Testing and Evaluation

(ENG520)

Table of Contents

Lesson 01

CONCEPT OF ASSESSMENT-I

Topic- 001: Measurement, Assessment and Evaluation

Measurement

Measurement is the process by which the attributes or dimensions of some object (both physical and abstract) are quantified. The tools used for this purpose may include test, observation, checklist, homework, portfolios, project etc. Measurement can be easily understood if we use this word to measure height and distance because these things are physical present in the existence so height can easily be measured by scale. But in the field of education, our variable are not physical and cannot be directly measured e.g. attitude, behavior, and achievement etc. these all are abstract, so there measurement is relatively difficult than those who have physical existence. The tool used for measuring the abstract variables cannot measure exactly like scale (thermometer). So, in this whole course, whenever the measurement word is used it means that tool which will be used for measuring student abilities and then converts it in numerical form.

Assessment

It means an appraisal of something to improve the quality of teaching and learning process for deciding what more can be done to improve the teaching, learning and outcomes.

Evaluation

Evaluation is the process of making a value judgment against intended learning outcomes and behavior, to decide the quality and extent of learning. Evaluation is always related to your purpose, you aligned your purpose of teaching with what students achieved at the end, with their quality and quantity of learning.

Example:

In a classroom situation, as a teacher, when you teach a chapter or unit to a class, first you made the objectives, either you will do it yourself as a teacher or you take it from a curriculum document of the particular subject. Objectives are also written at the start of the chapter or book which shows that at the end of the unit what the student will be able to do this, that is also referred as student learning outcomes. These SLOs can be checked by two aspects:

1. Assessment
2. Evaluation.

When you, as a teacher reflect on your teaching on a daily basis that what you teach yesterday was a good way to teach or not, what i taught was the students need, did they understand what I teach, and then you can decide changing which things can cause the improvement in the learning process of students. This is known as ASSESSMENT. At the end of the assessment, opinion is not about the individual, it is about the process from which the students or individuals are passing for the betterment of that process. Evaluation is that when the learning process is complete and you want to see what are your targets or objectives and how much my students achieved those objectives, then the tools are made and measures come from that tell that how much your student learn and what is the quality of their learning.

Difference in Measurement, Assessment and Evaluation

These terms are not different words for same concept, but a process, serving as prerequisite to each other and having unique purpose. Every mechanism or process starts with the measurement. In the field of education, measurements generate tools, i.e. test, observation, quiz, checklist, homework and portfolios. They can give information about students' learning. The information we got from the tool is MEASUREMENT. Now if that measurement is used for the process of making the teaching learning process better than it is ASSESSMENT.

EVALUATION

It is a systematic determination of a subject's merit, worth and significance, using criteria governed by a set of standards

Assessment and Evaluation don't exist in a hierarchy; they both are parallel and different in purpose. Measurement is the source to move towards assessment and evaluation because it provides the base and evidence to quantify the teaching learning process. The quantified number has no meaning until we do an assessment or evaluation. Assessment purpose is to make the teaching learning process better so that student learning improves and measurement purpose is to align the learning with a purpose.

Topic – 002: Classroom Assessment

The process of gathering, recording, interpreting, using and communicating information about a child's progress and achievement during the development of knowledge, concepts, skills and attitudes is called classroom assessment.

When we are teaching in a classroom, we are doing four things:

1. Developing students 'knowledge
2. Improving their concepts
3. Teaching the skills
4. Making their attitudes

And while doing so, we collect information about the development of these things, and then we record, interpret and use it to communicate about the learning progress of students. This whole procedure is called classroom assessment

Classification of Assessment

Assessment can be classified in four ways

1. Nature of Assessment
2. Format of Assessment
3. Use in Classroom Instruction
4. Method of Interpreting Results

1. Nature of Assessment

i. Maximum Performance Assessment
ii. Typical Performance Assessment

2. Format of Assessment

i. Fixed Choice Assessment
ii. Complex Performance Assessment

3. Use in Classroom Instruction

i. Placement Assessment
ii. Formative Assessment

iii. Diagnostic Assessment
iv. Summative Assessment

4. Method of Interpreting Results

i. Norm Referenced Assessment
ii. Criterion Referenced Assessment

Topic – 003: Types of Assessment

1. Types of assessment by nature
2. Types of assessment by format

1. Types of Assessment by Nature

i. Maximum Performance Assessment
ii. Typical Performance Assessment

Maximum Performance Assessment

Maximum performance assessment determines what individuals can do when performing at their best e.g. assess student in an environment when they exhibit their best performance. Procedure of this type is concerned with how well individual perform when they are motivated to obtain as high score as possible. This type of assessment includes aptitude tests and achievement tests. In achievement test, the student learn by themselves or we teach them and at the end, we want to see that how much student learn against our target so in this situation we make a test from which we can determine their best abilities. It is designed to indicate the degree of success in any past learning activity.

We use aptitude test when we want to predict the success in future learning activity, e.g. it is used when we want to see the interest of students in a particular field like medicine, sport and teaching. We know that different abilities used for going in different professions, we make a test depending on these abilities and then try to assess, in what abilities the students perform well.

Typical Performance Assessment

The second category is a typical performance assessment determines what an individual will do under natural conditions. This type of assessment includes attitude, interest, personality inventories, observational techniques and peer appraisal. Here the emphasis is on what students will do rather than what they can do.

2. Types of Assessment by Format

i. Fixed Choice Assessment
ii. Complex Performance Assessment

Fixed Choice Assessment

Fixed choice assessment is sued to measure the skills of people efficiently (means measure more skills in less time) and for this we usually use fixed choice items, i.e. multiple choice question, matching exercises, fill in the blanks and true false. It is called fixed choice because the person who is attempting the paper does not need to write the answer, just need to choice the answer. From these, we can assess

student abilities of lower level learning. Fixed choice assessment is used for efficient measurement of knowledge and skills.

Complex Performance Assessment

Complex performance assessment is used for measurement of performance in contexts and the problems valued in their own right. This includes hands on laboratory experiments, projects, essays, oral presentations. E.g. if want to measure the student's ability of writing an essay and this cannot be judged by fixed response items

Lesson 02

CONCEPT OF ASSESSMENT-II

Topic- 001: Use of Assessment in Classroom Instruction

Classification of assessment in terms of its uses in classroom instruction

i. Placement Assessment
ii. Diagnostic Assessment

Placement Assessment

Placement assessment determines prerequisite skills, degree of mastery of course goals and mode of learning. It is used when we want to assess student's prior knowledge so that we can decide what the level of student is. It is associated with a student's entry level performance to know either student have a sufficient knowledge required for a particular course or not. Through placement assessment, teacher can be able to know that where a student should be placed according to the present knowledge or skills. It determines the level of student knowledge at the beginning of session and helps teacher plan the lesson accordingly. In the classroom, the teacher can use placement assessment to assess the level of students' knowledge and skills and then make lesson plans keeping in mind the level and need of students accordingly. It also determines the interest and aptitude of the student regarding a subject and helps in selecting the correct path for the future.

Examples:

Readiness test: It is used to determine the students' knowledge or concept about a particular course of instruction or what is the level of students.
Aptitude test: It is used for the admission in a particular program.
Pretest: It is made according to the course objectives and determines the student present knowledge about them.
Self- report inventories: It determines the student level by interviewing or discussion.

Diagnostic Assessment

It determines the causes (intellectual, physical, emotional, environmental) of persistent learning difficulties. E.g. if you are having a headache, first you will try to cure it by yourself by taking a medicine and you got a relief, but if you did not get a relief by taking the medicine then either you change your medicine or you go to the physician or doctor. At first, doctor prescribed medicines, if you still have a headache you again go to the doctor, then the doctor suggests you the tests, i.e. blood test, urine test etc. Then finally by seeing the test reports the doctor able to recognize the reason or cause of the headache. And when doctor knows the root of your headache then he will prescribe you the medicine for that cause, this is diagnosis.

A diagnosis does not start at the first day, it is for the constant or continuous problems, e.g. if a student continues to experience failure in reading or mathematics or any other subject despite the use of prescribed alternative methods , then a diagnosis is indicated. Teachers try to find out what is the root of students' failure.

Topic- 002: Use of Assessment in Classroom Instruction – Formative and Summative Assessments

Formative Assessment:

It determines learning progress, provides feedback to reinforce learning, and correct learning errors. When we assess student during classroom instruction with a purpose to have a feedback that how can we make our teacher learning process better, that is formative assessment. In this assessment, we are not assessing what students learnt or not, rather we assess the process behind the students learning .The process behind the student learning includes a teaching method, book, if we make all these things according to the needs of students then learning will improve.

It is conducted during the academic session or teaching-learning process so that I can get a feedback about my way of teaching and how students are learning and decisions are made on the basis of results immediately. It is an ongoing process to modify teaching strategies on the basis of students need.

It provides feedback to teachers:

- About weakness and strength of learning process
- To modify their teaching practices
- To improve teacher-learning process

It also helps students to reflect on their weaknesses and encourages them for their successful learning. Formative assessment provides feedback to students who are struggling with a specific content area or concept. The main difference between formative and summative assessment is that the in formative assessment, improvement is in the process of learning rather than to certify students .We use different tools for formative assessment and it includes teacher made tests, custom made tests from textbook.

Summative Assessment

It comes at the end of the instructional session (course of the unit). It is designed to measure extend of achievement of intended learning outcomes. The primary utility of this type of assessment is to assign grades and certifying the level of mastery and expertise in a certain subject. It is usually done through teacher made achievement tests or alternative assessment techniques like portfolio to summarize the overall performance of the student at the end of the session. It is not compulsory to do it at the end of the semester, in semester system, there is a midterm and final term in one semester, and these both are summative assessment. It usually compares the student learning either with other students' learning (norm-referenced) or the standard for a grade level (criterion-referenced). Summative assessment includes teacher made survey test, performance, rating scales and product scales.

Topic- 003: Types of Assessment: Methods of Interpreting Results

- Norm-referenced Assessment
- Criterion-referenced Assessment

Norm-referenced Assessment

Norm-referenced assessment measures the students' performance, according to the relative position in any known group. E.g. ranks tenth in classroom group of 50 or top 5 students in a class. Relative position means a point defined with reference to another position (where a student stands compared to other students) rather than reporting students' achievement, it reports students standing among other students. NRT is utilized to discriminate between a certain groups of students. It is never used for certification or

issuing grades to students. The position of the student is generally represented by percentile scarce noting the percentage of students achieving same or low scores in the test.
Examples of NRT are NTS or CSS exams. E.g. if you achieved a 97% percentile on NTS test, it means there are 96% who scored lower than you.

Norm referenced test includes items with average difficulty and high discriminating power. This provides a large spread of scores which makes it easy to declare the relative positions of students. Because the purpose of NRT is not to certify grades, so the test made form it must be of average difficulty means test items should not be very easy or very tough. If all the items are difficult then no student will be able to solve it and we are not able to discriminate who are good students. And if all the items are easy then even the lower ability students can solve it then we are not able to discriminate it. Norm referenced test includes standardized aptitude, achievement tests, teacher-made survey tests, interest inventories, adjustment inventories.

Criterion-referenced Assessment
Criterion-referenced assessment describes student performance, according to a specific domain of clearly defined learning tasks, e.g. adds single-digit whole numbers. With this you do not compare student's performance with other students rather you compare the performance of all students with criteria (in our case that criteria are our learning outcomes). It is most commonly used in schools to report the achievement of learning outcomes against set goals rather than other students. It grades the students to pre-defined criteria and student's grades represent their mastery over content. Students with the same level of expertise achieve the same level of grades. A cut point is determined to distinguish between failed and successful students, regardless of score of highest and lowest achiever. It consists of teacher-made tests, custom made tests from the test publishers and observational techniques.

Lesson 03

ASSESSMENT, TESTING AND NATIONAL CURRICULUM

Topic- 007: Role of National Curriculum in Assessment

In national curriculums of Pakistan, learning of student is classified into four levels.

1. Competency
2. Standards
3. Benchmarks
4. Student learning outcome (SLOs)

Competency

It is a key learning area, for example, algebra, arithmetic, geometry, etc. in mathematics and vocabulary, grammar, composition, etc. in English.

Standards

These define the competency by specifying broadly, the knowledge, skills and attitudes that students will acquire, should know and be able to do in a particular key learning area during twelve years of schooling.

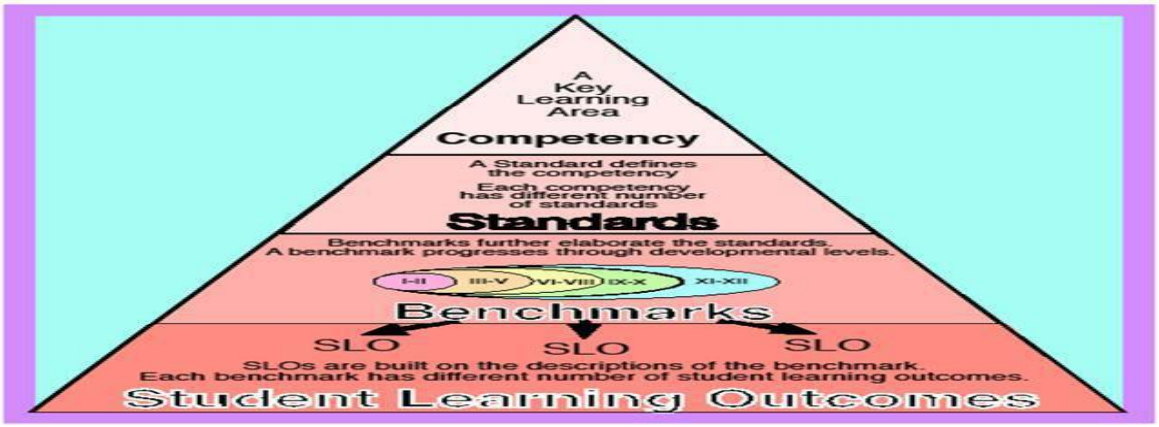
Benchmarks

The benchmarks further elaborate the standards, indicating what the students will accomplish at the end of each of the five developmental levels in order to meet the standard.

Student Learning Outcomes

These are built on the descriptions of the benchmarks and describe what students will accomplish at the end of each grade. It is the lowest level of hierarchy.

Topic- 008: Connecting all Four Levels in Curriculum



In the image above, SLOs are at the bottom which is the lowest level. All SLOs combined to make a benchmark and benchmarks convert into standards and then into competency.

Example:

Example is taken from curriculum of English subject.

Competency 1: Reading and thinking skills

Standard 1: All students will discover and understand a variety of text types through tasks which require multiple reading and thinking strategies for comprehension, fluency and enjoyment.

Benchmark 1: Use reading readiness strategies

Student learning outcome:

1. Articulate, identify and differentiate between the sounds of individual letters, digraphs and trigraphs in initial and final positions in a word.
2. Identify paragraph as a graphical unit of expansion, ¾ know that word in a sentence join to make sense in relation to each other.

Topic- 009: Modes of Assessment in Curriculum

Curriculum document provides specific guidelines for assessment.

Example:

6.1. The two forms of assessment recommended are:

6.1.1. Periodic/formative assessment through homework, quizzes, class tests and group discussions.

6.1.2. End of term/ summative assessment through final examination.

Purpose of Assessment and Curriculum-English 2006

The assessment system for the present curriculum should include:

- A clear statement of the specific purpose(s) for which the assessment is being carried out.
- A wide variety of assessment tools and techniques to measure student's ability to use language effectively.
- Criteria to be used for determining performance levels for the SLOs for each grade level.
- Procedures for interpretation and use of assessment results to evaluate the learning outcomes.

Form of Suitable Assessment Tools- English 2006

- MCQs
- Constructed response
  - Restricted response
  - Extended response
- Performance tasks

Lesson 04

TAXONOMIES OF EDUCATIONAL OBJECTIVES AND ASSESSMENT –I

Topic- 010: Taxonomies of Educational Objectives and Assessment

Every assessment regardless of its purposes rests on three important pillars:

1. A model for how students present knowledge and develop competence in the subject domain
2. Tasks or situations that allow the examiner to observe the students' performance
3. Inferences from performance evidence about the quality of learning

In developing a test to assess student learning, taxonomy provides a framework of categories with different hierarchical levels of outcomes.

Popular Taxonomies

1. Bloom's taxonomy of educational objective
2. Structure of Observed Learning Outcomes (SOLO)
3. Depth of Knowledge (DOK)

Topic- 011: Bloom's Taxonomy and SOLO Taxonomy

The taxonomy of structure of observed learning outcomes (SOLO) was initially developed by Biggs and Collis in 1982, and then well described in Biggs and Tang in 2007. It carries five different levels of competency of learners.

Levels of Structure of Observed Learning Outcomes (SOLO)

1. Pre-structural
2. Uni-structural
3. Multi-structural
4. Relational
5. Extended Abstract

Depth of Knowledge (DOK)

DOK (Depth of Knowledge) was presented by Webb in 1997, giving four levels of learning activities.

Levels of Depth of Knowledge (DOK)

1. Recall
2. Skill/Concept
3. Strategic Thinking
4. Extended Thinking

Bloom's Taxonomy of Learning Objectives

Bloom's Taxonomy was presented by Benjamin Bloom in the 1956, consists of a framework with most common objectives of classroom instructions.

1. Cognitive
2. Affective
3. Psychomotor

Cognitive Domain

- Knowledge
- Comprehension

- Application
- Analysis
- Synthesis
- Evaluation

Affective Domain

- Receiving
- Responding
- Valuing
- Organization
- Characterization

Psychomotor Domain

- Perception
- Set
- Guided Response
- Mechanism
- Complex covert Response
- Adaption
- Origination

Topic- 012: SOLO Taxonomy

Levels of SOLO

1. Pre-structural
2. Uni-Structural
3. Multi-structural
4. Relational
5. Extended Abstract

1. Pre-structural

Students are simply able to acquire bits of unconnected information and respond to a question in a meaningless way.

