises, alters, etc.
   - Example: Modifies instructions to meet the needs of learners.

7. **Origination:** This involves ability to develop an original skill that replaces the skill as initially learned. It is about creating new movement patterns to fit a a particular situation or specific problem. Learning outcomes emphasise creativity based upon highly developed skills.
   - Sample: creates, designs, originates, combines, composes, constructs, initiates, makes, etc.
   - Example: Develops a new hypothesis.

Activity 1.5

Dear student attempt the following questions so see whether you have been able to understand the Taxonomy of Educational Objectives.

(i) The student volunteers her answer by raising her hand in class. Which domain is best illustrated by this students' behavior?

(ii) You are a physical education teacher. You want your students to think that health and fitness are important and should be a part of their life style. In which domain of Taxonomy of Educational objectives would your goal would be best classified?

Module 2: Principles of Test Construction

Overview

We know that tests shape what students learn and how they learn. The construction of good tests requires specific skills and experience, which are neither easy to acquire nor widely available. This module serves as an introduction to principles of test construction. It offers you knowledge and skills in planning the tests and use of tables of specifications in constructing test items.

Learning outcomes

By the end of this module, you should be able to:

- plan classroom tests.

- construct tests that measure the extent to which students have achieved the learning objectives of the course.
- design test items that evaluate the appropriate level of learning outcomes.
- discriminate among testing methods and choose appropriate measures.
- develop meaningful formative communications with students.
- evaluate classroom tests for reliability and validity.

Learning Unit 2.1 Planning for Classroom Tests

Introduction

The key element to effective achievement testing is careful planning. It provides greater assurance that the test will measure relevant learning outcomes.

Resource 2.1

What is test planning?

Test planning involves the identification and specification of what is to be measured. It involves the following steps:

- Determining the purpose of the test;
- Identify and define the intended learning outcomes;
- Prepare the test specifications;
- Construct relevant test items.

A good test reflects the ***goals*** of the course. It is congruent with the skills that you want students to develop and with the content you emphasize in the class. A test that covers a much broader range of material than that covered in the class will be regarded as unfair by your students, even if you tell them that they are responsible for material that has not been discussed in class.

Table of specification of instructional objectives

What is a table of specifications?

This is a two-way chart which relates the instructional objectives to the course content and specifies the relative emphasis to be given to each type. The purpose of the table of specification is to provide assurance that the test will measure a representative sample of the learning outcomes and the subject matter topics to be measured.

Building a table of specifications

Building a table of specifications includes:

- Preparing a list of instructional objectives: This describes the type of performance the pupils are expected to demonstrate.
- Outlining the instruction content. The amount of detail to include in the content outline is somewhat arbitrary but is should be detailed enough to ensure adequate sampling during test construction and proper interpretations of results.
- Preparing the two way chart by listing the major content areas down the left side of the table.

You need to determine what proportion of the test items should b devoted to each objective and each content area. Here a number of factors should be considered:

- How important is each area in the total learning experience?
- How much time was devoted to each area during instruction?
- What relative importance do curriculum specialists assign to each area?

Activity 2.1

Select a topic from your subject of specialization and prepare a Table of specification that will guide you to assess the achievement of your students on the topic.

Resource 2.2

Learning Unit 2.2: Qualities of a Good Test

Validity

Meaning of validity

Validity refers to the appropriateness of the interpretations made from test scores and other evaluation results, with regard to a particular use. For example, if a test is to be used to describe pupil achievement, we would like to be able to interpret the scores as relevant and representative sample of the achievement domain to be measured, to predict future performance, should be an accurate estimate of future performance. Validity refers to the degree to which test scores serve their intended use. A valid assessment procedure is one which actually tests what it sets out to test, that is, one which accurately measures the behaviour described by the learning outcomes under scrutiny. Obviously, no one would deliberately construct an assessment item to test trivia or irrelevant material, but it is surprising just how often non-valid test items are in fact used. As we will see later in the review of assessment methods, validity-related problems are a common weakness of many of the

more widely-used methods. For example, a simple science question given to 14-year-old schoolchildren ('Name the products of the combustion of carbon in an adequate supply of oxygen') produced a much higher number of correct answers when the word 'combustion' was replaced by 'burning'. The original question had problems of validity in that it was, to some extent, testing language and vocabulary skills rather than the basic science involved.

Test validation

This refers to the procedures used to establish the validity of a test. Validity is viewed as a unitary concept, instead of talking of various types of validity, we talk of the various kinds of evidence. Validity is established by showing evidence for it.

1. Content related validity

This is concerned with how well the sample of test items represents the domain (area) to be measured. This can be done by comparing test items to the test specifications describing the task domain under consideration.

