# EDUCATIONAL ASSESSMENT AND EVALUATION

## B.Ed (1.5 Year)

**Units 1-9** **Code 8602**

**Credit Hours: 3**

**Department of Early Childhood Education and Elementary Teacher Education**
**Faculty of Education**
**Allama Iqbal Open University Islamabad**

| | |
|---|---|
| Edition..... | 1st |
| Year of Printing..... | 2016 |
| Quantity..... | 10000 |
| Composing, Layout..... | Mushtaq Hussain |
| Printer..... | AIOU-Printing Press, H-8, Islamabad. |
| Publisher ..... | Allama Iqbal Open University, Islamabad |

# COURSE TEAM

| | |
|---|---|
| **Chairperson:** | Prof. Dr. Nasir Mahmood |
| **Course Development Coordinator:** | Dr. Muhammad Tanveer Afzal |
| **Writers:** | Prof. Dr. Rehana Masrur<br>Dr. Naveed Sultana<br>Dr. Muhammad Tanveer Afzal<br>Dr. Muhammad Saeed<br>Muhammad Azeem<br>Muhammad Idrees |
| **Reviewers:** | Prof. Dr. Rehana Masrur<br>Dr. Naveed Sultana<br>Dr. Muhammad Tanveer Afzal |
| **Editor:**<br>**Course Coordinator:** | Dr. Muhammad Tanveer Afzal |
| **Composing:** | Mushtaq Hussain |

# CONTENTS

# FOREWORD

Learning is natural to the human beings, but in order to catalyze the process of learning the efforts of teachers contribute a lot towards educational attainments. The answer to the questions that to what extent the students have learned and which instructional techniques work better is not simple. These questions are vital to answer and answers need rigorous approach towards the measurement and assessment of the students' progress that consequently leads towards the better decision making. In order to ensure and enhance the effectiveness of teaching-learning process teachers need to get information regarding students' performance. Based upon this information teachers make critical instructional decisions for example whether to use a certain teaching method or not, whether the progress of students towards attainment of educational goals is satisfactory or not etc.

There is no exaggeration to say that classroom assessment is an integral and indispensible part of the teaching leaning process. Assessment provides comprehensive and objective information through which not only the learning of an individual student is recognized and responded but also through this information the overall effectiveness of an education program can be judged. Therefore, for a teacher it is highly significant to understand the concepts of measurement, assessment and evaluation as for as their role in instruction. He/she must also be able to plan and conduct procedures in an effective way and to interpret and use the information obtained through these procedures to maximize the effectiveness of teaching learning process.

For the optimization of the students learning it is mandatory that teachers can develop, administer, score and report the tests scores to the educational stakeholders, the validity and reliability of the classroom test developed by the teachers for the use in classroom can only be enhanced by exposing them to the process and procedures of test development. The experience towards the

measurement and development of the test may contribute towards the professional development of prospective and in-service teachers.

The development of this course intends towards the professional development of the prospective teachers in assessment and evaluation of the students. The knowledge and skills gained during the course may help them while practicing in the classroom and also help to develop more positive attitude towards assessment. In the end, I am happy to extend my gratitude to the course team, Course Development Coordinator, Unit Writers and Reviewers for the development of this course book despite of the time constraint. Any suggestions for the improvement of this course will be warmly welcomed.

Vice-Chancellor
AIOU

# PREFACE

Classroom tests play a central role in the assessment of student learning. Teachers use tests to assess the progress of the students learning. Tests provide relevant measures of many important learning outcomes and indirect evidence concerning others. They make expected learning outcomes explicit to students and parents and show what types of performance are valued. In order to ensure and enhance the effectiveness of teaching-learning process teachers need to get information regarding students' performance. Based upon this information teachers make critical instructional decisions for example whether to use a certain teaching method or not, whether the progress of students towards attainment of educational goals is satisfactory or not, what if a student is having learning deficiency, how to motivate a student etc. Classroom assessment primarily aims to yield the information regarding students' performance in order to help the teacher and/or stakeholders to determine a certain degree, to which a learner has acquired particular knowledge, has understood particular concepts or has mastered certain skill.