Example:

Question: What is your name?

Answer: What is your name?

2. Uni-Structural

Student shows concrete understanding of the topic. But at this level is only able to respond one relevant element from the stimuli or an item that is provided.

Indicative verbs: identify, memorize, do simple procedure

3. Multi- Structural

The student can understand several components, but the understanding of each remains discreet. A number of connections are made, but the significance of the whole is not determined. Ideas and concepts around an issue are disorganized and aren't related together.

Indicative verbs: enumerate, classify, describe, list, combine, do algorithms.

4. Relational

At this level, the learner is able to understand the significance of the parts in relation to the whole. Ideas and concepts are linked, and they provide a coherent understanding of the whole.

Indicative verbs: compare/contrast, explain causes, integrate, analyze, relate, and apply.

5. Extended Abstract

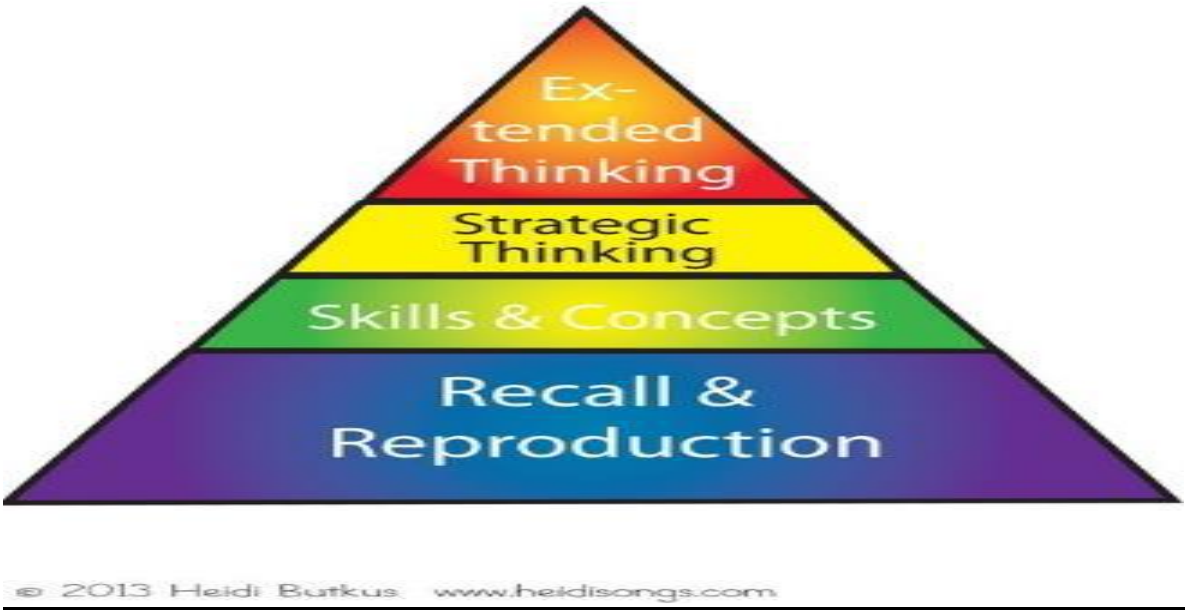
At this level, the learner is able to think hypothetically and can synthesize a material logically. Student make connections not only within the given subject area, but also understanding is transferable and generalizable to different areas.

Indicative verbs: theorize, generalize, hypothesize, reflect, generate.

Topic- 013: Depth of Knowledge

Levels of DOK

1. Recall
2. Skill/concept
3. Strategic Thinking
4. Extended Thinking



DOK measures the degree to which the knowledge brings about from students on assessments is as complex as what students are expected to know and do as stated in the curriculum.

Levels of DOK (Depth of Knowledge)

Recall

Recall of a fact, information, or procedure. The subject matter at this particular level usually involves working with facts, terms and/or properties of objects.

Key words: list, enlist, name, define etc.

Skill/Concept

It includes the engagement of some mental processing beyond recalling or reproducing a response. Use information or conceptual knowledge, two or more steps, not just recalling

Key words: Graph, separate, relate, contrast, narrate, compare etc.

Strategic Thinking

Items falling in this category demand a short-term use of higher order thinking processes, such as analysis and evaluation, to solve real-world problems with predictable outcomes.
Key words: Argue, critique, formulate

Extended Thinking

Learning outcomes to this level demand extended use of higher order thinking processes such as synthesis, reflection, assessment and adjustment of plans over time.
Key words: create synthesize, design and reflection.

Lesson 05

TAXONOMIES OF EDUCATIONAL OBJECTIVES AND ASSESSMENT –II

Topic- 014: Bloom‘s Taxonomy-I

There are three main domains of learning and all teachers should know and use them to construct lessons.

1. Cognitive Domain
2. Affective Domain
3. Psychomotor Domain

1. Cognitive or Thinking Domain

In 2000-01 revisions to the cognitive taxonomy were spearheaded by one of Bloom's former students, Lorin Anderson, and Bloom's original partner in defining and publishing the cognitive domain, David Krathwohl. One of the major changes that occurred between the old and the newer updated version is that the two highest forms of cognition have been reversed.

Levels of Cognitive Domain

Old Cognitive Domain

Knowledge

It is defined as the remembering of previously learned material. This may involve the recall of a wide range of facts, procedures, principals and generals, the recall of the procedures and the processes. Sample Question: Define the 6 levels of Bloom's taxonomy of the cognitive domain.

Comprehension

It is defined as the ability to grasp the meaning of the material. An individual can make use of the content or idea being communicated without necessarily related it to other content and seeing its fullest implications. Sample Question: explain the purpose of Bloom's taxonomy of the cognitive domain.

Application

It refers to the ability to use the previously learned material in new and concrete situations. The abstractions may be in the shape of universal ideas, rules of methods. Sample Question: write an instructional objective for each level of Bloom's taxonomy.

Topic- 015: Bloom‘s Taxonomy-II

Analysis

The breakdown of a concept into its constituent parts such that the relative hierarchy of the concept is made easy to understand or the relation between the parts of the concept is elaborated. Sample Question: compare and contrast the cognitive and affective domains.

Synthesis

There is a collection of the constituents or parts of a concept so as to make a whole. This is a stage when an individual is working with the pieces and grouping them in such a way as to formulate a pattern or structure not clearly there before. Sample Question: Design a classification scheme for writing educational objectives that combines the cognitive, affective, and psychomotor domains.

Evaluation

It is concerned with the ability to judge the value of the material for a given purpose. Judgments are made on the definite criteria. Sample Question: How far the different BISEs and universities are developing papers using Bloom's taxonomy? Support your answer with arguments.

Topic- 016: Revised version of Bloom‘s Taxonomy

The revised cognitive domains levels are:

Remembering

Exhibit memory of previously learned material by recalling facts, terms, basic concepts, and answer

Key verbs

Choose, define, find, how, label, list, match, name, omit, recall, relate, select, show, spell, tell, what, when, where, which, who, why

Understanding

Constructing meaning from different types of functions be they written or graphic messages, or activities.

Key verbs: Classify, compare, contrast, demonstrate, explain, extend, illustrate, infer, interpret, outline, relate, rephrase, show, summarize, and translate

Applying

Solve problems to new situations by applying acquired knowledge, facts, techniques and rules in a different way.

Key verbs:, Apply, build, choose, construct, develop, experiment with, identify, interview, make use of, model, organize, plan, select, solve and utilize.

Analyzing

Breaking materials or concepts into parts, determining how the parts relate to one another, or how the parts relate to an overall structure or purpose.

Key verbs: Analyze, assume, categorize, classify, compare, conclusion, contrast, discover, dissect, distinguish, divide, examine, function, inference and inspect

Evaluating

Making judgments based on criteria and standards through checking and critiquing.

Key verbs: Agree, appraise, assess, award, choose, compare, conclude, criteria, criticize, decide, deduct, defend, determine, disprove and estimate

Creating

Putting elements together to form a coherent or functional whole; reorganizing elements into a new pattern or structure through generating, planning, or producing.

Key verbs: Adapt, build, change, choose, combine, compile, compose, construct, create, delete, design, develop, discuss, elaborate, estimate and formulate

These categories range from simple to complex and from concrete to the abstract level of student‘s learning. It is assumed that the taxonomy represents a cumulative hierarchy, so that mastery of each simpler category is considered as a prerequisite to mastery of the next, more complex one.

Comparison of Bloom, SOLO and DOK

A comparison of hierarchical division in three taxonomies of learning objectives

SOLO Taxonomy	Bloom's Taxonomy	DOK Taxonomy
Level 1: Pre-structural Level 2: Uni-structural	Level 1 Knowledge	Level 1:REcall
Level 3:Multi-structural	Level 2:Comprehension	Level 2: Skill/Concept
Level 4: Relational	Level 3: Application	Level 3 Strategic Thinking
Level 5: Extended Response	Level 4: Analysis Level 5: Synthesis Level 6: Evaluation	Level 4: Extended Thinking

VU

Topic- 017: Instructional Objectives

Instructional Objectives as Learning Outcome

Instructional goals and objectives are stated in terms of actions to be taken. When viewing instructional objectives in terms of learning outcomes, we are concerned with products rather than process of learning.

Sources for Lists of Objectives

- Professional Associations standards
- State Content Standards
- Methods Books
- Year Books
- Encyclopedia of Educational Research
- Curriculum Frameworks
- Test Manuals

Criteria of Selecting the Final List of Objectives

Prepare tentative list of instructionally relevant learning outcomes

Review the list for:

- Completeness
- Appropriateness
- Soundness
- Feasibility

General Objectives

- Stating the general objectives is selecting the proper level of generality.
- Objective should be specific enough to provide the direction for instruction but not so specific that instruction is reduced to training
- Stating general objectives in general terms, we provide for the integration of specific facts and skills into complex response
- General statements gives teachers freedom in selecting the method and materials of instruction

List of general objectives shows the desired level of generality

- Knows basic terminology
- Understand concepts
- Relates concepts to everyday observations
- Applies principles to new situations
- Interpret graphs
- Demonstrate scientific attitude

Specific Learning Outcomes

Each General objective must be defined by a sample of specific learning outcome to clarify how students can demonstrate that they have achieved general objective. Until the general objective are further defined in this manner they will not provide adequate direction for assessment

Steps for Stating Specific Outcomes

- List below each general objective a representative sample of specific learning outcome that describes terminal performance students are expected to demonstrate.
- Begin each specific learning outcome with an action verb that specifies observable performance.
- Make sure that each specific learning outcome is relevant to the general objective it describes.
- Include enough SLOs to describe adequately the performances of students who have attained the objectives.
- Keep the SLOs sufficiently free of course content so that the list can be used with various units of study.
- Consult reference materials for the specific components of those complex outcomes that are difficult to define.
- Add a third level of specificity to the list of outcomes, if needed.

Lesson 06

PURPOSE OF TESTING-I

Topic- 018: Educational Decisions Making

Types of Educational Decisions

- Instructional
- Grading
- Diagnostic
- Selecting
- Placement
- Counseling and Guidance
- Program or Curriculum
- Administrative

These types of decisions are taken at different levels. Some are decided at the board/ administrative level while some are taken at school management level and other are taken in classrooms by teachers.

Instructional Decisions

Instructional decisions are the nuts and bolts types of decisions made in the classroom by teachersSuch decisions include deciding to:

- spend more time on specific units
- regroup student in class for better management
- instructional plans

Grading Decisions

Educational decisions based on grades are also made by the classroom teacher but much less frequently than instructional decisions. For the most students grading decisions are the most influential decision made about them.

Diagnostic Decisions

Diagnostic decisions are those made about a student‘s strengths and weaknesses and the reasons behind them. Teachers make diagnostic decisions based on information yielded by an informal teacher made test decisions about diagnostic nature can also be made with the help of standardized tests.

Selection Decisions

Selection decisions involve test data used in part for accepting or rejecting applicants for admission into a group, program, or institution.

Placement Decisions

Placement decisions are made after an individual has been accepted in a program. They determine where in a program someone is the best suited to begin with.

Program or curriculum decision

This type of decision is taken at policy level. Where it is decided if a lesion, unit or subject will continue or abandoned for the next academic session according to the national objectives of education.

Administrative Decisions

Administrative policy decisions may be made at school, district, state or national level. Based on measurement data, it includes financial decisions of schools.

Topic- 019: Types of Test

In classroom assessment different forms of assessments are utilized. Each form of test has its own benefits and disadvantages. The most common type of assessment used in classrooms is written assessment.

Types of Written Tests

- Verbal
- Non-verbal
- Objective
- Subjective
- Teacher Made
- Standardized
- Power
- Speed

Verbal

Emphasize reading, writing, or speaking. The most tests in education are verbal tests.

Non-verbal

It does not require reading, writing or speaking ability. The tests composed of numerals or drawings are an example of non-verbal test.

Objective

It refers to scoring of tests when two or more scorers can easily agree on whether the answer is correct or incorrect; the test is an objective one. True/ false, multiple choice and matching tests are examples of it

Subjective

When it is difficult for two scorers to agree on whether an item is correct or incorrect, the test is a subjective one. Essay tests are the example of it.

Teacher Made

It is constructed solely by teacher only to be used in the classroom. This type of test is custom designed according to need and issues related to specific class.

Standardized

Test constructed by measurement experts over a period of years. They are designed to measure broad national objectives and have a uniform set of instructions that are adhered to during each administration. Mostly it has tables of norms, to which a student performance may be compared to determine where the student stands in relation to a national sample of students at the same level of age or grade

Power

Tests with liberal time limits that allow each student to attempt each item. Items tend to be difficult

Speed

Tests with time limits so strict that no one is expected to complete all items. Items tend to be easy.

Topic- 020: Norm Referenced Assessment (NRT)

General purpose of assessment is to gather information to make better and more informed decision. The utility of that information is what differentiate among thetypes of assessments.

Norm Referenced Assessment (NRT)

Type of test which tells us where a student stands compared to other students. It helps to determine a student‘s place or rank among a group of similar students. Such kind of test is called norm-referenced test.

Dimensions

- It provides estimate of ability in a variety of skills in much shorter time.
- NRT tends to be general. It measures variety of skills at same time but fails to measure them thoroughly.
- It is hard to make decisions regarding the mastery of student‘s skill in subject.
- It provides estimate of ability in a variety of skills in much shorter time. NRT is much difficult for students to solve. On average only 50% students are able to get an item right in a test.

Topic- 021: Criterion Referenced Assessment (CRT)

This type of test tells us about student’s level of proficiency in or mastery of some skill or set of skills. This is achieved by comparing a student‘s performance to a standard mastery called a criterion. Test that yields such information is called criterion referenced test.

Dimensions

- CRT tends to be specific. It measures particular a set of skill at one time and focus on level of achievement of that skill. CRT gives clear picture regarding the mastery of student‘s skill in subject.
- It measures skill more thoroughly, so naturally it takes more time comparing to NRT in measuring the mastery of said skill.
- Items included in CRT are relatively easier. Around 80% of the students are expected to respond item correctly in the test.
- CRT compares students ‘performance to the standards indicative of mastery.
- Breadth of content sampled is narrow and covers very few objectives.

Lesson 07

PURPOSE OF TESTING –II

Topic- 022: Characteristics of Criterion Referenced Assessment

Sampled content in CRT is much more comprehensive, usually three or more items are used to cover single objective.

- The meaning of the score does not depend upon on comparison with other scores.
- It flows directly from the connection between the items and the criterion.
- Items are chosen to reflect the criterion behavior. Emphasis is placed upon the domain of relevant responses.
- Number succeeding or failing or range of acceptable performance used.

Example: 90% proficiency achieved, or 80% class reached 90% proficiency.

Topic- 023: Difference between NRT and CRT

Basis of Comparison

- Comparison targets
- Selection of items
- Meaning of success
- Average item difficulty
- Score distribution
- Reported scores

Comparison Targets

In CRT, the examinee‘s performance is compared to an external standard of competence. While in NRT, examinee‘s performance is typically compared to that of other examinees.

Selection of Items

Items included in CRT are of specific nature and designed for the student skilled in particular subjects. In NRT items are of general knowledge nature. Student should be able to answer it, but superficial knowledge is sufficient to respond the item correctly.

Meaning of Success

In CRT, an examinee is classified as a master or non-master. There is no limit to the number of pass or fail. In NRT, examinee‘s opportunity for success is relative to the performance of the other individuals who take the test.

Average Item Difficulty

In CRT, the average item difficulty is fairly high. Examinees are expected to show mastery. In NRT, the average item difficulty is lower. Tests are able to spread out the examinees ‘and provide a reliable ranking.

Score Distributions

In CRT, a plot of the resulting score distribution will show most of the scores clustering near the high end of the score scale. In NRT, broader spread of scores is expected, with a few examinees earning very low or high scores and many earning medium scores.

Reported Scores

In CRT, classification of the examinee is measured as master/non-master or pass/fail. In NRT, percentile ranks or scale scores are frequently used.

Topic- 024: Formative Assessment

Formative assessment provides feedback and information during the instructional process, while learning is taking place and occurring. Formative assessment measures student progress, but it can also assess your own progress as an instructor.

Types of Formative Assessment

• Observations during in-class activities; of student‘s non-verbal feedback during lecture.
• Homework exercises as review for exams and class discussions
• Reflections journals that are reviewed periodically during the semester
• Question and answer sessions, both formal (planned) and informal (spontaneous)
• Conferences between the instructor and student at various points in the semester
• In-class activities where students informally present their results
• Student feedback collected by periodically answering specific question about the instruction and their self-evaluation of performance and progress

Lesson 08

PURPOSE OF TESTING-III

Topic- 025: Functions of Formative Assessment

Functions of Formative Assessment

- Focus of measurement in formative assessment is predefined segment of instruction.
- Limited sample of learning tasks are addressed.
- The difficulty of item varies with each segment of instruction.
- Formative assessment is conducted periodically during the instructional process.
- Results of formative assessment are used to improve and direct learning through ongoing feedback.