2. Criterion related evidence

How well the test scores predict future performance or estimates current performance on some other measures other than the test itself. Compare test scores with another measure of performance obtained at a later date (for prediction) or with another measure of performance obtained at the same time (for estimating the present status).

3. Construct related evidence

How well test performance can be interpreted as a meaningful measure of some characteristic or quality. A construct is a psychological quality that we assume exists in order to explain some aspects of the behaviour, e.g. Mathematical reasoning, anxiety, creativity. Here you need to define the construct first. This includes both content validation and criterion related evidence. Then correlate the test scores with other measures known to correlate with your construct and measures which do not correlate with the construct.

Factors Influencing Validity

- Unclear directions – directions that do not clearly indicate to the pupil how to respond to the items
- Inappropriate level of difficulty of the test items – in NRT items that are too easy or too difficult will not provide reliable discrimination among pupils and therefore lower validity.
- Poorly constructed items
- Inadequate time limits – If the time is too short then the test will be a speed test rather than the test to measure the intended domain.

- Test too short – a test is only a sample of the many questions that might be asked. Thus a test which is too short fails to provide a representative sample of performance.

Reliability

Meaning of reliability
The reliability of an assessment procedure is a measure of the consistency with which the question, test or examination produces the same results under different but comparable conditions.

A reliable test should give similar results even through different testers administer it, different people score it, different forms of the test are given, and the same person takes the test at two or more different times. It is obviously important to have reasonably reliable assessment procedures when a large number of individual markers assess the same question (e.g. in national school examinations, or with many postgraduates marking lab work). A student answer which receives a score of 75 per cent from one marker and 35 per cent from another, for example, reveals a patently unreliable assessment procedure.

To help produce reliability, the questions which comprise an assessment should (ideally) test only one thing at a time and give the candidates no choice. The assessment should also adequately reflect the desired outcomes of the teaching unit. Note that the reliability and validity factors in an assessment are in no way directly linked - a test or examination, for example, may be totally reliable and yet have very low validity, and vice versa.

Note
- Reliability refers to the results obtained with an evaluation instrument and not to the instrument and not to the instrument itself – Any particular instrument may have a number of different reliabilities, depending on the group involved and the situation in which it is used.
- Reliability is a necessary but not a sufficient condition for validity. A test that is highly consistent test results may be measuring the wrong thing or may be used in inappropriate ways.
- Reliability is primarily statistical. The logical analysis of a test will provide little evidence concerning the reliability of scores; computations must be made to establish reliability.

Methods of Estimating Reliability

1. Test – retest method

- The test-retest method is essentially a measure of examiners reliability, an indication of how consistently examinees perform on the same set of tasks.
- To estimate reliability by means of the test-retest method, the same test is administered twice to the same group of pupils with a given time interval between two administration. The resulting test scores are correlated, and this correlation coefficient provides a measure of stability that is, it indicates how stable the result are over the given period of time.

2. Equivalent – forms method

- This method uses two different but equivalent forms of the test which have nearly the same level off difficulty
- The two forms of the test are administered to the same group of pupils in close succession, and the resulting test scores are correlated. This, it indicates the degree to which both forms of the test are measuring the same aspects of behaviour.

3. Split – half method

- In this method, the test is administered to a group of pupils in the usual manner and then divided in half for scoring purpose.
- The usual procedure is to score the even – numbered and the odd – numbered items separately. This produces two scores for each pupil, when correlated, provide a measure of *internal consistency*.

4. Kuder – Richardson method

- This is the method of estimating the reliability of test scores from single administration by means of formula some of these formulas are those developed by Kuder and Richardson (1937).
- These formulas provide a **measure of internal consistency** but do not require splitting the test in half for scoring purposes.

(i) K – R 20

$$r = \frac{k}{k-1}\left[1 - \frac{\sum pq}{s^2}\right]$$

where k= number of test items

p = proportion of correct responses to a particular item.

q = proportion of incorrect responses to that item so p plus q always equal to one (p+q=1).

s^2 = variance of the scores on the test.

- The formula is applicable only to tests scored dichotomously (0 or 1, one point for each correct answer and no points for an in correct answer.

(ii) K – R 21

$$r = \frac{k}{k-1}\left[1 - \frac{\overline{X}\left(k - \overline{X}\right)}{ks^2}\right]$$

where $\overline{X}$ = mean of the class (group)

- The computation of K–R20 requires information about the difficulty (proportion of correct responses) of each item in the test.
- If the test items do not vary widely in difficulty, a good approximation of the quality pq can be obtained from information about the test mean and the number of items. The formula K–R21 estimates the value of K-R20 .

$$r = \frac{k}{k-1}\left[1 - \frac{\overline{X}\left(k - \overline{X}\right)}{ks^2}\right]$$

- One limitation of the K–R21 is that it always gives an underestimate of the K-R20 reliability coefficient when the items vary in difficulty.