The competency of the teachers to develop, administer, score and interpret the results is the prime consideration of the tomorrow's classrooms. Therefore, it is necessary to enhance the knowledge and skills of the prospective teachers towards the development and use of assessment tools. This particular course comprised of nine units. The concept of measurement, assessment and evaluation is elaborated in the first unit, the test items are developed in-line with the objectives/learning outcomes, so objectives are discussed in unit two. The third and fourth units of the textbook are about different types of tests and techniques used by the teachers. The characteristics of assessment tools such as validity and reliability are

explained in the sixth and seventh units. The 8th and 9th units of textbook are about the interpretation and reporting of the test scores. The text includes relevant examples for the elaboration of the concepts and the activities are placed for the hands on works, which consequently, help to develop the attitude and the skills of the prospective teachers.

In the end, I am thankful to the course team and especially the course development coordinator for this wonderful effort.

**Dr. Naveed Sultana**
**Chairperson**
Department of Secondary Teacher Education

**Dr. Muhammad Tanveer Afzal**
Course Development Coordinator

# COURSE OBJECTIVES

Classrooms are busy places. Every day in every classroom, teachers make decisions about their pupils, the success of their instruction and perform a number of other tasks. Teachers continually observe, monitor, and review learners' performance to obtain evidence for decision. Evidence gathering and classroom marking are necessary and ongoing aspects of teaches' lives in classroom. And decisions based on this evidence serve to establish, organize, and monitor classroom qualities such as pupil learning, interpersonal relations, social adjustment, instructional content and classroom climate. Keeping in view the tasks teachers have to perform in classroom, this course has been organized to follow the natural progression of teacher' decision making form organizing the classroom as a social setting, to planning and conducting instruction to the formal assessment of pupil learning, to grading and finally to communicating results to an ongoing part of teaching therefore this course covers the broad range of assessments. The course intends to achieve the following objectives.

## COURSE OBJECTIVES

After studying this course the prospective teachers will be able to:

1. Understand the concepts and application of classroom assessment.
2. Integrate objectives with evaluation and measurement.
3. Acquire skills of assessing the learning outcomes.
4. Interpret test scores.
5. Know about the trends and techniques of classroom assessment.

**UNIT–1**

# MEASUREMENT, ASSESSMENT AND EVALUATION

**Written By:**
**Prof. Dr. Rehana Masrur**

**Reviewed By:**
**Dr. Naveed Sultana**

## CONTENTS

# INTRODUCTION

In order to ensure and enhance the effectiveness of teaching- learning process teachers need to get information regarding students' performance. Based upon this information, teachers make critical instructional decisions for example whether to use a certain teaching method or not, whether the progress of students towards attainment of educational goals is satisfactory or not, what if a student is having learning deficiency, How to motivate a student etc. Measurement, testing, assessment and evaluation primarily aims to yield the information regarding students' performance in order to help the teacher and/or stakeholders to determine a certain degree, to which a learner has acquired particular knowledge, has understood particular concepts or has mastered certain skill. This information is used to scaffold the next step in the learning process.

There is no exaggeration to say that measurement, assessment and evaluation collectively form an integral and indispensible part of the teaching leaning process. Measurement, assessment and evaluation provides comprehensive and objective information through which not only the learning of an individual student is recognized and responded but also through this information the overall effectiveness of an education program can be judged, maintained and/or enhanced.

Therefore for a teacher, it is highly significant to understand the concepts of measurement, assessment and evaluation as for as their role in instruction. He/she must also be able to plan and conduct procedures in an effective way and to interpret and use the information obtained through these procedures to maximize the effectiveness of teaching learning process.

# OBJECTIVES

After studying this unit, the prospective teacher will be able to:

- indicate the primary differences among the terms measurement, assessment and evaluation
- explain the types of assessment used in the classroom milieu
- compare and contrast the assessment *for* learning and assessment *of* learning
- summarize the need for assessment
- highlight the role of assessment in effective teaching-learning process
- describe major characteristics of classroom assessment
- identify the core principles of effective assessment

## 1.1 Concept of Measurement, Assessment and Evaluation

Despite their significant role in education the terms measurement, assessment, and evaluation are usually confused with each other. Mostly people use these terms interchangeably and feel it very difficult to explain the differences among them. Each of these terms has a specific meaning sharply distinguished from the others.

**Measurement:** In general, the term measurement is used to determine the attributes or dimensions of object. For example, we measure an object to know how big, tall or heavy it is. In educational perspective measurement refers to the process of obtaining a numerical description of a student's progress towards a pre-determined goal. This process provides the information regarding how much a student has learnt. Measurement provides quantitative description of the students' performance for example Rafaih solved 23 arithmetic problems out of 40. But it does not include the qualitative aspect for example, Rafaih's work was neat.