Topic- 026: Summative Assessment

Summative assessment takes place after the learning has been completed and provides information and feedback that sums up the teaching and learning process. Typically, no more formal learning is taking place at this stage, other than incidental learning which might take place through the completion of projects and assignments. Summative assessment is more product-oriented and assesses the final product, whereas formative assessment focuses on the process toward completing the product. Once the project is completed, no further revisions can be made. If, students are allowed to make revisions, the assessment becomes formative.

Types of Summative Assessment

• Examinations (major, high-stakes exams)
• Final examinations (a truly summative assessment)
• Term papers (drafts submitted during the semester would be a formative assessment)
• Projects (project phases submitted at various completion points could be formatively assessed)
• Portfolios (could also be assessed during its development as a formative assessment)
• Performances

Topic- 027: Functions of Summative Assessment

- Focus of measurement in summative assessment is on course or unit objectives.
- Broad sample of all objectives is used in summative assessment.
- This type of assessment uses wide range of difficulty while selecting items for the test.
- Summative assessment is done at the end of the unit or the course.
- Most important functions of summative assessment is to assign grade, certification of accomplishment and evaluation of teaching.

Lesson 09

TABLE OF SPECIFICATION

Topic- 028: Table of Specification

One of the tools used by teachers to develop a blueprint for the test is called ―table of specification; in the other word table of specification is a technical name for the blue print of the test. It is the first formal step to develop a test.

Concept of Table of Specification

- It helps a teacher in allotting the questions to different content areas and Bloom‘s learning categories in a systematic manner.
- The blueprint is meant to ensure content validity. Content validity is the most important factor in constructing an achievement test. (will be discussed in later unit)
- A unit test or comprehensive exam is based on several lessons and/or chapters in a book supposedly reflecting a balance between content areas and learning levels (objectives).

Two Ways of Table of Specification

A table of specifications consists of a two-way chart or grid relating instructional objectives to the instructional content.

Table of specification performs two important functions

1. It ensures the balance and proper emphasis across all content areas covered by teacher.
2. It ensures the inclusion of items at each level of the cognitive domain of Bloom's Taxonomy.

Learning objectives	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation	Weightage %age
1							
2							
3							
4							
5							

Topic- 029: Concept of Table of Specification

It helps a teacher in allotting the questions to different content areas and Bloom‘s learning categories in a systematic manner.

Learning objectives	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation	Weightage %age
1							
2							
3							
4							
5							

Topic- 030: Elements and Appropriateness in Table of Specification

Carey (1988) listed six major elements that should be attended to in developing a table of specifications for a comprehensive end of unit exam:

i. Balance among the goals selected for the exam (weighing objectives)
ii. Balance among the levels of learning (higher order and lower order)
iii. The test format (objective and subjective)
iv. The total number of items
v. The number of test items for each goal and level of learning
vi. The enabling skills to be selected from each goal framework.

A table of specifications incorporating these six elements will result in a "comprehensive posttest that represents each unit and is balanced by goals and levels of learning"

Checklist for Appropriateness of Table of Specification

- Are the specifications in harmony with the purpose of the test?
- Do the specifications indicate the nature and limits of the achievement domain?
- Do the specifications indicate the types of learning outcomes to be measured?
- Do the specifications indicate the sample of learning outcomes to be measured?
- Is the number of test items indicated for the total test and for each subdivision?
- Are the types of items to be used appropriate for the outcomes to be measured?
- Is the difficulty of the items appropriate for the types of interpretation to be made?
- Is the distribution of items adequate for the types of interpretation to be made?
- If sample items are included, do they illustrate the desired attributes?
- Do the specifications, as a whole, indicate a representative sample of instructionally relevant tasks that fits the used to be made of the results?

Topic- 031: Balance Among Learning Objectives and Their Weightin table of specification

In developing a test blueprint first of all it is necessary to select some learning. Objectives and among this list of learning objectives some objectives are more important in sense that more time of instruction is spent on them while some other are less important in terms of time spent on them in classroom so in developing table of specification balance among these learning objectives is important, for this purpose we need to weigh the learning objectives for calculating their relative weightage in test.

Step 1: Instruction Time

To do the calculations for the instruction time for columns of the table of specifications the teacher must use the following formulas for each objective in the table.

Time in class spent on objective (min) / total time for the instruction being examined (min)

Percentage of instruction time= $\frac{\textit{time spent on objective, content, theme (min)}}{\textit{Total time for the instruction being examined (min)}}$

Percentage of instruction time= $\frac{250}{1000}$

Percentage of instruction time = 25%

Step 2: Examine value

Then the instructor should look at the number of test items/score to be allocated to objective/content/theme 1.

Let us assume total marks of the are100. Then 25 marks should be allocated to questions related to objective/content/theme 1.

Step 3

Percent of instruction time = Percent of examination value (within ±2 percent, if not, redo test)

25 ±2 = 25 ±2

It can be a bit tricky if the total marks of the test are 50. Then 25% of 50 will be 12.5 marks. Point total of questions for objective / total points * on examination = % of examination value

Topic- 032: Balance among Learning Objectives and their Weight in table of specification: Example

We have learnt to give weightage to the content area in a table of specification. Now we will look at an example to develop table of specification practically. Following is the table of specification comprised of topics to be covered in the test and their weightage that represent percentage of marks for each topic.

Topics/Level	Knowledge	Comprehension	Application	Marks
Pakistan Movement Time: (100/500)*100 = 20%				
Geography of Pakistan Time: (150/500)*100 = 30%				
Climate Change Time: (150/500)*100 = 20%				
Industries Time: (50/500)*100 = 10%				
Economy Time: (50/500)*100 = 10%				
Total (Time: 500/Marks: 50)				

Let us consider that we have to develop a test of 50 marks according to the above discussed table of specification then distribution of marks for each topic is as under.

Topics/Level	Knowledge	Comprehension	Application	Marks
Pakistan Movement Time: (100/500)*100 = 20%				10 (20%)
Geography of Pakistan Time: (150/500)*100 = 30%				15 (30%)
Climate Change Time: (150/500)*100 = 20%				15 (30%)
Industries Time: (50/500)*100 = 10%				5 (10%)
Economy Time: (50/500)*100 = 10%				5 (10%)
Total (Time: 500/Marks: 50)				50 (100%)

Then we have to consider the importance of each topic for cognitive level of questions according to Bloom‘s Taxonomy.

Topics/Level	Knowledge	Comprehension	Application	Marks
Pakistan Movement Time: (100/500)*100 = 20%	5 (50%)	2 (20%)	3 (30%)	10 (20%)
Geography of Pakistan Time: (150/500)*100 = 30%	2 (10%)	6 (40%)	7 (50%)	15 (30%)
Climate Change Time:		7 (50%)	8 (50%)	15 (30%)

(150/500)*100 = 20%				
Industries Time: (50/500)*100 = 10%	1 (10%)	1 (20%)	3 (70%)	5 (10%)
Economy Time: (50/500)*100 = 10%	1(20%)	1 (20%)	3 (60%)	5 (10%)
Total (Time: 500/Marks: 50)	9 (18%)	17 (34%)	24 (48%)	50 (100%)

Lesson 10

SELECTION OF TEST

Topic- 033: Selecting Pre-designed

Published test are designed and conducted in such a manner that each and every characteristic is pre planned and known. There are many published tests available for school use. The two most value to the instructional program are:

i. Achievement tests
ii. Aptitude tests

There are hundreds of tests available for each type. Selecting the most appropriate one is important task. In some cases published tests are used by teachers. But more frequently these are used by provincial or national testing programs.

In classrooms most used published tests are:

i. Achievement tests
ii. Reading test

Published tests commonly used by provincial or national testing programs are:

- Aptitude tests
- Readiness tests
- Placement tests

Topic- 034: Standards for Selecting Appropriate Test –I

Test users should select tests that meet the purpose for which they are to be used and that are appropriate for intended population.

- First define the purpose for testing and the population to be tested and select the test accordingly.
- Investigate the potentially useful sources of information, in addition to the test scores, to validate the information provided by tests.
- Read the materials provided by test developers and avoid using tests for which unclear or incomplete information is provided.
- Become familiar with how and when test was developed and tried out.

Topic- 035: Standards for Selecting Appropriate Test –II

Test users should select tests that meet the purpose for which they are to be used and that are appropriate for intended population.

- Read independent evaluations of a test and of possible alternative measures.
- Examine specimen sets, disclosed tests or sample questions directions, answer sheets, manuals and score reports before selecting the tests.
- Select and use only those tests for which the skills needed to administer the test and interpret scores correctly are available.

Topic- 036: Fairness in Selecting Appropriate Test

- Evaluate the procedures used by test developers to avoid potentially insensitive content or language.
- Review the performance of test takers of different races, gender, and ethnic groups when sample of sufficient size are available.
- Evaluate the extent to which the performance differences may have been caused by inappropriate characteristics of test.
- Use appropriately modified forms of tests or administration procedures for test takers with handicapping conditions.

Lesson 11

CHARACTERISTICS OF A GOOD TEST-I

Topic- 037: Characteristics of Good Test: Validity, Reliability and Usability

The most essential characteristics of are:

- Validity
- Reliability
- Usability

Validity

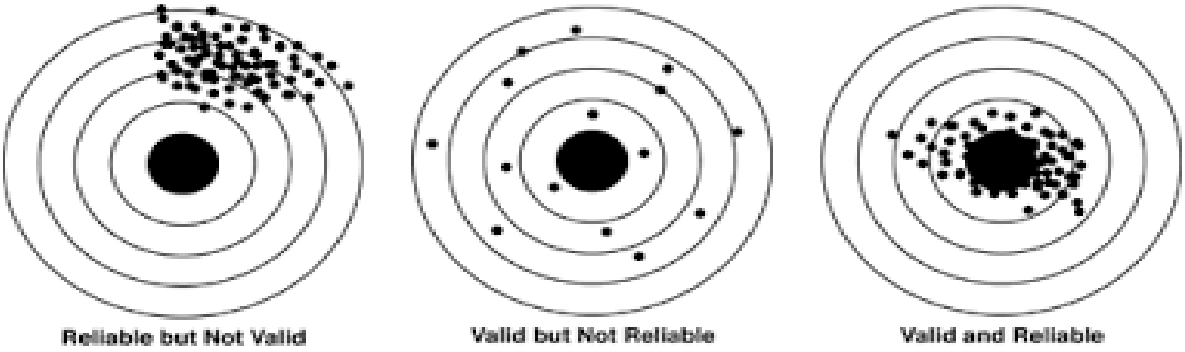
Validity is an evaluation of adequacy and appropriateness of the interpretation and uses of results. It determines if a test is measuring what it intended to measure.

Reliability

Reliability refers to the consistency of assessment results.

Reliability vs. Validity

Reliability of measurement is needed to obtain the valid results, but we can have reliability without validity. Reliability is necessity but not sufficient condition for validity.



Usability

In addition to validity and reliability, an assessment procedure must meet certain practical requirements which include feasibility, administration environment and availability of results for decision makers.

Topic- 038: Nature of Validity

Following are different points that describe the nature of validity.

1. Appropriateness of the interpretation of the results

Validity is referred as ―validity of test but it is in fact validity of the interpretation and use to be made of the results.

2. Validity is a matter of degree

It does not exist an all or none basis. It is best considered in term of categories that specify degree, such as high, moderate or low validity

3. Validity is specific to some particular use or interpretation

No assessment is valid for all purposes. An arithmetic test may have high degree of validity for computational skill and low degree for arithmetical reasoning

4. Validity is a unitary concept

Validity does not have different types. It is viewed as a unitary concept based on different kind of evidences

5. Validity involves an overall evaluative judgment

It requires an evaluation in terms of the consequences of interpretations and uses of assessment results

Topic- 039: Evidences of Validity: Content Validity

Three Evidences of Validity

- Content
- Construct
- Criterion

Meaning

How well the sample of assessment tasks represents the domain of the tasks to be measured.

Procedure

It compares the assessment tasks to the specifications describing the task domain under consideration

Method

1. Classroom instruction determines which intended learning outcomes (objectives) are to be achieved by students.
2. Achievement domain specifies and delimits a set of instructionally relevant learning tasks to be measured by an assessment.
3. Instructional and assessment priorities specifies the relative importance of learning objectives to be assessed.

Lesson 12

CHARACTERISTICS OF A GOOD TEST-II

Topic- 040: Evidences of Validity: Construct Validity

Meaning

How well a test measures up to its claims? A test designed to measure *depression* must only measure that particular construct, not closely related ideals such as *anxiety* or *stress.*

Procedure

Develop a test framework;

- Defining construct
- Identifying sub-constructs
- Listing indicators of each sub-construct
- Writing test items for each indicator

Construct: Essay Writing

Sub-construct	**Meaning/Scope**	**Indicators**
Introduction Paragraph	It introduces the main idea, captures the interest of reader and tells why topic is important.	1. Single sentence called the thesis statement is written 2. Background information about your topic provided 3. Definitions of important terms written
Supporting Paragraphs	Supporting paragraphs make up the main body of your essay	1. List the points about main idea of essay. 2. Write separate paragraph for each supporting point. 3. Develop each supporting point with facts, details, and examples.
Summary Paragraph	Concluding paragraph comes after you have finished developing your ideas.	1. Restate the strongest points of 2. Restate the main idea 3. Give your personal opinion or suggest a plan for action.

Method

There are two methods to confirm construct validity of a test.

1. Expert judgment

There are experts in the field. For the above example, people who are expert in essay writing will be considered to assess the construct validity of the table and the table will be revised under their guidance.

2. Factor analysis

In this, we group the questions by keeping in view the responses of respondents on them.

Topic- 041: Evidences of Validity: Criterion Validity

Meaning

It demonstrates the degree of accuracy of a test by comparing it with another test, measure or procedure which has been demonstrated is needed to be valid.

Concurrent validity

This approach allows one to show the test is valid by comparing it with an already valid test

Predictive

It involves testing a group of subjects for a certain construct, and then comparing them with results obtained at some point in the future.

Procedure

It compares assessment results with another measure of performance obtained at a later date (for prediction) or with another measure of performance obtained concurrently (for estimating present status)

Method

The degree of relationship can be described more precisely by statistically correlating the two sets of scores. The resulting *correlation coefficient* provides numerical summary of relationship

Topic- 042: Evidences of Validity: Consequence Validity

Meaning

How well use of assessment results accomplishes intends purposes and avoids unintended effects

Procedure

Evaluate the effects of the use of assessment results on teachers and students. Both, the intended positive effects (e.g., increased learning) and possible unintended negative effects (e.g. dropout of school) need to be evaluated

Considerations

• Does the assessment artificially constrain the focus of student‘s study?

• Does the assessment encourage or discourage exploration and creative modes of expression?

Factors in Test or Assessment Itself

- Unclear directions
- Reading vocabulary and sentence structure too difficult
- Ambiguity
- Inadequate time limits (construct irrelevant variance)
- Overemphasis of easy to access aspects of domain at the expense of important, but hard to access aspects
- Test items inappropriate for the outcomes being measured
- Poorly constructed test items
- Test too short
- Improper arrangement of items
- Identifiable pattern of answers

Lesson 13

CHARACTERISTICS OF A GOOD TEST-III

Topic- 043: Nature of Reliability

Reliability refers to the consistency of measurement.

- Reliability refers to the results obtained with an assessment instrument and not to the instrument itself.
- An estimate of reliability always refers to particular type of consistency (stability, equivalence, internal consistency)
- Reliability is necessary but not sufficient condition for validity.
- Reliability is primarily statistical (range +1 and -1).

Topic- 044: Method of Estimating Reliability

1. Stability:

Consistency over period of time

2. Equivalence:

Over different forms of assessment

3. Internal consistency:

Within the assessment itself

Determining Reliability by Correlation Methods

In determining reliability, it would be *desirable* to obtain two sets of measures *under identical conditions* and then to compare the results. The reliability coefficient resulting from each method must be interpreted according to type of consistency being investigated.

Method to Estimate Reliability

- Test-Retest (stability)
- Equivalent Forms (equivalence)
- Test- retest with Equivalent Forms (stability and equivalence)
- Split Half (internal consistency)
- Kuder- Richardson(internal consistency)
- Cronbach Alpha (internal consistency)
- Inter-rater reliability (consistency of rating)

Topic- 045: Test-retest Method

- Test-retest method is a measure of stability
- It gives the same test twice to the same group with any time interval between tests, Time interval can range from several minutes to the several years

Test- retest	
September 25	**October 15**
Form A	Form A
1. Item a yes	2. Item a yes
2. Item b no	2. Item b no
3. Item c yes	3. Item c yes

Time interval is key point in this type

- Short interval will provide inflated coefficient of reliability.
- Very long interval will influence results by instability and actual changes in students over time.

Lesson 14

CHARACTERISTICS OF A GOOD TEST-IV

Topic- 046: Method of Estimating Reliability: Equivalent Form Method

1. Equivalent Forms method
2. Test- Retest with Equivalent Forms

Equivalent Forms method

• The equivalent forms method is a measure of equivalence.

• It gives two forms of the test to the same group in close succession.

September 25	**September 25**
Form A	Form B
1. Item a yes	2. Item d yes
2. Item b yes	2. Item e yes
3. Item c No	3. Item f No

Test- retest with Equivalent Forms

• Test- retest with the equivalent forms is a measure of stability and equivalence.

• It gives two forms of the test to the same group with increased interval between forms

Equivalent Forms Method	
September 25	**September 25**
Form A	Form B
1. Item a	2. Item a
2. Item b	2. Item b
3. Item c	3. Item c
Test- retest with Equivalent Forms	
September 25	**October 15**
Form A	Form B
1. Item a	2. Item a
2. Item b	2. Item b
3. Item c	3. Item c
Score = 82	Score= 74

Topic- 047: Method of Estimating Reliability: Split Half Method

Split Half Method

• Split half method is a measure of internal consistency.