Factors Influencing Reliability Measures

Consideration of the factors influencing reliability not only will help us interpret more wisely the reliability coefficients of standardised test but also should be us in constructing more reliable classroom tests.

1. **Length of test -** The longer the test is, the higher its reliability will be. This is because longer test will provide a more adequate sample of behaviour being measured, and the scores apt to be less distorted by chance factors such as guessing.
2. **Spread of Scores:** Other things being equal, the larger the spread of scores is the higher the estimate of reliability coefficient result when individual and to stay in the same relative position in a group from one testing to another, it naturally follows that anything that reduces the possibility of shifting positions in the group also contributes to larger reliability coefficients.
3. **Difficulty of test:** Test that are too easy or too difficult for the group members taking it will tend to produce scores of low reliability. This is because both easy and difficult test result in a restricted spread of scores.
4. **Objectivity:** The objectivity of a test refers to the degree to which equally competent scores obtain the same results. Most standardised

tests of aptitude are high in objectivity. The test items are of the objective type and resulting scores are not influenced by scores judgement or opinion.

Practicability

For most purposes, assessment procedures should be realistically **practical** in terms of their cost, time taken, and ease of application. For example, with a large class, it may be convenient to use only a paper-and-pencil test rather than set up numerous practical testing situations. Note, however, that such compromises can reduce the validity of the assessment.

Fairness

To be **fair** to all students, an assessment must accurately reflect the range of expected behaviours as described by the published syllabus. It is also highly desirable that students should know exactly **how** they are to be assessed. Indeed, it is now advised that students have a **right** to information such as the nature of the materials on which they are to be tested (i.e. content and outcomes), the form and structure of the test or examination, the length of the examination, and the value (in terms of marks) of each component of the course.

Usefulness to students

Students should also find assessment **useful** , in that it contributes to the effectiveness of their learning. It should do this in two different ways, namely, by getting them to carry out tasks that facilitate learning (revising material covered in the lessons) and in providing them with feedback on how they are progressing, thus helping them to identify their strengths and weaknesses.

Learning Unit 2.2 Classification of tests

For the purposes of measurement, tests fall into two general categories: selection-type - those in which students select the correct response and supply type - those in which students must formulate their own answers. The cognitive capabilities required to answer selection items are different from those required by supply items, regardless of content.

In principle, both selection and supply items can be used to test a wide range of learning objectives. In practice, most people find it easier to construct selection items to test recall and comprehension and supply items to test higher-level learning objectives. Selection items that require students to do such things as classify statements as fact or opinion go beyond rote learning, and focused essay questions can easily stay at the recall level.

Selection type of items include:

- Multiple choice items
- True false items
- Matching items

Supply Types of items require students to supply answers. These include

- Short answer items
- Completion items
- Essay items

Key characteristics of test items

Multiple Choice Items

Multiple-choice items offer the most versatility of all item types. Their uses include testing for factual recall as well as measuring level of understanding and the application of concepts. Multiple-choice items may be used to test all levels of thinking, although higher level items are more difficult to compose. As a result, the ease of constructing items that only require lower levels of thinking often leads to tests that address only these levels.

Construction of Multiple Choice Items

Writing the Stem

The term *stem* refers to the part of the item that asks the question. The terms *responses* , *choices* , *options* , and *alternatives* refer to the parts of the item that are used to answer the question.

Rules of Stem Writing

- The stem of the item should be written first.
- Write the stem as a single, clearly-stated question or problem. Stems without verbs fail to present problems clearly.
- As much of the question as possible should be included in the stem. Students should be able to understand the question without reading it several times or having to read all the options.
- Don't obscure the question by adding irrelevant information to the stem. In higher-level questions the stem will normally be longer than in lower-level questions, but you should still strive for brevity.
- State the question in positive form whenever possible, as students often misread negative-form questions. If you must compose a negative stem, emphasize the negative words with formatting such as underlining or all capital letters. Beware of using double negatives, such as "Which of these is not the least important element of aviation safety programs?"