**Testing:** A test is an instrument or a systematic procedure to measure a particular characteristic. For example, a test of mathematics will measure the level of the learners' knowledge of this particular subject or field.


**Assessment:** Kizlik (2011) defines assessment as a process by which information is obtained relative to some known objective or goal. Assessment is a broad term that includes testing. For example, a teacher may assess the knowledge of English language through a test and assesses the language proficiency of the students through any other instrument for example oral quiz or presentation. Based upon this view, we can say that every test is assessment but every assessment is not the test.

The term 'assessment' is derived from the Latin word 'assidere' which means 'to sit beside'. In contrast to testing, the tone of the term assessment is non-threatening indicating a partnership based on mutual trust and understanding. This emphasizes that there should be a positive rather than a negative association between assessment and the process of teaching and learning in schools. In the broadest sense assessment is concerned with children's progress and achievement.

In a comprehensive and specific way, classroom assessment may be defined as:

> *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*. (NCCA, 2004)

In short, we can say that assessment entails much more than testing. It is an ongoing process that includes many formal and informal activities designed to monitor and improve teaching and learning.

**Evaluation:** According to Kizlik (2011) evaluation is most complex and the least understood term. Hopkins and Antes (1990) defined evaluation as a continuous inspection of all available information in order to form a valid judgment of students' learning and/or the effectiveness of education program.

The central idea in evaluation is "value." When we evaluate a variable, we are basically judging its worthiness, appropriateness and goodness. Evaluation is always done against a standard, objectives or criterion. In teaching learning process teachers made students' evaluations that are usually done in the context of comparisons between what was intended (learning, progress, behaviour) and what was obtained.


Evaluation is much more comprehensive term than measurement and assessment. It includes both quantitative and qualitative descriptions of students' performance. It always provides a value judgment regarding the desirability of the performance for example, Very good, good etc.

| | |
|---|---|
| Kizlik 2011 | http://www.adprima.com/measurement.htm |

| | |
|---|---|
| **Activity 1.1:** | Distinguish among measurement, assessment and evaluation with the help of relevant examples |

## 1.2 Classroom Assessment: Why, What, How and When

According to Carole Tomlinson "Assessment is today's means of modifying tomorrow's instruction." It is an integral part of teaching learning process. It is widely accepted that effectiveness of teaching learning process is directly influenced by assessment. Hamidi (2010) developed a framework to answer the Why; What, How and When to assess. This is helpful in understanding the true nature of this concept.

**Why to Assess**: Teachers have clear goals for instruction and they assess to ensure that these goals have been or are being met. If objectives are the destination, instruction is the path to it then assessment is a tool to keep the efforts on track and to ensure that the path is right. After the completion of journey assessment is the indication that destination is ahead.

**What to Assess**: Teachers cannot assess whatever they themselves like. In classroom assessment, teachers are supposed to assess students' current abilities in a given skill or task. The teacher can assess students' knowledge, skills or behaviour related to a particular field.

**Who to Assess**: It may seem strange to ask whom a teacher should assess in the classroom, but the issue is of great concern. Teachers should treat students as 'real learners', not as course or unit coverers. They should also predict that some students are more active and some are less active; some are quick at learning and some are slow at it. Therefore, classroom assessment calls for a prior realistic appraisal of the individuals teachers are going to assess.

**How to Assess**: Teachers employ different instruments, formal or informal, to assess their students. Brown and Hudson (1998) reported that teachers use three sorts of assessment methods – selected-response assessments, constructed-response assessments, and personal-response assessments. They can adjust the assessment types to what they are going to assess.

**When to Assess:** There is a strong agreement of educationists that assessment is interwoven into instruction. Teachers continue to assess the students learning throughout the process of teaching. They particularly do formal assessments when they are going to make instructional decisions at the formative and summative levels, even if those decisions are small. For example, they assess when there is a change in the content; when there is a shift in pedagogy, when the effect of the given materials or curriculum on learning process is examined.