• It gives test once. Score two equivalent halves of test, correct correlation between halves to fit whole test by spearman –brown formula

Split Half Method		
Sum number of odd items correct	Sum number of even items correct	September 25
Item 1	Item 2	

		1. Item 1
Item 3	Item 4	2. Item 2
Item 5	Item 6	3. Item 3
.	.	4. Item 4
.	.	5. Item 4
.	.	6. Item 4
.	.	
Odd	Even	
Score = 40	Score= 42	Total Score=82

Split half reliabilities tend to be higher than equivalent form reliabilities because split half method is based on the administration of single assessment.

Topic- 048: Method of Estimating Reliability: Kuder-Richardson Method

1. Kuder- Richardson methods and Coefficient Alpha
2. Inter-Rater Method

Kuder- Richardson methods and Coefficient Alpha

• Kuder- Richardson methods and Coefficient Alpha is also a measure of internal consistency.

• It gives test once. Score total test and apply Kuder- Richardson

As with the split half method, these formulas provide an index of internal consistency but do not require splitting the assessment in half for scoring Purposes. One formula KR20 is applicable only when student responses are scored dichotomously (0 or 1). It is the most useful with traditional test items scored correct or incorrect. The generalization of KR20 for assessments that have more than dichotomous, right-wrong scores is called Coefficient Alpha.

Inter-rater Method

• Inter-rater method is a measure of consistency of ratings

• It gives a set of student's responses requiring judgmental scoring to two or more raters and has them independently score the responses

Lesson 15

TYPES OF ASSESSMENT TOOLS-I

Topic- 049: Anecdotal Records

Many outcomes in the cognitive domain, such as those pertaining to knowledge, understanding, and thinking skills, can be measured by paper pencil tests. But there are still many learning outcomes that require informal observation of natural interactions.

Types of assessment tools

Learning outcomes aspects of development can generally be assessed by:

1. Observing students as they perform and describing or judging that behaviors (Anecdotal record).
2. Asking their peers about them and assessing social relationships (Peer appraisal).
3. Questioning them directly and assessing expressed interests (Self-appraisal).
4. Measuring progress by recorded work (portfolio).

Anecdotal records

Impressions gained through observation are apt to provide an incomplete and biased picture, however, unless we keep an accurate record of our observations. The method to do so is called anecdotal records. Anecdotal records are factual descriptions of meaning incidents and events that the teacher observes.

Topic- 050: Effective Use of Anecdotal Records

One should keep in mind the following points to use anecdotal records effectively.

- Determine in advance what to observe but be alert of unusual behavior.
- Analyze observational records for possible sources of bias.
- Observe and record enough of the situation to make behavior meaningful.
- Make record of the incident as soon after the observation is possible.
- Limit each anecdote to a brief description of a single incident.
- Keep the factual description of the incident and your interpretation of it separate.
- Record both positive and negative behavioral incidents.
- Collect a number of anecdotes on a student before drawing inferences concerning typical behavior.
- Obtain practice in writing anecdotal records.

Topic- 051: Advantages and Limitations of Anecdotal Records

Following are the advantages of anecdotal records.

- It depicts actual behaviors in natural situations.
- It facilitates gathering evidence on the events that are exceptional but significant.
- It is beneficial for the students with less communication skills.

Limitations of Anecdotal Records

Following are the limitations of anecdotal records.

It takes long time to maintain.

It is subjective in nature.

Anxiety may lead to wrong observation.

Lesson 16

TYPES OF ASSESSMENT TOOLS-II

Topic- 052: Peer Appraisal

In this procedure students rate their peers on the same rating device used by their teacher. It depends on greatly simplified procedures.

Techniques of peer appraisal

There are two widely used techniques in this area are:

1. Guess who
2. Sociometric

Guess who technique

In this technique teacher use a positive or negative behavior of a student as an example. Other students from same group try to guess the statement with that characteristic correctly. Generally behaviors used, for example are positive in nature to avoid any adverse effect on the student pointed out in the example. The guess that technique is based on the nomination method of obtaining peer ratings and is scored by simply counting the number of mentions each student receives on each description.

Sociometric technique

Sociometric technique is a method for assessing the social acceptance of individual students and the social structure group. It is based on students' choice of companion for any group situation. This form was used to measure student's acceptance as seating companions, work companions and play companions.

There are few important principles of sociometric choosing:

1. The choices should be real choices that are the natural part of classroom activities.
2. The basis for the choice and restriction on the choosing should be made clear.
3. All students should be equally free to participate in the activity or situation.
4. The choice of each student make must be kept confidential.
5. The choices should actually be used to organize or rearrange the group.

Topic- 053: Portfolio

Systematic collection of students' work into portfolios can serve a variety of instructional and assessment purposes. The value of portfolios depends heavily on the clarity of purpose the guidelines for the inclusion of materials, and the criteria to be used in evaluating portfolio.

Portfolio of students work

A portfolio is a collection of student work selected to serve a particular purpose, such as documentation of student growth. It is purposeful collection of pieces of student's work.

Key steps in defining and using portfolios

Specify purpose

Provide guidelines for selecting portfolios

- Define student's role in selection and self-evaluation.
- Specify evaluation criteria.

Use portfolios in instruction and communication.

Strengths of portfolios

The can be readily integrated with the instruction.

Provide opportunity to student‘s to show what they can do.
Encourage to become reflective learner.
Help in setting goal and self-evaluation
Help teacher and student to collaborate and reflect on student‘s progress.
Effective way to communicate with parents.
Provide mechanism for student centered and student-directed conferences with parents.
Provide concrete examples of student’s development and current skills.

Weaknesses of portfolios

- Can be time consuming to assemble.
- Hard to use in summative assessment.
- Difficult to compare results.
- Very low reliability.

Topic- 054: Purpose of Portfolio

Fundamentally, two global purposes for creating portfolios of students work: for student‘s assessment and instruction. It can be used to showcase student‘s accomplishment and document the progress.

Instructional purposes:

When primary purpose is instruction, the portfolio might be used as means of:

1. Helping students develop and refine self-evaluation skills.
2. Providing teacher with more reflecting information regarding students‘progress.
3. Set criteria of excellence between teacher and student.
4. Student directed conferences with parents.
5. Access to student thought process and awareness of standards.
6. Teachers ‘student to communicate with different audience.

Assessment purposes:

When emphasis is on assessment it is important to distinguish between formative and summative roles of assessment.

1. It can be used for formative purposes to measure progress.
2. Basis for certifying accomplishment.
3. For system accountability mechanism.

Current accomplishment and progress

When the focus is on accomplishments, portfolios usually are limited to finished work and may cover only a relatively small period of time. When focus is on demonstrating growth and development the time frame is longer. It will include multiple version of same work over time to measure progress.

Showcase and documentation portfolios

It contains student selected entries. It demonstrates students’ ability to choose his best work which demonstrates his ability to do a task. It intended to provide evidence about breath as well as depth of learning. It needs to be more inclusive and not just limited to special strength of student.

Finished and working portfolios

It implies that work is complete for specific audience. A job application portfolio is anexample. It is finished product for specific audience.

Guidelines for portfolios entries

Guidelines should specify:

- The uses that will be made of the portfolio.

- Who will have to do it?
- What type of work is appropriate to include.
- What criteria will be used in evaluating the work?
- Should define timeline for the portfolios.
- Minimum and maximum numbers of entries.

Lesson 17

CREATING FIXED-CHOICE TEST ITEMS: MCQS-I

Topic- 055: Selection of Item in a Test

Selecting Item Format

In accordance with the table of specifications, the teacher decides the item type most appropriate for measuring a stated level of cognitive domains. There are basically two major types of items objective and subjective types. Within objective type items there are: true-false, short answer, multiple choice items, and matching exercises. Which are most appropriate in view the nature of the content, nature of cognitive processes and mental level of the students. The format of the item necessarily proceeds from the test blueprint. The blueprint indicates the kinds of skills and the balance of test content to be measured
The format of the item necessarily proceeds from the test blueprint. The blueprint indicates the kinds of skills and the balance of test content to be measured.

The selection of item format should be based on the kinds of skills to be measured and not on some personal like or dislike for a particular item format. The use of multiple-choice questions may make sense for large group testing the knowledge of the mechanics of English. This type of item is not generally appropriate, though, as a direct measure of writing skill. If the intent is to determine whether an examinee can write a clear, coherent essay, then an essay or the free-response format is clearly more appropriate than a multiple-choice format. There is no inherent goodness or badness in any type of question format. The choice must be made on the basis of the behavior to be tested.

Types of Objective Type Items

1. Selection Type
2. Supply Type

In selection type, true-false or alternative form, matching exercises, and multiple choice items are included. Another category of objective type items that is supply type consisted of completion or fill in the blanks and short answer items. The selection of any one or combination of these is made on the basis of their suitability for measuring the desired learning outcome.

Topic- 056: Characteristics of MCQs

The multiple choice items are generally recognized as the most widely applicable and useful types of objective test items that can measure knowledge, comprehension and as well as application level learning outcomes.
Multiple choice items consist of two parts:

i. A Stem
ii. No. of options or alternatives

Stem

The problem may be stated as a direct question or an incomplete statement and is called the stem of the item.

Alternatives

The list of suggested solutions which may include words, numbers, symbols or phrases are called alternatives (also called choices or options). The correct alternative in each item is called the answer

while all incorrect or less appropriate alternatives are called distractors or foiled and the student‘s task is to select the correct or best alternative from all the options.

Format of MCQs

Whether to use a direct question or incomplete statement in the stem depends on several factors.

Direct questions: the direct questions format is easier to write, is more natural for younger students and is more likely to present a clearly formulated problem.

Example of direct question:

Which of the following cities is the capital of Pakistan?

a) Islamabad
b) Karachi
c) Lahore
d) Quetta

Example of incomplete question:

The capital of Pakistan is:

a) Islamabad
b) Karachi
c) Lahore
d) Quetta

Selection of Format

A common procedure is to start each stem as a direct question and shift to the incomplete statement form only when the clarity of the problem can be retained and greater conciseness achieved.

Format of MCQs

When we get beyond the simple aspects of knowledge, represented by questions of who what when and where variety, answers of varying degree of acceptability are the rule rather than the exception.

The question of why variety tends to reveal a number of possible reasons, some of which are better than others. The questions of how variety usually reveals several possible procedures some of which are more desirable then the others. Measures of achievement in these areas then become a matter of selecting the best answer. This type is useful for measuring SLOs that require understanding, application, or interpretation of factual information.

Example of best answer type MCQs

Which of the following factors contributed most to the selection of Islamabad as capital of Pakistan?

a) Central location
b) Good climate
c) Large population
d) Good highways

Best answer type MCQs

The best answer type MCQs tend to be more difficult than the correct answer type. This type is used to measure more complex level of learning in students.

Topic- 057: Uses of MCQs –I

The MCQs is the most versatile types of test item available. It can measure varieties of learning outcomes from simple to complex, and it is adaptable to the most types of matter content.

Despite the wide applicability of the MCQs there are SLOs, such as the ability to organize and present ideas that cannot be measured with any form of selection item. For such skills constructed and restricted response questions are used.

Uses of MCQs in measuring knowledge outcomes

Learning outcomes in the knowledge are so prominent in all school subjects, and MCQs can measure variety of these outcomes. Some of the typical uses are:

Knowledge of Terminology

student can be asked to show their knowledge of particular term by selecting a word that has same meaning as given term.

Example: Which of the following words has same meaning as the word Egress?

a) Depress

b) Enter

c) Exit

d) Regress

Example: Which of the following statements best define the word degree?

a) An expression of disapproval

b) **An act of leaving an enclosed place**.

c) Proceeding to higher level

Topic- 058: Uses of MCQs –II

Knowledge of Principles

Knowledge of principle is an important learning outcome in most school subjects. MCQs can be constructed to measure knowledge of principle as easily as those designed to measure facts.

Example: The principle of capillary action helps explain how fluids;

a) Enter solution of lower concentration

b) Rise in fine tube

c) Escape through small openings

Knowledge of Methods

This includes diverse areas such as knowledge of labs, knowledge of methods underlying communication, computational and performance skills. In some cases we might want to measure knowledge of procedures before we permit students to practice in a particular area. In other cases knowledge of methods may be an important learning outcome in their own right.

Example: To make legislation the prime minister of Pakistan must have the consent of:

a) Parliament

b) Ministry of law

c) Military command

d) Supreme court

These uses and their examples have merely scratched the surface of MCQs measuring knowledge outcomes. MCQs can be used to measure the much more complex nature of knowledge. But that is left for more advanced course.

Lesson 18

CREATING FIXED-CHOICE TEST ITEMS: MCQS-II

Topic- 059: Uses of MCQs-III

Another learning outcome basic to all school subjects is the knowledge of specific facts. It provides the basis for developing understanding, thinking skills and other complex learning. MCQs designed to measure specific facts can take many forms, that questions of who, what, when, and where variety are most common.

Example: Who was 1st astronaut to land on moon?

a) Buzz Aldrin

b) Neil Armstrong

c) Yuri Gagarin

Example: What was the name of space shuttle that landed on the moon?

a) Apollo

b) Atlas

c) Midas

d) Polaris

Example: When did first man landed on the moon?

a) 1962

b) 1966

c) 1967

d) 1969

Topic- 060: Advantages and Limitations of MCQs –I

The MCQ is one of the most widely applicable test items for measuring knowledge, achievement. It can effectively measure various types of knowledge and complex learning outcomes. It is free from some of the common shortcomings which are characteristics of the other test items like ambiguity and vagueness usually associated with short questions.

Example of vague question:

Quid -e -Azam was born in _____.

There can be multiple correct answers for this question. This is poorly constructed item.

Quid -e -Azam was born in:

a) Karachi

b) Lahore

c) Peshawar

d) Dehli

Topic- 061: Advantages and Limitations of MCQs –II

MCQs reduces the risk of guessing the correct answer. You have to know the correct answer. There is a high chance of wrong answer if you solely depend on guesses.

Example: quid -e -Azam was born in 1867. **True/False**

The student will receive score even if he didn‘t know the correct year of birth Quid –e- Azam if he/she tick n false.

In which year Quid- e- Azam was born?

a) 1867

b) 1876
c) 1878
d) 1887
MCQs items reduced the probability of guessing as compared to other form of item.

Topic- 062: Advantages and Limitations of MCQs –III

Advantages of MCQs

- Ensure objectivity, reliability and validity; preparations of questions with colleagues provide constructive criticism.
- Increase significantly the range and variety of facts that can be sampled in given time.
- Provide precise and unambiguous measurement of the higher intellectual processes.
- Provide detailed feedback for both students and teachers.
- MCQs are easy and rapid to score.

Limitations of MCQs

- Take long time to construct in order to avoid arbitrary and ambiguous questions.
- Require careful preparation to avoid multitude of questions testing only recall.

Lesson 19

CREATING FIXED-CHOICE TEST ITEMS: MCQS-III

Topic- 063: Suggestions for Constructing MCQs -I

The general applicability and the superior qualities of multiple choice test items are realized most fully when care is taken in their construction. This involves formulating clearly stated problems, identifying plausible alternatives, and removing irrelevant clues to the answer.

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

Explanation: often the stem of the test placed in MCQs form are incomplete statements that are make little sense until all the alternatives have been read. This is not MCQs but rather and true false question placed in MCQs form.

Example: **Poor item**

South America

a) Is flat, arid country

b) Imports coffee from the United States

c) Has larger population then Europe

d) Was settled by colonist from Spain

Example: **Better item**

Most of the South America was settled by colonists from

a) England

b) France

c) Holland

d) Spain

Topic- 064: Suggestions for Constructing MCQs –II

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

Explanation: Clear stem increase the probability of the item as well as reduce the reading time required.

Example: **Poor item**

Most of the Indian subcontinent was settled by colonists from Britain. How would your account for the large number of colonists settling there?

a) They are adventurers

b) They were in search of wealth

c) They wanted lower taxes

d) They were seeking religious freedom

Example: **Better item**

Why did Britishers settled in India?

a) For adventures

b) For wealth

c) For lower taxes

d) For religious freedom

Topic- 065: Suggestions for Constructing MCQs –III

Try to avoid the negative statements, unless the significant learning outcome requires it.

Explanation: These avoid the possibility of student overlooking ―no, or ―least and similar words used in negative items.

Example: **Poor item**

Which of the following cities is not located in north Islamabad?

a) Abbottabad

b) Gilgit

c) Lahore

d) Mingora

Example: **Better item**

Which of the following cities is located in south Islamabad?

a) Abbottabad

b) Gilgit

c) Lahore

d) Mingora

Topic- 066: Suggestions for Constructing MCQs –IV

All alternatives should be grammatically consistent with the stem of item.

Explanation: in the following examples note how the better version results from a change in the alternatives in order to obtain grammatical consistency. The main function of this rule is to prevent irrelevant clues from entering.

Example: **Poor item**

An electric transformer can be used

a) For strong electricity

b) To increase the voltage of alternating current

c) It converts electrical energy into mechanical energy

d) Alternating current is changed to direct current

Example: **Better item**

An electric transformer can be used to

a) Produce strong electricity

b) Increase the voltage of alternating current

c) Convert electrical energy into mechanical energy

d) Change alternating current to direct current

Lesson 20

CREATING FIXED-CHOICE TEST ITEMS: MCQS-IV

Topic- 067: Suggestions for Constructing MCQs –V

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

Explanation: Including more than one correct answer in a test item and asking students to select all the correct alternatives has 2 shortcomings.

a. Such items are usually no more than a collection of true and false item presented in MCQ form.

b. The number of alternatives selected as correct answers varies from one student to another.

Example: **Poor item**

Pakistan borders on:

a) India

b) Tajikistan

c) Saudi Arabia

d) China

Example: **Better item**

Pakistan borders on:

a) India T/F

b) Tajikistan T/F

c) Saudi Arabia T/F

d) China T/F

Topic- 068: Suggestions for Constructing MCQs –VI

All distractors should be plausible. The purpose of distractor is to confuse the uniformed.