Rules of Response Writing

- Students should be able to select the right response without having to sort out complexities not relating to the correct answer.
- Students should not be able to guess the correct answer from the way the responses are written.
- Write the correct response immediately after writing the stem and make sure it is unquestionably correct.
- Write the incorrect responses to parallel the correct response in length, complexity, phrasing, and style. Increase the believability of the incorrect responses by including superfluous information and by basing them on logical fallacies.
- Randomise the placement of the correct response. Be sure that the correct responses do not follow a pattern.
- Avoid clues that give away the correct answer, such as grammatical/syntactical mismatches between the stem and options; key words that appear only in the stem and correct response; stating correct options in textbook language and incorrect options in everyday language; using absolute terms in incorrect responses; using two incorrect options with the same meaning; and providing clues from other questions.
- Avoid using "all of the above" or "both A and B" as responses, since these options make it possible for students to guess the correct answer with only partial knowledge.
- All of the options must be plausible; humorous give-away options defeat the purpose of having multiple options.

True – False or Alternative Response Items

- The alternative – response test item consists of a declarative statement that the pupil is asked to mark true or false, right or wrong, correct or incorrect, yes or no, fact or opinion, agree or disagreed, etc. The true-false items are used in measuring;
- The ability to identify the correctness of statements or definitions of terms;
- The ability to distinguish fact from opinion;
- The ability to recognize cause and effect relationship;

Advantages

- True False Items are easy to prepare.
- They permit covering more content area than most other item types.
- They are easy to score accurately and quickly.

Limitations

- True/false items may not give a true estimate of the students' knowledge since half can be correct by chance.
- They are extremely poor for diagnosing student strengths and weaknesses.
- They are generally considered to be "tricky" by students.
- They tend to be either very easy or very difficult.
- They do not distinguish between varying degrees of learning well

Effective Practices in Constructing True/False Items

- Keep the language simple and clear. Avoid ambiguous and trick items.
- Use a relatively large number of items (75 or more when the entire test is T/F).
- Be aware that extremely long or complicated statements will test reading skill more than content knowledge.
- Avoid the use of negatives, especially double negatives.
- Make sure that the statements used are entirely true or entirely false.
- Use certain key words sparingly since they may provide clues to the correct answers. The words *all, always, never, every, none,* and *only* usually indicate a false statement, while the words *generally, sometimes, usually, maybe* , and *often* are often used in true items.
- Use precise terms, such as 50% of the time, rather than less precise terms, such as *several, seldom* , and *frequently* .
- Use more false than true items, but not more than 15% more. (False items tend to discriminate more than true items.)

Matching Items

Matching items are a type of multiple-choice question, and the same principles apply to writing them. It is extremely difficult to write matching items that test higher-order learning. The connections that students make between two concepts may reflect only a barely understood association rather than a full appreciation of the relationship.

Advantages

Matching items have the following advantages. They:

- are generally quite brief and uninvolved.
- are especially suitable for who, what, when, and where questions.
- can be used to have students discriminate among and apply concepts.
- permit efficient use of space when there are a number of similar types of information to be tested.
- are easy to score accurately and quickly.

Disadvantages

Matching Items are:

- difficult to use to measure learning beyond recognition of basic factual knowledge.
- usually poor for diagnosing student strengths and weaknesses.
- appropriate in only a limited number of situations.
- difficult to construct since parallel information is required

Suggestions for Constructing Matching Items

- Use only homogeneous material in a set of matching items (i.e, dates and places should not be in the same set) to reduce the possibility of guessing the correct answers.
- Place the more involved expressions in the stem and keep the responses short and simple.
- Supply directions that clearly state the basis for the matching, indicating whether or not a response can be used more than once, and stating where the answer should be placed.
- Make sure that there are never multiple correct responses for one stem (although a response may be used as the correct answer for more than one stem).
- Avoid giving grammatical clues to the correct response.
- Arrange items in the response column in some logical order—alphabetical, numerical, chronological—so that students can find them easily.
- Avoid breaking a set of items (stems and responses) over two pages.
- Use no more than 15 items in one set.
- Provide more responses than stems to make process-of-elimination guessing less effective.
- Use capital letters for the response signs rather than lower-case letters.

Completion Items

On the whole, they offer little advantage over other item types unless the need for specific recall is essential.

Advantages

- Useful in assessing mastery of factual information when a specific word or phrase is important to know.

Disadvantages

- Cannot test higher-order learning.
- Tend to test only rote, repetitive responses.

Suggestions for Constructing Completion Items

- Providing clear and concise cues about the expected response in the statement.
- When possible, providing explicit directions as to what amount of variation will be accepted in the answers.
- Avoiding using a long quote with multiple blanks to complete.
- Requiring only one word or phrase in each blank.
- Facilitating scoring by having the students write their responses on lines arranged in a column to the left of the items.
- Asking students to fill in only important terms or expressions.
- Avoiding providing grammatical clues to the correct answer by using a/an, etc., instead of specific modifiers.