**How much to Assess**: There is no touchstone to weigh the degree to which a teacher should assess students. But it doesn't mean that teachers can evaluate their students to the extent that they prefer. It is generally agreed that as students differ in ability, learning styles, interests and needs etc so assessment should be limited to every individual's needs, ability and knowledge. Teachers' careful and wise judgment in this regard can prevent teachers from over assessment or underassessment.

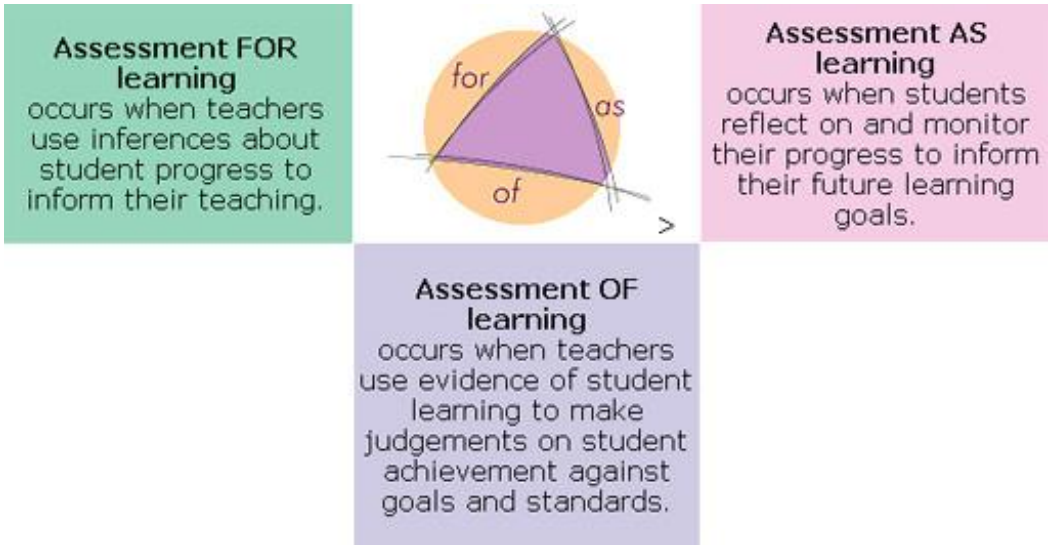
**Activity:** Critically discuss the significance of decisions that teachers take regarding classroom Assessment.

## 1.3 Types of Assessment

***"As coach and facilitator, the teacher uses formative assessment to help support and enhance student learning, As judge and jury, the teacher makes summative judgments about a student's achievement..."***

Atkin, Black & Coffey (2001)

Assessment is a purposeful activity aiming to facilitate students' learning and to improve the quality of instruction. Based upon the functions that it performs, assessment is generally divided into three types: assessment *for* learning, assessment *of* learning and assessment *as* learning.




**a) Assessment *for* Learning (Formative Assessment)**

Assessment *for* learning is a continuous and an ongoing assessment that allows teachers to monitor students on a day-to-day basis and modify their teaching based on what the students need to be successful. This assessment provides students with the timely, specific feedback that they need to enhance their learning. The essence of formative assessment is that the information yielded by this type of assessment is used on one hand to make immediate decisions and on the other hand based upon this information; timely feedback is provided to the students to enable them to learn better. If the primary purpose of assessment is to support high-quality learning then formative assessment ought to be understood as the most important assessment practice.