Explanation: to the student who has not achieved the learning outcome being tested, the distractor should be as attractive as the correct answer. If properly constructed, each distractor will be selected by some students. If the distractor is not selected by anyone, it is not contributing to the functioning of the item and should be eliminated or revised.

Example: **Poor item**

Who wrote national anthem of Pakistan?

a) Allama Iqbal

b) Christopher Columbus

c) Hafeez Jullandhuri

d) Ibrar ul Haq

Example: **Better item**

Who wrote national anthem of Pakistan?

a) Allama Iqbal

b) Habib Jalib

c) Hafeez Jullandhuri

d) Munir Niazi

Topic- 069: Suggestions for Constructing MCQs –VII

Verbal association between the stem and the correct answer should be avoided.

Explanation: Frequently a word in the correct answer will provide an alternative clue because it looks or sounds like the word in the stem of the item. However, word similar to those in the stem might be included in the distractors to increase their plausibility. Students who depend on rote memory and verbal association will then led away, from rather than to the correct answer.

Example: **Poor item**

Which of the following agencies should you contact to find about a flood?

a) National flood relief
b) Local radio station
c) Pakistan post
d) Pakistan weather bureau

Example: **Better item**

Which of the following agencies should you contact to find about a flood?

a) Disaster management office
b) Radio station
c) Post office
d) Weather bureau

Topic- 070: Suggestions for Constructing MCQs –VIII

The relative length of the alternatives should not provide a clue to the answer.

Explanation: the best we can hope for in equalizing the length of the test item‘s alternatives is to make them approximately equal. But because the correct answer is usually needs to be qualified, it tends to be longer than the distractors unless a special effort is made.

Example: **Poor item**

What is the major purpose of United Nations?

a) To maintain peace among people of the world
b) To establish international law
c) To provide military control
d) To form new governments

Example: **Better item**

What is the major purpose of United Nations?

a) To maintain peace among people of the world
b) To develop new system of international law
c) To provide military control of new nations
d) To establish democratic forms of governments

Lesson 21

CREATING FIXED-CHOICE TEST ITEMS: TRUE/ FALSE AND ITS USES-I

Topic- 071: True/ False Items

Alternate-response test items consist of declarative statement that the student is asked to mark true or false, right or wrong, correct or incorrect, or the like.

Uses of True-false (Alternate form questions)

For measuring such relatively simple learning outcomes, a single declarative statement is used with anyone of several methods of responding.

Example: Directions, read each of the following statements. If the statement is true, encircle T and if statement is false encircle F.

A river is bigger than a stream. T F

Founded is the past tense of found. T F

Dozen is equivalent to 20. T F

Example: Directions, read each of the following statements. If the answer is yes, encircle Y and if answer is no encircle N.

Is 51% of 38 more than 19? Y N

Is 50% of 4/10 equal to 2/5? Y N

If 60% of a number is 9, is the number smaller than 9? Y N

One of the most useful functions of true and false items is in measuring the students' ability to distinguish fact from opinion.

Example: Directions, read each of the following statements. If the is a fact encircle F and if statement is an opinion encircle O.

Current constitution of Pakistan is written in 1973. F O

Pakistan progressed most under dictator rule. F O

18th amendment decentralized the ministry of education. F O

Topic- 072: Uses of True/ False Items

Another aspect of understanding that can be measured by true and false item is ability to recognize cause and effect relationship. This type of item usually contains true propositions in one statement, and the student is to judge whether the relationship between them is true or false.

Example: Directions, read each of the following statements, both parts of statements are true. You are to decide whether the second part explains why the first part is true. If it does, encircle Yes. If it does not, encircle No.

Leaves are essentials because they shade the tree trunk. Yes No

Some plants do not need sunlight because they get food from other plants. Yes No

It could be used to measure some simple aspect of logic.

Example: Directions, read each of the following statements. If the statement is true encircle T. if the statement is false encircling F. Also, if the converse of the statement is true circle CT, if the converse of statement is false circle CF. Be sure to give 2 answers for each statement.

All trees are plant. T F CT CF

All parasites are animals. T F CT CF

All eight legged animals are spiders. T F CT CF

Topic- 073: Advantages and Limitations of True-False

True-False (Alternative Form) Questions

Alternative form of question requires students to select any one of the two given categories. The categories may be true-false, yes-no, correct-incorrect or fact-opinion. Alternative form of items is most suitable for measuring lower level learning outcomes. True-false items can be used in different forms

Advantages of Alternative Form (True-False)

True-false questions are well suited for testing lower level learning outcomes, like a student's ability to;

i. Identify the correctness of factual statements, e.g. Earth is a planet.
ii. Definition of terms e.g. Photosynthesis is the process by which leaves make food for plants.
iii. Statement of principles, e.g. Earth is revolving around the sun.
iv. Distinguish facts from opinion, e.g. Islam is the official religion of Pakistan.
v. Recognize cause-and-effect relationship

From a teacher's point of view alternative-form questions are very useful when;
A lot of content is to be covered in a fairly short amount of time.
The time available for scoring is very short.

Limitations of Alternative Form (True-False)

Most commonly observed limitations are:

i. All learning outcomes cannot be measured through alternative form questions. They are generally limited to lower level learning outcomes.
ii. Ease of guessing correct answers when the answer is not known. With only two choices\ (true or false) the student could expect to guess correctly on half of the items for which correct answers are not known.
iii. There is sometimes a tendency to take quotations from the text with a minor change in wording.
iv. There may also be a tendency to include trivial material from the text.
v. True-false items are prone to high guessing and can only be used for measuring lower level learning outcomes

Topic- 074: Suggestions for Constructing True-False (Alternative Form) Items –I

Most important tsk in formulating statements is free from ambiguity and irrelevant clues. There are list of things to avoid when phrasing the statement.

1. Avoid broad general statements if they are to be judged true or false.

Explanation: Most broad generalizations are false unless qualified, and the use of qualifiers provides clues to the answer.

Example: **Poor item**

The president of Pakistan is *usually* elected to his/her office.

Example: **Better item**

According to constitute of Pakistan, the president of Pakistan is elected by parliament.

Topic- 075: Suggestions for Constructing True-False Items –II

Avoid trivial statements.

Explanation: To obtain the statements that are clearly true and false, we turn to specific statements of fact that fits the criterion but have little significance from learning point of view.

Example: **Poor item**

Mamnoon Hussain is 12th president of Pakistan.

Example: **Poor item**
India declared war on Pakistan on September 3rd, 1965.

CREATING FIXED-CHOICE TEST ITEMS: TRUE/ FALSE AND ITS USES-II

Topic- 076: Suggestions for Constructing True-False Items –III

Avoid the use of negative especially double negative statements.

Explanation: Students tend to overlook negative words such as no or not, and double negative contributes to statement's ambiguity. If absolutely necessary, it should be underlined or put in italic so that students do not overlook it.

Example: **Poor item**

None of the steps in the experiment was unnecessary.

Example: **Better item**

All the steps in the experiment were necessary.

Topic- 077: Suggestions for Constructing True-False Items –IV

Avoid long and complex sentences.

Explanation: A test item should indicate whether a student has achieved the knowledge or understanding being measured. Long, complex sentences tend also to measure the extraneous factor of reading comprehension. If avoiding long and complex statement is not possible then to may be necessary to change to another item form in order to avoid a complex sentence structure.

Example: **Poor item**

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

Example: **Better item**

Litmus paper turns red in an acid solution.

Topic- 078: Suggestions for Constructing True-False Items –V

Avoid including two ideas in one statement unless cause and effect relationship are being measured.

Explanation: Students can get confused and answer for only one statement. Other statement can have different answers. It is better to use two simple statements instead of including two ideas in one statement.

Example: **Poor item**

Pakistan could not qualify for world cup hockey because of poor resources, low talent and government support. (T)

Pakistan could not qualify for world cup hockey because of poor resources, low talent. (F)

Pakistan could not qualify for world cup hockey because of poor resources. (T)

Topic- 079: Suggestions for Constructing True-False Items –VI

Avoid using opinion that is not attributed to some source.

Explanation: A statement of the opinion cannot be marked true or false, and it is unfair to expect students to guess how the teacher will score such items, or to respond to opinion statements as statements of facts.

Example: **Poor item**

All anti-state activists should be hanged

Example: **Better item**

National action plan allows the justice system to give death penalty to anti-state activists if proven guilty.

Topic- 080: Suggestions for Constructing True-False Items –VII

Avoid using true statements and false statements that are unequal in length.

Explanation: There is a natural tendency for true statements to be longer because such statements must be precisely phrased in order to be absolutely true. This can be overcome by lengthening the false statements through the use of qualifying phrase similar to those found in true statements.

Lesson 23

CREATING FIXED-CHOICE TEST ITEMS: MATCHING EXERCISES-I

Topic- 081: Matching Exercises

The matching type exercises consist of two parallel columns, first column premises and the second one is responses and the directions for matching the two columns. Matching test items are also selection items specially used to measure a student's ability to identify the relationship between a set of similar items, each of which has two components. Such as words and their definitions, symbols and their meanings, dates and events, people and their accomplishments, etc.

Matching exercise is economical methods when used with content which has the sufficient homogeneous factual information. In developing matching items, there are two columns of material. Matching exercise is used when measuring a student's ability to identify the relationship between a set of similar items, each of which has two components.

Instructions: A list of premises includes a list of prominent Pakistanis and the list of responses give their reason of fame. Match the given names with their respective reason of fame. There can be more than one reason of fame listed for one name and any of the given names may not correspond to any of the reasons of fame.

List A (Premises)	**List B (Responses)**
Quaid-e-Azam	a. Player
Jahanger Khan	b. Prime Minister
Abul Qadeer Khan	c. Statesman
Parveen Shakir	d. Poet
Benazir Bhutto	e. Scientist
	f. Civil Servant

Topic- 082: Uses of Matching Exercises

The typical matching exercise is limited to measuring factual information, based on simple associations. It is a compact and efficient method of measuring such simple knowledge outcomes. Examples of relationships considered important by teachers, in a variety of fields, including the following:

Persons	Achievements
Dates	Historical events
Terms	Definitions
Rules	Examples
Symbols	Concepts
Authors	Titles of books
Persons	Achievements
Foreign words	Local equivalents
Machines	Uses
Plants or animals	Classifications
Principles	Illustrations
Objects	Name of objects
	Functions

It is also being used with pictorial materials in relating pictures and words to identify positions on maps, charts and diagrams. Regardless of the dorm of presentation, the student‘s task is essentially to relate two things that have logical association. This restricts the use of matching exercise to small area of student‘s achievement.

Topic- 083: Advantages and Limitations of Matching Exercises

Advantages of Matching Exercises

- The major objective of matching exercise is its compact form, which makes it possible to measure a large amount of related factual material in a relatively short time.
- Another advantage is ease of construction. Poor matching items can be rapidly constructed, but good matching item requires a high degree of skills.

Limitations of Matching Exercises

- It is restricted to the measurement of factual information, based on rote learning.
- It is highly susceptible to the presence of irrelevant clues.
- Difficulty of findings homogenous material that is significant from the viewpoint of our objectives and learning outcomes.

Lesson 24

CREATING FIXED-CHOICE TEST ITEMS: MATCHING EXERCISES-II

Topic- 084: Suggestions for Constructing Matching Exercises –I

Use only homogeneous material in the single matching exercise.
Explanation: This is the most violated rule of developing a matching exercise. Homogeneity is a matter of degree. What is homogeneous to one group may be heterogeneous to another.

Topic- 085: Suggestions for Constructing Matching Exercises –II

Include an unequal number of responses and premises, and instruct the students that responses may be used once, more than once, or not at all.
Explanation: This will make all the responses eligible for selection for each premise and will decrease the likelihood of successfully guessing.
Keep the list of items to be matched brief and place the shorter responses on the right.
Explanation: It is easier to maintain homogeneity in a brief list. Four to seven items in each column seems best. Placing shorter responses on the right also contributes to more efficient test tasking.
Instructions: A list of premises includes a list of prominent Pakistanis and the list of responses give their reason of fame. Match the given names with their respective reason of fame. There can be more than one reason of fame listed for one name and any of the given names may not correspond to any of the reasons of fame.

List A (Premises)	List B (Responses)
Quaid-e-Azam	a. Player
Jahanger Khan	b. Prime Minister
Abul Qadeer Khan	c. Statesman
Parveen Shakir	d. Poet
Benazir Bhutto	e. Scientist
	f. Civil Servant

Topic- 086: Suggestions for Constructing Matching Exercises –III

Arrange the list of responses in logical order. Place words in alphabetical order and numbers in sequence.
Explanation: This will contribute to the ease with which student can scan the responses in the searching for the correct answers. It will also prevent them from detecting possible clues from the arrangements of the responses.

Example

Directions: On the line to the left of each historical event in column A, write the letter from Column B that identifies the time period when the event occurred. Each date in Column B may be used once, more than once, or not at all.
Instruction: A list of premises includes a list of prominent Pakistanis and the list of responses give their reason of fame. Match the given names with their respective reason of fame. There can be more than one reason of fame listed for one name and any of the given names may not correspond to any of the reasons of fame.

List A (Premises)	List B (Responses)
Abul Qadeer Khan	Civil Servant
Benazir Bhutto	Player
Jahanger Khan	Poet

Parveen Shakir	Prime Minister
Quaid-e-Azam	Scientist
	Statesman

Topic- 087: Suggestions for Constructing Matching Exercises –IV

Indicate the directions the basis for matching the responses and premises.

Explanation: By following this suggestion ambiguity and confusion can be avoided.

Testing time will be saved, student will not need to read through the entire list of premise and responses and then reason out the basis for matching. Place all of the items for one matching exercise on the same page.

Explanation: This will prevent the students from missing the responses appearing on another page and generally adds to the speed and efficiency of test administration.

CREATING FIXED-CHOICE TEST ITEMS: SHORT QUESTIONS-I

Topic- 088: Developing Short Answer/Completion Questions –I

There are two types of item formats which are included in the category of supply type: short answers and completion and fill in the blanks, under the range of objective test items. Both the short answers and completion items are forms of supply item that can be answered by a word, phrase, number or symbol. They only differ in form of presentation. Short answers are presented in the form of direct questions while completion items are incomplete statements. Both of them can be answered with a word, phrase, number, or symbol. Such items are frequently used for measuring knowledge of terminology, specific facts, principles and procedures etc.

Topic- 089: Developing Short Answer/Completion Questions –II

The short answer and completion items are forms of "supply" items both are supply type test items that can be answered by a word, phrase, number or symbol. Short answers are presented in the form of direct questions while completion items are incomplete statements.
Example:
Short Answer: What is the name of man who invented the light bulb? (Thomas Edison)
Completion: The name of the man who invented the light bulb is _____.
This category includes the problem in arithmetic, mathematics science and other areas whose solution must be supplied by the students.

Topic- 090: Uses of Short Answer/Completion Questions

The short answer test item is suitable for measuring a wide variety of relatively simple learning outcomes. The following outcomes and test items some of its common uses.

Simple interpretation of data.

i. How many syllables are there in word Argentina? (4)
ii. In the number 612, what value does the 6 represent? (600)
iii. If an airplane flying northwest made a 180 degree turn, what direction would it be heading? (Southeast)

More complex interpretations can be made when the short answers item is used to measure the ability to interpret diagrams, charts graphs and pictorial data. More notable exceptions to the general rule that short answer items are limited to measuring simple learning outcomes are found in the areas of mathematics and science.

Lesson 26

CREATING FIXED-CHOICE TEST ITEMS: SHORT QUESTIONS-II

Topic- 091: Advantages of Short Answer/Completion Questions

Advantages of short answer/completion questions

- Teacher wants students to complete a large number of items in a fairly short time (unless they involve working complex mathematical problems).
- Teacher wants to control the possibility of guessing. Since the student has to generate the answers.
- The possibility of guessing the correct answers to these questions is greatly reduced when compared with true-false questions.

Limitations of short answer/completion questions

- A potential problem with both short-answer and completion items is that difficult to frame questions for one specific answer unless the items are well written.
- Not usable to measure complex learning outcomes.

Topic- 092: Suggestions for Constructing Short Answer/Completion Questions –I

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

Example:

Poor: An animal that eats the flesh of other animals is? (Carnivorous)

Better: an animal that eats the flesh of other animals classified as (Carnivorous)

Topic- 093: Suggestions for Constructing Short Answer/Completion Questions –II

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

Example:

Poor: Chlorine is a _____. (Halogen)

Better: Chlorine belongs or a group of elements that combines with the metals to form salts. It is therefore called as_____. (Halogen)

CREATING FIXED-CHOICE TEST ITEMS: SHORT QUESTIONS-III

Topic- 094: Suggestions for Constructing Short Answer/Completion Questions –III

A direct question is generally more desirable than an incomplete statement.
Example:
Poor: Pakistan gained its independence in _____. (1947)
Better: When did Pakistan gained its independence? (1947)
Best: In what year did Pakistan gained its independence? (1947)

Topic- 095: Suggestions for Constructing Short Answer/Completion Questions –IV

If the answer is to be expressed in numerical units, indicate the type of answer.
Example:
Poor: if water melon weighs 1kg 200 grams each. How much 3 watermelons will weighs? (3.5kg and 100 grams or 3600 gram)
Better: if water melon weighs 1kg 200 grams each. How much 3 watermelons will weighs? (3.6kg)

Topic- 096: Suggestions for Constructing Short Answer/Completion Questions –V

When completion items are used, do not include too many blanks or start with a blank space. Also avoid asking trivial information in blank space.
Example:
Poor: (warm blooded) animals that are born (alive) and (suckle) their young are called (mammals).
Better: warm blooded animals that are born alive and suckle their young are called (mammals).