Short Answer Items

These are suitable for measuring a wide variety of relatively simple learning outcomes: They can be used for measuring knowledge of terminology, knowledge of specific facts, knowledge of principles and knowledge of procedure or method and Simple interpretations of data.

Advantages of Short Answer Items

- Easy to construct partly because of the relatively simple learning outcomes they usually measure.
- The requirement for pupils to supply the answer reduces the chances of guessing.

Limitations

They have got two major limitations

- Unsuitable for measuring complex learning outcomes
- The difficulty of scoring esp. when the question is not carefully phrased

Suggestions for Constructing Short- Answer Items

- Word the item so that the required answer is both brief and specific.
- Do not take statement directly from textbooks to use as a basis for short answer items.
- A direct question is generally more desirable than an in complete statement.
- If the answer is to be expressed in numerical its, indicate the type f answer wanted.
- Blanks for answers should be equal in length and in column to the right of the question.
- When completion are used, do not include too may blanks.

Essay Items

The distinctive feature of essay question is the freedom of response. Pupils are free to select, relate and present ideas in their own words. Although this freedom enhances the value of essay questions as a measure of complex achievement, it introduces scoring difficulties that make them inefficient as a measure of factual knowledge. Essay items are primarily used to measure these learning outcomes that cannot be measured by objective test items. Many instructors consider essay questions the ideal form of testing, since essays seem to require more effort from the student than other types of questions.

Essay responses allow us to see our students' thought processes that lead to the answers. We may be testing at some higher level of Bloom's taxonomy of thinking—perhaps within the level of synthesis—but discover in a student's answer that he/she lacked the knowledge required to begin synthesis.

Advantages of Essay Questions

- The essay question measures complex learning outcomes that cannot be measured by other means.
- The extended response question emphasised on the integration and application of thinking and problem solving skills.
- Because the pupils must present their answers in their own handwriting the essay test is often regarded as a device for improving wring skills.
- Ease of construction

Limitations

- The unreliability of scoring
- The amount of time required for scoring the answers.
- They provide limited sampling of content area.

Suggestions for Increasing the Reliability of Essay Grading

- Read a few papers before you actually start grading in order to get an idea of the range of quality.
- Some instructors select "range finder" papers—middle range A, B, C and D papers to which they refer for comparison.
- Stop grading when you get too tired or bored. When you start again, read over the last couple of papers you graded to make sure you were fair.
- Conceal the student's name while you grade the response. If you know the identity of the student, your overall impressions of that student's work will inevitably influence the scoring of the test.

- If there is more than one essay question on the test, grade each essay separately rather than grading a student's entire test at once. Otherwise, a brilliant performance on the first question may overshadow weaker answers in other questions (or vice-versa).
- Remain open to legitimate interpretations of the questions different from your own. If students misinterpret the intent of your question, or if your standards are unrealistically high or low, you should alter your model response in light of this information.

Suggestions for Constructing Essay Questions

- Restrict the use of essay question to those learning outcomes that cannot be satisfactorily measured by objective items
- Formulate question that will call forth the behaviour specified in the learning outcomes.
- Phrase each questions so that the pupils tasks is clearly indicated
- Indicate an appropriate time limit for each question
- Avoid the use of optional questions.

General Principles of Test Construction

- Construct a test so as it contributes to improved teaching-learning practices. Tests can influence course objectives, methods and/or prerequisite entering behaviours.
- Base test questions on a representative sample of course content and specific learning outcomes to be measured.
- Make test questions at an appropriate level of difficulty.
- When constructing test questions, keep ambiguity low and reduce answering errors that might be attributable to using complex sentence structure or difficult vocabulary levels.
- Construct test questions so students obtain the correct answer only if the desired learning outcome is attained.

Practical Approaches to Test Construction

- **Note possible test questions throughout the term** ; writing a few after each class is ideal.
- **Avoid trick questions** ; if the majority of students get a question wrong, it was likely a poor question.
- **Keep questions as brief as possible** to eliminate the need for speed reading and writing.
- **Use a variety of question types** .
- **Group similar type questions** so students don't constantly shift response patterns.
- If using a **series of questions** in which answering successfully depends on knowing the correct answer to a previous item **, grading should take early errors into account** and students should be informed of this.

- **Arrange items in order of difficulty** to avoid discouraging students at the exam beginning.
- **Weight points according to question type**, amount of learning assessed and time students should spend answering.
- **Avoid patterns in the response key**.
- Strive to **include items** which **test the higher levels of thinking and learning**.
- **Minimize such qualifying words as 'always' and 'never'**.
- **Use the positive statement** whenever possible.
- **Don't give too many clues** to answers in preceding or subsequent items.