Lesson 28

CREATING CONSTRUCTED RESPONSE TEST ITEMS

Topic- 097: Creating Constructed Response Test Items

Subjective test item
Restricted response essay type items

Subjective test items

Subjective test items (essay questions) are constructed response type questions that can be the best way to measure the students' higher order thinking skills. Such as applying, organizing, synthesizing, integrating, evaluating, or projecting while at the same time providing a measure of writing skills. Essay items can vary from very lengthy (5-10 pages), open-ended to limited or restricted response (one page or less).

Essay type questions are divided into two types

1. Restricted Response Items
2. Extended Response Items

Pose a specific problem for which student needs to recall suitable information, organize it, derive a defensible conclusion, and express it within the given limits of the questions.

Example: List the similarities and differences in the process of cell division in meiosis and mitosis?

Variety of learning outcomes can be checked by using this format of essay question. Some of which are;

1. Analysis of relationship.
2. Compare and contrast positions.
3. Explain cause-effect relationship.
4. Organize data and support a viewpoint.
5. Formulate hypotheses.
6. Point out strengths and weaknesses.
7. Integrate data from various resources.

Teacher can use this type of questions under the following conditions when:

1. Supplying information is required instead of simple recognition.
2. Limited numbers of content areas are needed to be tested.

Topic- 098: Guidelines for Constructing Restricted Response Essay Type Items

- Statements should not be quoted directly from the text.

- Evaluate essay responses anonymously.
- Frame questions so that the examinee‘s task is explicitly defined.
- Specify the value and an approximate time limit for each question.
- Employ a larger number of questions that require relatively short answers rather than only a few questions that require long answers.
- Do not employ optional questions.
- Verify a question‘s quality by writing a trial response to the question.
- Prepare a tentative scoring key in advance of considering examinee responses.
- Score all answers to one question before scoring the next question.
- Make prior decisions regarding treatment of such factors as spelling and punctuation.

Lesson 29

CREATING EXTENDED RESPONSE TEST ITEMS

Topic- 99: Extended Response Essay Type Items

Subjective test items (essay type)

It allows the students to determine the length and complexity of the response. This type of questions is most suitable for measuring higher level mental process skills like synthesis and evaluation.

Subjective type items can be used like:

Q1. Give examples to identify different forms of governments, which form of government is most suitable in the socio-economic and cultural context of our country? Keep your response limited to 5000-6000 words.

Some modifications can be made according to the age and grade of students to make the questions applicable to the level of the students. For example a question for students of grade V can be.

Example: What are the teachings of Islam about the respect of elders? How do these teachings help us to make a good society? Marks will be given for correct information, your own point of view.

Advantages of essay type items.

Extended response questions have following advantages over other types of question formats:

Effective for assessing higher order abilities: analyze, synthesize and evaluate.

It is comparatively less time consuming to develop such items.

Emphasizes essential communications skills

Guessing is eliminated.

Limitations of essay type items.

These advantages come with the following limitations:

1. The scoring is unreliable and time consuming.
2. Limited sampling of the content is possible.

Topic- 100: Guidelines for Writing Essay Type Items

Following guidelines are for writing essay type items when developing a test.

1. Frame questions so that the examinee‘s task is explicitly defined.
2. Specify the value and an approximate time limit for each question.
3. Do not employ optional questions.
4. Employ a larger number of questions that require relatively shorter answers rather than only a few questions that require long answers.
5. Verify a question‘s quality by writing a trial response to the question.
6. Prepare a tentative scoring key in advance of considering examinee responses.
7. Score all answers to one question before scoring the next question.
8. Make prior decisions regarding treatment of such factors as spelling and punctuation.
9. Evaluate essay responses anonymously.

Topic- 101: Scoring Rubrics for Essay Type Items

Scoring of essay items is a time consuming and difficult process. Reliability of the test demands that scoring should be consistent not only by the rater at different times, but by two independent raters as well. Scoring rubrics are descriptive scoring schemes that are developed by teachers or other evaluators to guide the analysis of writing products or processes of students’ piece of writing or any performance.

Scoring of essay time items focus on increasing the objectivity (consequently reliability) of marking. Judgments concerning the quality of a given writing sample may vary depending upon the criteria established by the individual evaluator. Writing samples are just one example of performance that may be evaluated using scoring rubrics. Scoring rubrics have also been used to evaluate group activities, extended projects and oral presentations. Writing samples are just one example of performance that may be evaluated using scoring rubrics. Scoring rubrics have also been used to evaluate group activities, extended projects and oral presentations.

Scoring Rubrics for Essay Type Items

Scoring rubrics provide at least two benefits in the evaluation process.
They support the examination to the extent to which the specified criteria has been reached.
They provide feedback to students concerning how to improve their performance.

Topic- 102: Types of Scoring Rubrics

There are two general methods for scoring subject-matter essays: the analytic (scoring key, point or rating) method and the holistic (also called global, sorting or rating) method.

Analytic Scoring Rubric

The analytic scoring requires developing a list of major elements that students are expected to include in an ideal answer of a question. Next, to decide the number of points awarded to students when they include each element. Analytic type of rubric is used to score an essay type question on different points and significantly it works best on the restricted response type essay question
In developing analytic rubric identify the certain elements of the answer which are more appropriate to the learning objectives of a course.
When assigning number to each element, be sure the total points should match with the essay‘s total value in relation to the overall number of points on test.
When a student gives a partially correct answer will get a partial credit award. But assigning partial credits can increase the inconsistency in scoring and decrease the reliability of scoring process.
Crafting a partial credit scoring may be difficult. After reading a few papers the pattern of students‘errors and misconceptions start emerging then craft a partial credit scoring rubric and use it to score al papers

Crafting a partial credit scoring may be difficult. After reading a few papers the pattern of students‘errors and misconceptions start emerging then craft a partial credit scoring rubric and use it to score al papers

	Elements of questions	Scoring key
Q1.	(a) What is meiosis and mitoses?	2 points
	(b) Compare these two.	
	Write four similarities.	4 points
	Write four differences.	4points

Factors	Sub factors	Indicators
1. Examine problems	Diagnose problem and gather information	Identify facts
	Integrate the information and	Organize ideas

	able to summarize	Define terms
1. Look for evidence	Relevance (evidence pertinent to the issue)	Identify evidence
	Consistency (supporting material consistent with each other)	Describe evidence
Draw conclusions *Make comparative judgments from data and able to adjust opinions when new facts are found and reject information that is incorrect or irrelevant.*	Analyze and interpretation of data	Combine ideas and information in new ways
	Make comparisons	Write similarities
	Correlate result	Write differences
	Look for reasons	Use relationship between phenomena
		Present solution
		Give well-reasoned conclusion
Decision making	Look for plausible alternatives to the conclusion drawn	Explore possibilities
		Formulate purposeful judgment
	Recognize and correct discrepancies	Reject the wrong answer and make correction

Topic- 103: Holistic Scoring Rubric

The holistic scoring rubric requires making a judgment about the overall quality of each student‘s response to an item. No need to mark each specific content element that a student included in the answer. Holistic rubric is probably more appropriate for extended response essay type items involving a student‘s abilities to synthesize and create and when no definite answer can be pre-specified. One way to implement holistic rubric is to decide beforehand on the number of quality categories into which you will sort the student‘s answers.

A second better way of using the holistic method is to craft a holistic rubric, which defines the qualities of paper that belong in each category. For example, defining what is in paper ―A‖ or in paper ―B‖ etc. A third refinement is to select specimen papers, which are good examples of each scoring category. Then you can compare the student‘s paper with the pre-specified specimens that define each category level. A fourth way of implementing holistic rubric is to read the answer completely and compare with another to decide which the best, the next best is and so on. This will result in the rough ranking of all the papers. This approach of holistic rubric cannot be applicable to a large number of papers.

Among these four approaches first three are consistent with a grading philosophy of criterion referenced or absolute quality standards. While the fourth one is consistent with the norm reference or relative standard grading philosophy

	Levels of Rubric
Q1. Critically evaluate the various approaches to research by Identifying sound and unsound reasoning in scientific and lay contexts.	**4. Outstanding** Discerning in judging the validity of findings as warranted or not by evidence and research design. Can articulate the basic implications of identified strengths and weaknesses of methods. **3. Effective** Can differentiate sound from flawed research methods and evaluate the validity of inferences based on available evidence. **2. Adequate** Recognizes major flaws in research. Critical judgment exercised only when pressed, elicited, or when prior (closely held) assumptions are challenged. **1. Ineffective** Unable to recognize inappropriate research methods or invalid inferences from evidence. Likely to accept results more on basis of preconceived notions, prejudice or style of presentation than on the basis of a critical assessment of the evidence, concepts, and methods.

Lesson 30

ANALYZING THE TEST-I

Topic- 104: Theories of Test Development

There are two widely perceived theories in psychosocial measurement, classical test theory (CTT) and item response theory (IRT) as both of these theories represents two different measurement frameworks.

1. Item difficulty
2. Item discrimination
3. Chance of guessing

Classical test theory is more focused on finding the true score of an individual on a test and called as true score theory

Observed Score= True score + Error

Item response theory is known as latent trait theory as its focus is to find the item characteristics as a function of ability

Topic- 105: Item Analysis: Information Provided by Item Analysis

In this phase statistical methods are used to identify any test items that are not working well.

If an item is too easy, too difficult, failing to show a difference between skilled and unskilled examinees, or even scored incorrectly, an item analysis will reveal it.

The two most common statistics reported in an item analysis are

1. Item difficulty (which is a measure of the proportion of examinees who responded to an item correctly).
2. Item discrimination (which is a measure of how well the item discriminates between examinees who are knowledgeable in the content area and those who are not).

An additional analysis that is often reported is the distractor analysis. The distractor analysis provides a measure of how well each of the incorrect options contributes to the quality of a multiple choice item. In Item Analysis, analyses are conducted for the purpose of providing information about the items, rather than the test takers.

Item analysis results can be presented graphically or numerically.

The graphic presentation consists of response curves showing the test taker‘s estimated probability of a particular response as a function of the test taker‘s score on a measure of the general type of skills or knowledge measured by the item.

The numerical presentation includes statistics that measure the difficulty of the item and the extent to which it discriminates between strong and weak test takers.

Topic- 106: Appropriate Time for Item Analysis

There are two stages at which items can be analyzed

- After administration, but before scoring
- After scores have been reported

After administration, but before scoring

Item analysis done at this stage of the process helps the test developers identify errors in the scoring key or serious defects in the items—errors or defects serious enough to exclude the item from scoring. The item analysis enables test developers to focus their attention on a relatively small subset of items, helping them to make any necessary corrections in the scoring key before the test takers‘ scores are computed.

After scores have been reported

Item analysis done at this stage of the process helps test developers select items for reuse in future forms of the tests. If the scores on a future form of the test will be linked to scores on the current form through common items. Item analysis is especially useful in selecting a set of common items that represents the full range of difficulty of the items on the test.

Lesson 31

ANALYZING THE TEST-II

Topic- 107: Test Theories in Item Analysis

1. Classical Test Theory (CTT)
2. Item Response Theory (IRT)

Classical Test Theory (CTT) and Item Analysis

Classical Test Theory (CTT) has relatively weak theoretical assumptions, which make it easy to apply in many testing situations

Relatively weak theoretical assumptions not only characterize CTT but also its extensions (e.g., generalizability theory).

CTT‘s major focus is on test-level information, item statistics (i.e., item difficulty and item discrimination) are also an important part of the CTT model.

Item Response Theory (IRT) and Item Analysis

―Item Response Theory‖ (IRT) presents a model for expressing the association between an individual's response to an item and the underlying latent variable (often called "ability" or "trait")

The latent variable, expressed as theta (θ), is a continuous one-dimensional construct that explains the covariance among item responses

People at higher levels of latent trait have a higher probability of responding correctly or endorsing an item.

Topic- 108: Item Difficulty in Classical Test Theory (CTT)

Regardless of the theoretical weakness of CTT in terms of its circular dependency of item and person statistics, practical solutions are determined within the framework of CTT for some otherwise difficult problems.

Item Difficulty in CTT

CTT does not raise a complex theoretical model to relate an examinee‘s ability to the probability of success on a particular item. CTT collectively considers a pool of examinees and empirically examines their success rate on an item. This success rate of a particular pool of examinees on an item, well known as the p value of the item, is used as the index for the item difficulty

In CTT, the item difficulty index p (p value), is the proportion of examinees correct on an item, expresses item difficulty on an Item.

Item difficulty in CTT is simply calculated by the percentage of students that correctly answered the item as refers to the p value which range from .00 to 1.00.

The values closer to 1 more easy will be the item and conversely the values near to .00 the more difficult will be the item. The values lie somewhere in the middle i.e. 0.4 to 0.6 will refer to moderate item difficulty index.

Example of Item Difficulty in CTT

The percentage of students that correctly answered the item:

The range is from 0% to 100%, or more typically written as a proportion of 0.0 to 1.00. The higher the value, the easier will be the item.

Calculation: Divide the number of students who got an item correct by the total number of students who answered it.

Ideal value: Slightly higher than midway between chance (1.00 divided by the number of choices) and a perfect score (1.00) for the item.

For example, on a four-alternative, multiple-choice item, the random guessing level is 1.00/4 = 0.25; therefore, the optimal difficulty level is .25 + (1.00 - .25) / 2 = 0.62. P-values above 0.90 are very easy items and should be carefully reviewed based on the instructor‘s purpose.
For example, if the instructor is using easy ―warm-up‖ questions or aiming for student mastery, P-values below 0.20 are very difficult items and should be reviewed for possible confusing language, removed from subsequent exams, and/or identified as an area for re-instruction.
If almost all of the students get the item wrong, the problem is either with the item or students were not able to learn the concept.
If an instructor is trying to determine the top percentage of students that learned a certain concept, this highly difficult item may be necessary.
The optimal item difficulty depends on the question-type and on the number of possible distractors. Many test experts believe that for a maximum discrimination between high and low achievers, the optimal levels (adjusting for guessing) are:
1. 2 alternatives true and false = 0.75
2. 3 alternatives multiple-choice = 0.67
3. 4 alternatives multiple-choice = 0.63
4. 5 alternatives multiple-choice = 0.60

Items with difficulties less than 30% or more than 90% definitely need attention. Such items should either be revised or replaced. An exception might be at the beginning of a test where easier items (90% or higher) may be desirable.

Topic- 109: Item Discrimination in CTT

The ability of an item to discriminate between higher ability examinees and lower ability examinees is known as item discrimination
It is the relationship between how well students did on the item and their total exam score.
The range is from –1.00 to 1.00.
The higher the value, the more discriminating the item. A highly discriminating item shows that the students who had high exams scores got the item correct students who had low exam scores got the item incorrect.
Items with discrimination values near or less than zero should be removed from the exam. This indicates that students who overall did poorly on the exam did better on that item than students who overall did well.
• Acceptable range: 0.20 or higher
• Ideal value: The closer to 1.00 the better

Item discrimination is also considered as the difference between the percentage correct for these two groups.
Item discrimination can be calculated by ranking the students according to total score and then selecting the top 27% and the lowest 27% in terms of total score
For each item, the percentage of students in the upper and lower groups answering correctly is calculated. The difference is one measure of item discrimination (IDis).
The formula is:
ID is = (Upper Group % Correct) – (Lower Group % Correct)
The following levels may be used as a guideline for acceptable items.
Negative ID is Unacceptable – check item for error
D= (UG-LG) / n.
The higher the discrimination index, the test item can discriminate better between students with higher test scores and those with lower test scores.

- D: 0.0 – 0.19 – poor item – to be Revised;
- D: 0.2 – 0.29 – acceptable;
- D: 0.3 – 0.39 – good;
- D: >0.4 – excellent

ANALYZING THE TEST-III

Topic- 110: Item Characteristic Curve (ICC) in Item Response Theory

Item Characteristic Curve (ICC) in IRT

To analyze items using IRT, the main thing need to consider is item characteristic curve (ICC). The item characteristic curve is considered as the basic building block of item response theory;

Methodological properties of an ICC

1. The difficulty which under item response theory describes the item functions along the ability scale. For example an easy item functions among the low-ability examinees and a hard item functions among the high-ability examinees
2. The second property is discrimination, which describes how well an item can differentiate between examinees having abilities below the *item location* and those having abilities above the item location
3. An item characteristic curve is the graphical representation of the probability of answering an item correctly with the level of ability on the construct being measured.

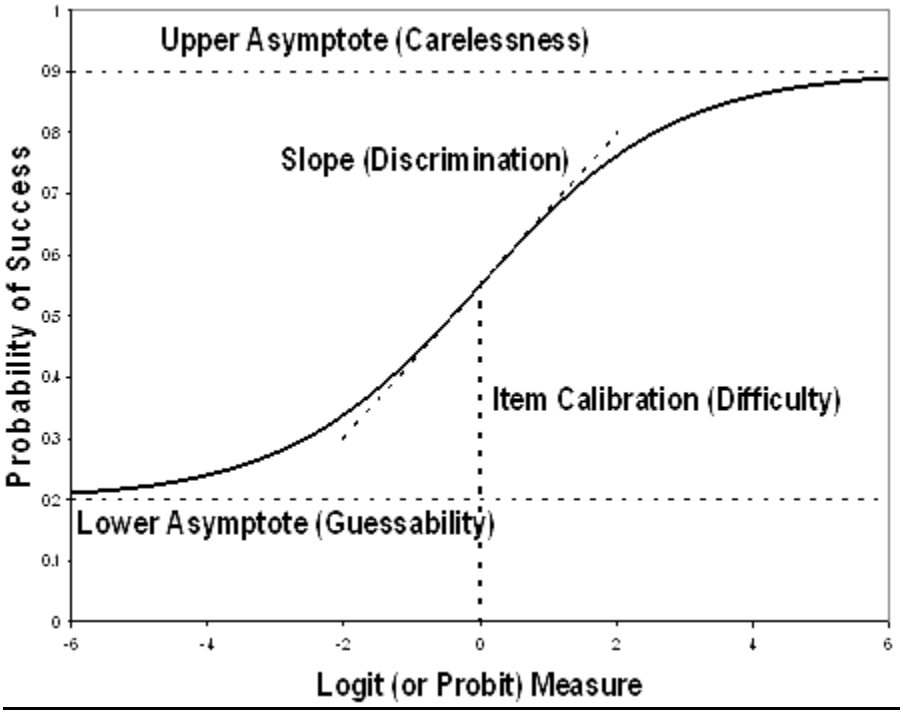
Item Characteristic Curve (ICC) in IRT

It gives a picture of:

- The item difficulty
- Discrimination power
- The probability of answering correctly by guessing

As the item difficulties are defined in relation to ability levels, on the same scale. If we know person ability then we can predict how that person is likely to perform on an item, without administering the item to the person.

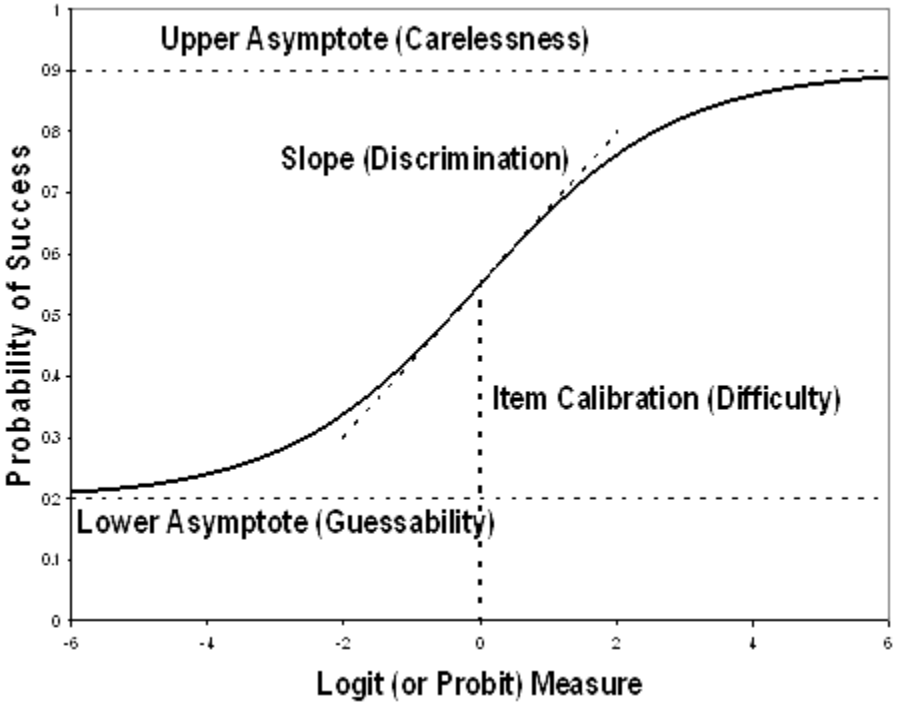
A typical item characteristic curve (ICC)



Topic- 111: Item Difficulty in Item Response Theory

Item Difficulty in Item Response Theory

The application of Item difficulty in IRT is defined as the ability at which the probability of success on the item is .5 on a logit scale, which is also known as threshold difficulty.
An item that has a high level of difficulty will be less likely to be answered correctly by an examinee with low ability than an item that has a low level of difficulty (i.e., an easy item).
A typical item characteristic curve (ICC)



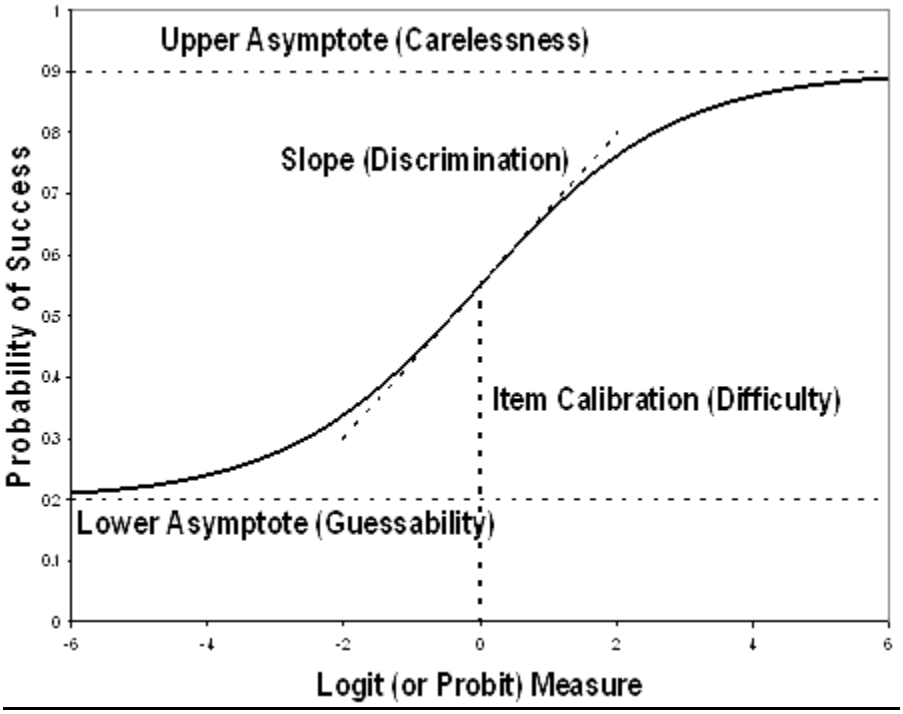
Three item characteristic curves are presented on the same graph. All have the same level of discrimination but differ with respect to difficulty.
Three item characteristic curves are presented on the same graph. All have the same level of discrimination but differ with respect to difficulty.
The left-hand curve represents an easy item because the probability of correct response is high for low-ability examinees and approaches 1 for high-ability examinees..
The center curve represents an item of medium difficulty because the probability of correct response is low at the lowest ability levels, around 0.5 in the middle of the ability scale and near 1 at the highest ability levels.
The right-hand curve represents a hard item. The probability of correct response is low for most of the ability scale and increases only when the higher ability levels are reached.
Even at the highest ability level shown (+3), the probability of correct response is only 0.8 for the most difficult item.

Topic- 112: Item Discrimination in Item Response Theory (IRT)

Item Discrimination in Item Response Theory (IRT)

The items on a test might also differ in terms of the degree to which they can differentiate individuals who have high trait levels from individuals who have low trait levels.
This property is reflected in the steepness of the item.
The steeper the curve, the better it discriminates.

A typical item characteristic curve (ICC)



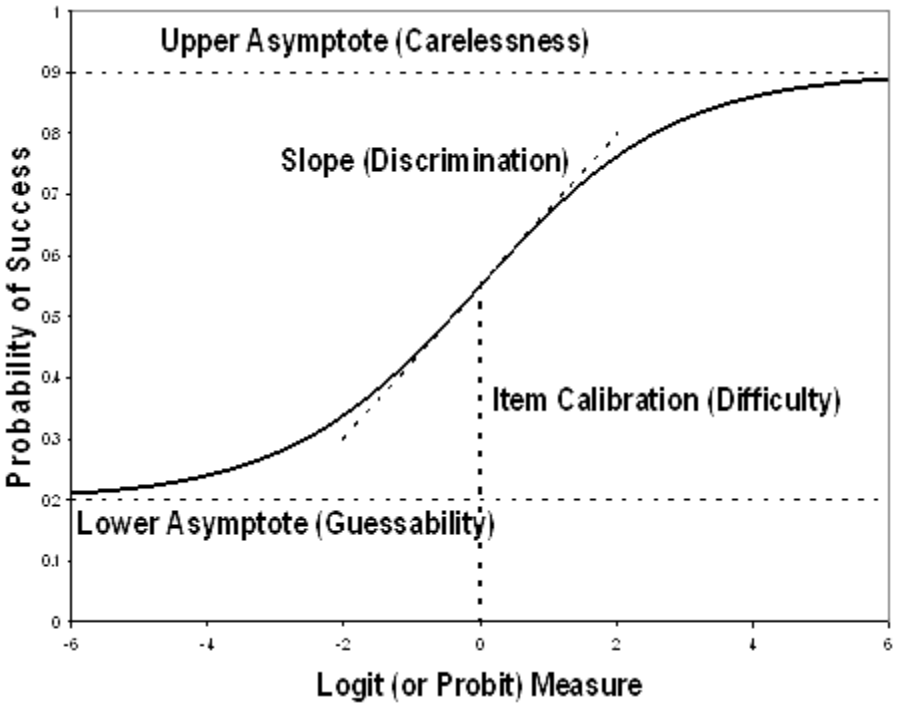
Items with steep ICC are more discriminating as compare to relatively flatter curves.
The figure contains three item characteristic curves having the same difficulty level but differing with respect to discrimination.
The upper curve has a high level of discrimination since the curve is quite steep in the middle where the probability of correct response changes very rapidly as ability increases.
The figure contains three item characteristic curves having the same difficulty level but differing with respect to discrimination.
The middle curve represents an item with a moderate level of discrimination. The slope of this curve is much less than the previous curve and the probability of correct response changes less dramatically as the ability level increases.
The figure contains three item characteristic curves having the same difficulty level but differing with respect to discrimination.
The probability of correct response is near zero for the lowest-ability examinees and near 1 for the highest ability examinees. The third curve represents an item with low discrimination.
The figure contains three item characteristic curves having the same difficulty level but differing with respect to discrimination.
The third curve represents an item with low discrimination. The curve has a very small slope and the probability of correct response changes slowly over the full range of abilities shown.

Topic- 113: Probability of Guessing in Item Response Theory (IRT)

Guessing means giving an answer or making a judgment about something without being sure of all the facts. Guessing is a standard test-taking strategy presented to examinees taking a multiple choice assessment. If test scores are based simply on the number of questions answered correctly, then a random guess increases the chance of a higher score.

In IRT this parameter of an item is also known as G (guessing) parameter which allow to detect the potential possibility of guessing in an item

A typical item characteristic curve (ICC)



Examinees guess because they do not have adequate knowledge or ability to provide correct answer
There are two types of guessing

- Blind guessing
- Informed guessing

Blind guessing:

Where an examinee chooses an answer at random from among the alternatives offered.

Informed guessing:

Where the examinee draws upon all his knowledge and abilities to choose the answer most likely to be correct.

Item writers should be conscious of guessing and not write item that could be prone to guessing.
IRT method of item analysis should be employed to eliminate those items prone to guessing

Lesson 33

ADMINISTERING TEST-

Topic- 114: Item Characteristic Curve (ICC) in Item Response Theory

We have spent a lot of time in learning writing instructional objectives, table of specification, selection of questions format, writing questions, aligning questions and objectives to ensure GOOD TEST. But imagine after all this hard work, if care is not taken while assembling all good work to a test actually presented to students or test administration is not done the way it should be then whole hard work will be wasted.

Issues in test administration

1. Cheating
2. Poor testing conditions
3. Test anxiety
4. Errors in test scoring procedure.

What to do?

It is equally important to control all he factors other than test itself, to collect trustable evidence of student learning by addressing test administration issues.

The way in which the test is administered is very important to meet the goal of producing highly valid, reliable results.

Once a test is ready then next step is to administer it.

Teacher has to help the students psychologically by maintaining a positive test-taking attitude, clarifying the rules; penalties for cheating, reminding them to check their copies, minimizing distractions and giving time warnings.

Cheating, poor testing conditions, and test anxiety, as well as errors in test scoring procedures contribute to invalid test results.

Accurate achievement data are very important for planning curriculum and instruction.

Test scores that overestimate or underestimate students‘ actual knowledge and skills cannot serve these important purposes.

So it is worth a little more time to properly assemble and administer a test.

Topic- 115: Assembling Test: Recording Items

The preparation of test items is greatly facilitated if the items are properly recorded.

Recording test items

The card contain information concerning the instructional objectives, specific objective, difficulty index, discrimination index and the content measured by the item should be prepared for each item to maintain its record.

Topic- 116: Assembling Test: Packing the Test –I

1. Packing the test

Once you have measureable instructional objectives, test blueprint, written test items matching instructional objectives, so you are ready to

- package the test and
- reproduce the test

Assembling Test (Packaging the Test)

Packing of test involves:

1.Grouping together items of similar format
2. Arranging test items from easy to hard
3. Properly spacing items

4. Keeping items and options on the same page
5. Position illustration near description (if diagram given)
6. Randomness of answer key
7. Determine how students record answers (separate sheet or same sheet)
8. Providing space for test taker‘s details
9. Proofread the test (typographical and grammatical error)
10. Test directions

1. Grouping together items of similar format
Makes easy to understand
Save response time
Once set of instruction per format is enough

2. Arranging test items from easy to hard
Increases the possibility of good start for poor starters
Build confidence
Motivates
Reduces test anxiety

Topic- 117: Assembling Test: Packing the Test –II

1. Packing the test
Following are some more points.
Space the item for easy reading
Enough space between lines and questions
Suitable answer space
Standard font type and size

4. Keeping items and options on the same page
Minimizes the likelihood of misprint
Saves respondent from unnecessary hassle
Save test time.

5. Position illustration near description (if diagram given)
Put all related questions on same page
Diagram above the question

Lesson 34

ANALYZING THE TEST-II

Topic- 118: Assembling Test: Packing the Test –III

1. Packing the test
Following are some more points.
6. Randomness of answer key
Check key, equal distribution of correct answer, use all options equally
Avoid guessing
Avoid confusion
7. Determine how students record answers (separate sheet or same sheet)
Brings uniformity attempt with clarity
Prepares for standardized tests
8. Providing space for test taker‘s details
Identification ensured
Makes easy to combine different parts of test (if applicable)

Topic- 119: Assembling Test: Packing the Test –IV

1. Packing the test
Following are some more points.
9. Proofread the test (typographical and grammatical error)
Save time during test
Avoid clues/confusion
Increase confidence in test
10. Test directions
Conveys expectation
Help in time management
Conveys scoring policy and priorities

Topic- 120: Assembling Test: Reproducing the Test

In schools it is usually by photocopying and quality of test copies may vary considerably.

Reproducing the Test

Reproduction of test involves:
1. Knowing the photocopying machine
2. Specifying copying instructions
3. Filing original test
1. Knowing the photocopying machine
Expert operator
Toner quality
Atomization facility in machine
Ensure uniformity of legibility, shades etc.
2. Specifying copying instructions
Randomly checking of every 19th copy
Paper size
Margins
Ordering of pages
Stapling

Packing
3. Filing original test
Master copy for reference and reuse
Use as reference in random checking
Keep with you during test

Lesson 35

ANALYZING THE TEST-III

Topic- 121: Assembling Test: Things to Remember When Administering Test –I

Things to remember when administering test

Test is ready, get students ready.

1. Maintain a positive attitude
2. Maximize achievement motivation
3. Equalize advantages
4. Avoid surprises
5. Clarity the rules
6. Rotate distribution
7. Remind students to check their copies
8. Monitor students
9. Minimize distributions
10. Give time warning
11. Collect test uniformly

Maintain a positive attitude

1. Assure test is from taught
2. Students convenience and limitations addressed
3. Guide to reduce anxiety
4. Eliminate non-actors affecting achievement

Maximize achievement motivation

1. Encourage to do best, mitigate the fear
2. Highlight value of giving best
3. Reduce panic
4. Encourage serious thinking

Topic- 122: Assembling Test: Things to Remember When Administering Test –II

1. Things to remember when administering test

Equalize advantages (test-wise)

1. Discourage guessing
2. Discourage no answer
3. Discourage multiple answers
4. Instructions like, don‘t spend more time on difficult items, do easy questions first, must read after completing, etc.

Avoid surprises

Advance notice of tests; discuss test structure, and ways of preparing for test

1. Brings out stable achievement
2. What students can perform the best?

Topic- 123: Assembling Test: Things to Remember When Administering Test –III

1. Things to remember when administering test

Clarity the rules

1. Time limits
2. Restroom policy

3. Special requirement of students
4. Test distribution
5. Consulting others
6. Borrowing things

Rotate distribution

1. Left to right, right to left
2. Front to back, back to front
3. Multiple person distribution
4. Starting at same time

Remind students to check their copies

1. Order of pages
2. Number of pages
3. Quality of print
4. Replacement if needed before start of test
5. Ensuring recording of name and date

Topic- 124: Assembling Test: Things to Remember When Administering Test -IV

In this session student will learn:

1. Things to remember when administering test

In this session student will learn:
1. Things to remember when administering test

Monitor students (test invigilation)

1. Penalties of cheating
2. Disturbing others
3. Seating position/posture
4. Protecting test from others
5. Cheating material handing
6. Jurisdiction of invigilation staff

Minimize distributions

1. Avoid noise
2. Avoid instruction after start of test
3. Avoid in-out movement
4. No talk between invigilation staff- silence

Time warning and test collection

1. Half time
2. 2/3 time
3. 30, 15 and 5 min reminder
4. Use minimum words to announce time warning
5. Give time to close
6. Announce ―stop writing‖
7. Collecting test in reverse order of distribution

Lesson 36

SCORING TEST-I

Topic- 125: Scoring Criteria: Scoring Rubric

Scoring Criteria

Planning how responses will be scored leads to rethinking and clarifying the questions so that students have a clearer idea of what is expected.

Clear specification of scoring criteria in advance of administering essay questions can contribute to improved reliability and validity of the assessment.

Scoring Rubric

A rubric is an explicit set of criteria used for assessing a particular type of work or performance and provides more details than a single grade or mark. Scores levels identified in a scoring rubric must be descriptive, not merely judgmental in nature.

Example: Define the level of rubric as ―Writing is clear and thoughts are complete‖ as compared to ―excellent‖.

Topic- 126: Scoring Rubric for Essay Type Questions

Scoring Rubric for Essay Type Questions

Scoring of essay items is a time consuming and difficult process. Reliability of the test demands that scoring should be consistent not only by the rater at different items but by two independent raters as well.

Judgments concerning the quality of a given writing sample may vary depending upon the criteria established by the individual evaluator.

Writing samples are just one example of performance that may be evaluated using scoring rubric. Scoring rubric has also been used to evaluate group activities, extended projects and oral presentations.

Basic Steps to design Rubric

1. Identify a learning goal.
2. Choose outcomes that may be measured.
3. Develop or adapt existing rubric.
4. Share it with students.

Topic- 127: Elements of Rubric

Elements of Rubric

A rubric includes:

1. Score
2. Criteria
3. Level of performance
4. Descriptors

Score

A system of number or values assigned to a work often combines with a level of performance.

High numbers are for best performance like 4, 5 or 6 whereas down to 1 or 0 are the lowest score in a performance assessment.

Criteria

It tells us that which feature, trait or dimension to be measured and include a definition and example to make clear the meaning of each trait to be assessed.

Level of Performance

There are adjectives to describe the performance levels. These levels tell students what they are expected to do.
Descriptors can be used with levels of performance to achieve objectivity but they can be used without them as well.

Descriptors

These are details for each level of performance to reliable and unbiased scoring.

Topic- 128: Holistic Scoring Rubric

Holistic Scoring Rubric

These rubrics are good for evaluating overall performance on a task. All criteria are assessed as a single score. As only one score is given so holistic rubrics are easy to score.

When to use:

1. There is no correct answer/response to ask a task e.g. creative work.
2. Focus is on overall quality, proficiency or understanding of a specific content or skill.
3. The assessment is summative e.g. at the end of the semester or major.
4. Assessing significant numbers e.g. 150 student portfolio.

Holistic Rubric for Essay Type Questions

Holistic rubric is probably more appropriate for extended response essay type items involving a student‘s abilities to synthesize and create and when no definite answer can be pre-specified.

SCORING TEST-II

Topic- 129: Approaches to Apply Holistic Scoring Rubric

One way to implement holistic rubric is to decide beforehand on the number of quality categories into which you will sort the student‘s answers. A second better way of using the holistic method is to craft a holistic rubric, which defines the qualities of paper that belong in each category. For example, defining what an ―A‖ paper is a ―B‖ paper is and so on.

A third refinement is to select specimen papers, which are good examples of each scoring category. Then you can compare the student‘s paper with the pre-specified specimens that define each category level.

A fourth way of implementing holistic rubric is to read the answer completely and one with an-other to decide which are the best. The next best and so on.

This will result in the rough ranking of all the papers this approach of holistic rubric cannot be applicable to a large number of papers.

Among these four approaches first three are consistent with a grading philosophy of criterion referenced or absolute quality standards

While the fourth one is consistent with the norm reference or relative standard grading philosophy

Holistic Scoring Rubric

	Levels of Rubric
Q1. Critically evaluate the various approaches to research by Identifying sound and unsound reasoning in scientific and lay contexts.	**4. Outstanding** Discerning in judging the validity of findings as warranted or not by evidence and research design. Can articulate the basic implications of identified strengths and weaknesses of methods. **3. Effective** Can differentiate sound from flawed research methods and evaluate the validity of inferences based on available evidence. **2. Adequate** Recognizes major flaws in research. Critical judgment exercised only when pressed, elicited, or when prior (closely held) assumptions are challenged. **1. Ineffective** Unable to recognize inappropriate research methods or invalid inferences from evidence. Likely to accept results more on basis of preconceived notions, prejudice or style of presentation than on the basis of a critical assessment of the evidence, concepts, and methods.

Score	Description
5	Demonstrates complete understanding of the problem. All requirements of task are included in response.
4	Demonstrates considerable understanding of the problem. All requirements of task are included.
3	Demonstrates partial understanding of the problem. Most requirements of task are included.
2	Demonstrates little understanding of the problem. Many requirements of task are missing.
1	Demonstrates no understanding of the problem.

0	No response/task not attempted.

Advantages
1. Quick scoring
2. Provides overview of student achievement

Disadvantages
1. Does not provide detailed information.
2. May be difficult to provide one overall score.

Topic- 130: Analytic Scoring Rubric

Analytic Scoring Rubric

Each criterion is assessed separately by using different descriptive ratings. Each criterion is given a separate score.

Final score is made up of adding each component parts. It takes more time to score but gives detailed more feedback.

When to use:
1. Several faculties are collectively assessing student work.
2. Outside audience will be examining rubric scores.
3. Profiles of specific strength/weakness are desired.

Analytic Rubric for Essay Type Questions

Analytic scoring requires a list of major elements that students are expected to include in an ideal answer of a question.

Next, to decide the number of points award to students when they include each element.

Analytic rubric is used to score an essay type of question on different points and significantly it works best on the restricted response type essay question.

In developing analytic rubric identify the certain elements of the answer which are more appropriate to the learning objectives of a course.

When assigning number to each element, be sure the total points should match with the essay‘s total value in relation to the overall number of points on test.

Student will get a partial credit award gives on a partially correct answer but partial credits can increase the inconsistency in scoring and decrease the reliability of scoring process.

Crafting a partial credit scoring may be difficult after reading a few papers. The pattern of students‘ errors and misconceptions are emerged then craft a partial credit scoring rubric and use it to score all papers.

	Elements of questions	Scoring key
Q1.	(a) What is meiosis and mitoses?	2 points
	(b) Compare these two.	
	Write four similarities.	4 points
	Write four differences.	4points

Criterion	**Beginning 1**	**Developing 2**	**Accomplished 3**	**Exemplary 4**	**Score**
#1	Description reflecting	Description reflecting	Description reflecting	Description reflecting	

	beginning level of	movement towards mastery level of	achievement of mastery level	highest level of	
#2	**performance**	**performance**	**performance**	**performance**	
	Description reflecting beginning level of performance	**Description reflecting movement towards mastery level of performance**	**Description reflecting achievement of mastery level of performance**	**Description reflecting highest level of performance**	
#3	**Description reflecting beginning level of performance**	**Description reflecting movement towards mastery level of performance**	**Description reflecting achievement of mastery level of performance**	**Description reflecting highest level of performance**	

Topic- 131: Advantages and Disadvantages of Analytic Rubric

Advantages

1. More detailed feedback.
2. Scoring more consistent across students and grades.

Disadvantages

1. Time consuming to score.

Suggestions for scoring constructed response questions

1.Prepare an outline of the expected answer in advance.
2. Use the scoring rubric that is most appropriate.
3. Decide how to handle factors that are irrelevant to the learning outcomes being measured.
4. Evaluate all responses to one question before going on to the next one.
5. When possible, evaluate the answers without looking at the student‘s name.
6. If especially important decisions are to be based on the results, obtain two or more independent ratings.

Lesson 38

STANDARDIZED TEST-I

Topic- 132: Standardized Achievement Test

A standardized achievement test a fixed set if items to measure defined achievement domain, specific direction for administrating and scoring test, and norms based on representative groups of individuals.

Most publishes achievement test are called standardized achievement tests. These typically are norm referenced tests. Quite a few criterion referenced achievement tests are also been published.
Achievement tests are used as part if broader assessment system or alone. They provide relatively inexpensive means of measuring broad achievement goals.
Standardized achievement tests are often customized to include characteristics of both norm and criterion referenced test.
Standard content and procedure makes it possible to give an identical test to individuals in different places at different times.
Equivalent forms are included in many standardized tests, which make it possible to repeat the test without fear that the test takers will remember the answers from first testing.

Topic- 133: Characteristics of Standardized Achievement Test

Characteristics of Standardized Achievement Test

1. The test items are of highly technical quality. They have been developed by educational and test specialists, pretested and selected on the basis of difficulty discriminating power, and relationship to a clearly defined and rigid set of specifications.
2. Directions for administering and scoring are so precisely stated that the procedures are standard for different users of test.
3. Norms based on national samples of students in the grade where the test is intended for use are provided as aids in interpreting the test scores.
4. Equivalent and comparable forms of the test are usually provided, information concerning the degree to which the forms are comparable.
5. A test manual is used as guides for administering the test, evaluating its technical qualities, and interpreting and using the results.

Topic- 134: Standardized Test Versus Informal Classroom Test

Standardized tests and carefully constructed test are common in many ways. The main differences between the two types are:

1. The nature of the learning outcomes and the content measured.
2. Quality of test items
3. Reliability of the tests
4. Procedure for administering and scoring
5. Interpretation of scores

The standardized tests inflexibility makes it less valuable for those purposes for which the informal classroom tests is so admirably suited.

1. Evaluating the learning outcomes and content unique to particular class or school
2. Evaluating student‘s day to day progress and their achievement on the work units or varying sizes.
3. Evaluating knowledge of current developments in the rapidly changing content areas such as science and social studies.

STANDARDIZED TEST-II

Topic- 135: Standardized Test Batteries and Guidelines for SAT Batteries

Standardized test batteries

Standardized achievement tests are frequently used in the form of survey test batteries. A battery consists of a series of individual tests all standardized on same national sample of students. Test batteries include subjects according to educational level of students.

Guidelines for SAT batteries

Achievement test batteries focus on the basic skills measuring important outcomes of the program. Content oriented test in basic achievement batteries have broad coverage but limited sampling in each content area and may tend to become outdated more quickly.

The selection of battery should be based on its relevance to the schools objectives. Diagnostic batteries should contain a sufficient number of test items for each type of interpretation to be made.

Topic- 136: SAT in Specific Area, Separate Content Oriented Test, Reading Test

SAT in specific area

There are separate tests designed to measure achievement in specific areas. This includes tests of course content, reading tests.

Separate content oriented tests

Attention should be directed on appropriateness for particular course in which it is to be used. Standardized tests of specific knowledge are seldom as relevant and useful as well constructed teacher made test in same area.

Reading test

Such tests commonly measure

1. Vocabulary
2. Reading comprehension
3. Rate of reading

No two reading tests are exactly alike. They differ in the material that the reader is expected to comprehend, in the specific reading skills tested and in the adequacy with which each skill is measured.

Reading survey test measure only some of the outcomes of reading instruction, the mechanics of reading is measured by diagnostic reading tests.

In addition to matching the objectives of instruction, test selection should also take into account all the possible uses to be made of the results.

Topic- 137: Concept of Interpreting Test Scores

Concept of Interpreting Test Scores

Test scores can be interpreted in the terms of:

1. Types of tasks that can be performed (Criterion referenced or standard based)
2. Relative position held in some reference group (norm reference)

Interpreting test scores

The properties of physical measuring scales lacks in educational measurement. A student who receives a score of zero does not have zero knowledge of that subject.

A true zero point in achievement cannot be usually established.

60 correct items on a simple vocabulary test does not have the same meaning as 60 items correct on more difficult one or any other subject or study skills. However, this arbitrary starting point prevents

us from claiming that a zero indicates no achievement at all or 100 represent twice the achievement of a score of 50.

Lesson 40

STANDARDIZED TEST-III

Topic- 138: Method of Interpreting Test Scores

Raw scores

A raw score is a numerical summary of a student‘s test performance, it is not very meaningful without further information.

1. A raw score is simply the number of points received on a test.
2. ―0‖ in raw scores does not mean absence of trait.

If a student in our class answered 35 items correctly on arithmetic test, and therefore has a raw score of 35. This statement creates a lot of questions in our mind.

1. What is 35 mean?
2. Is that a good score?
3. How many items were there on the test?
4. What kind of arithmetic problems were presented?
5. How difficult was test?
6. What is student‘s position in his class?

Answer to these or similar questions are needed to make any raw score meaningful.

Raw score can be converted into:

1. A description of specific tasks that the student can perform (Criterion reference interpretation)
2. Some type of derived scores to indicate the student‘s relative position in a clearly defined reference group.

Topic- 139: Criterion-Referenced Interpretation

Criterion-referenced Interpretation

In case of standardized test, interpretation can only be made with reference to the constructions on which test was based by the developer.

This is primarily useful in mastery testing where a clearly defined and delimited domain of learning tasks can be most readily obtained.

Such interpretations must be made with caution because these tests were typically designed to discriminate among individuals rather than describe the specific tasks they can perform.

Criterion-referenced interpretations of test results are most meaningful when the test has been specifically designed for this purpose e.g. designing a test that measures a set of clearly stated learning tasks.

Topic- 140: Guidelines for Criterion-Referenced Interpretation

Guidelines for Criterion-referenced Interpretation

Are the achievement domains (objective or content clusters) homogenous, delimited, and clearly specified? If not, avoid specific descriptive statements.

Are there enough items for each type of interpretation? If not, make tentative judgment and or combine items into larger content clusters for interpretation.

In constructing the test, were the easy items omitted to increase the discrimination among individual? If so, remember that descriptions of what low achievers can do will be severely limited.

Does the test use selection type items only? If so, keep in mind that a proportion of correct answers may be based on guessing.

Do the test items provide a directly relevant measure of the objectives? If not, base the interpretation on hat the items actually measured.
Example: ―ability to identify misspelled words‖ rather than ―ability to spell‖. They are related but not the same process.

Topic- 141: Norm-Referenced Interpretation

Norm-referenced Interpretation

This interpretation tells us how an individual compares with other persons who have taken the same test e.g. ranking of scores from highest to lowest and to note where an individual's score falls. Standardized tests typically have been designed for norm-referenced interpretations which involves converting the raw scores to derived scores by means of table of norms.

Derived Scores

A derived score is a numerical report of test performance on a score scale that has well-defined characteristics and yields normative meaning.

Criteria Most Desired in Norms

Test norms should be
1. Normal.
2. Representative.
3. Up to date.
4. Comparable.
5. Adequately described

Cautions in Interpreting Test Scores

A test score should be interpreted
1. In terms of the specific test form which it was derived.
2. In light of all of the student's relevant characteristics.
3. According to the type of decision to be made.
4. As a band of scores rather than a specific value.
5.A test score must be verified by supplementary evidence.

Lesson 41

HIGH STAKE TESTING AND ISSUES-I

Topic- 142: High Stake Testing

The use of test and assessment to make decisions that are of prominent educational, financial, or social impact.

Decision based on high stake testing

1. Promotion to next grade
2. Awarding of diploma or degree
3. Evaluation of school performance
4. Incentives and accountability of school staff

Topic- 143: High Stake Testing in Pakistan

High Stake Testing in Pakistan

In school education in Pakistan, high stake testing is conducted at each level from primary to higher secondary.

1. Grade 5 and 8 examination by Punjab Examination Commission.
2. Grade 10 and 12 examination by BISEs

Test construction in High Stake testing

1. The format of items are same as used in CRT in classroom
2. Item of developed by professional employed in dedicated organization
3. Item banks are developed
4. Psychometric properties of item are tested
5. The process goes on round the year

Criticism on high stake testing

High stake tests have same issues as classroom test but much larger impact and consequences.

Topic- 144: Recommendations for Effective HST

1. Recommendations for effective High Stake Testing

Recommendations for effective High Stake Testing

- Protection against high stake decisions based on single test
- Adequate resources and opportunity to learn
- Validation for each intended separate use
- Full disclosure of likely consequences
- Alignment between test and curriculum
- Validity of passing scores and achievement levels
- Appropriate attention towards language difference between examinees
- Appropriate attention towards examinees with disabilities
- Careful adherence to explicit rules for determining which students are to be tested
- Ongoing evaluation for intended and unintended effects of high stake testing

Lesson 42

HIGH STAKE TESTING AND ISSUES-II

Topic- 145: Preparation for Effective HST

Preparation for effective High Stake Testing

1. Focus on task, not on your feelings towards it
2. Inform presents and students about importance of the test.
3.Teach test taking skills as part of regular instruction
4. As the test day approaches respond to students questions openly and directly
5. Take advantage of whatever preparation material is available

Topic- 146: Institutions Involve in Assessment-I

Institutions involve in assessment

1. Examination commission
2. Board of intermediate and secondary education (BISEs)
3. Board of Technical Education

Scope of Examination Bodies

1. Examination commission conduct examination at grade 5 and 8 level
2. BISEs hold SSC and HSSC annual examinations
3. Boards of Technical Education conduct examination of various diplomas and certificates etc.

Inter board committee of chairman (IBCC) is a forum to discuss matters relating to development and promotion of intermediate and secondary education and technical education in Pakistan.
For high stake examination at primary and elementary level Baluchistan and Punjab has established examination commissions
National Education Assessment System (NEAS) provides a countrywide picture of situation of education and report to federal policy makers.

Topic- 147: Institutions Involve in Assessment-II

1. Institutions involve in assessment

Following institutions are involved in assessment.

1. National Education Assessment System (NEAS)
2. Provincial Education Assessment Centre (PEAS) in KPK and Sindh
3. Examination Commission (Punjab and Baluchistan)

National Education Assessment system (NEAS) works for promoting quality learning among children of Pakistan by carrying out fair and valid national assessment.
Objectives of NEAS

1. Informing policy
2. Monitoring standards
3. Identifying correlation of achievement
4. Directing teachers‘ efforts and raising students‘ achievement

The areas centers of NEAS were established in all provinces and areas. They are still working except in Punjab which was merged in PEC.

Topic- 148: National Education Assessment System

1. National Education Assessment system

National Education Assessment system

The National Education Assessment System has been institutionalized in Pakistan at national level with the cooperation of provincial and area Assessment Centers.

NEAS was established as five years development project with the financial assistance of the World Bank and Development for International Development (DfID) in year 2003. NEAS is subordinate office under the ministry of Federal Education & Professional Training.

Objectives of NEAS

1. Informing policy: the extent to which geography and gender are linked to inequality in student performance.
2. Monitoring standards: How well the curricula are translated into knowledge and kills
3. Identifying correlation of achievement: the principal determinants of student performance
4. Directing teachers' efforts and raising students'achievement: assisting teachers to use data to improve student performance.

Working of NEAS

Every year NEAS conduct large scale assessment at primary and elementary level. The content areas it usually covers are reading and writing of language, mathematics and science.

Completion of four cycles of Assessment on a large sale i.e. in 2005, 2006, 2007 and 2008. The assessment results of cycle in 2016 are about come.