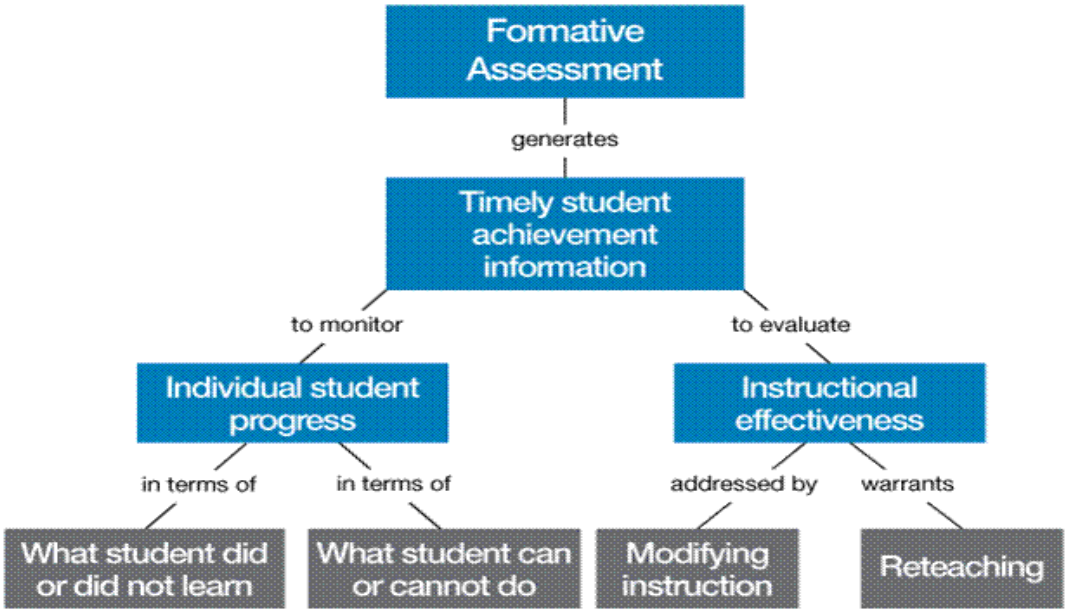
| | |
|---|---|
| The National Center for Fair and Open Testing (1999). | The Value of Formative Assessment. http://www.fairtest.org/examarts/winter99/k-forma3.html |

Assessment for learning has many unique characteristics for example this type of assessment is taken as “practice." Learners should not be graded for skills and concepts that have been just introduced. They should be given opportunities to practice. Formative assessment helps teachers to determine next steps during the learning process as the instruction approaches the summative assessment of student learning. A good analogy for this is the road test that is required to receive a driver's license. Before the final driving test, or summative assessment, a learner practice by being assessed again and again to point out the deficiencies in the skill

Another distinctive characteristic of formative assessment is student involvement. If students are not involved in the assessment process, formative assessment is not practiced or implemented to its full effectiveness. One of the key components of engaging students in the assessment of their own learning is providing them with descriptive feedback as

they learn. In fact, research shows descriptive feedback to be the most significant instructional strategy to move students forward in their learning. Descriptive feedback provides students with an understanding of what they are doing well. It also gives input on how to reach the next step in the learning process.

Role of assessment for learning in instructional process can be best understood with the help of following diagram.




Source:
http://www.stemresources.com/index.php?option=com_content&view=article&id=52&Itemid=70

Garrison, & Ehringhaus, (2007) identified some of the instructional strategies that can be used for formative assessment:

- **Observations. Observing students' behaviour and tasks can help teacher to identify** if students are on task or need clarification. Observations assist teachers in gathering evidence of student learning to inform instructional planning.
- **Questioning strategies.** Asking better questions allows an opportunity for deeper thinking and provides teachers with significant insight into the degree and depth of understanding. Questions of this nature engage students in classroom dialogue that both uncovers and expands learning.
- **Self and peer assessment.** When students have been involved in criteria and goal setting, self-evaluation is a logical step in the learning process. With peer evaluation, students see each other as resources for understanding and checking for quality work against previously established criteria.

- **Student record keeping** It also helps the teachers to assess beyond a "grade," to see where the learner started and the progress they are making towards the learning goals.

**b) Assessment *of* Learning (Summative Assessment)**

Summative assessment or assessment of learning is used to evaluate students' achievement at some point in time, generally at the end of a course. The purpose of this assessment is to help the teacher, students and parents know how well student has completed the learning task. In other words summative evaluation is used to assign a grade to a student which indicates his/her level of achievement in the course or program.

Assessment of learning is basically designed to provide useful information about the performance of the learners rather than providing immediate and direct feedback to teachers and learners, therefore it usually has little effect on learning. Though high quality summative information can help and guide the teacher to organize their courses, decide their teaching strategies and on the basis of information generated by summative assessment educational programs can be modified.

Many experts believe that all forms of assessment have some formative element. The difference only lies in the nature and the purpose for which assessment is being conducted.

**Comparing Assessment *for* Learning and Assessment *of* Learning**

| **Assessment *for* Learning (Formative Assessment)** | **Assessment *of* Learning (Summative Assessment)** |
|---|---|
| Checks how students are learning and is there any problem in learning process. it determines what to do next. | Checks what has been learned to date. |
| Is designed to assist educators and students in improving learning? | Is designed to provide information to those not directly involved in classroom learning and teaching (school administration, parents, school board), in addition to educators and students? |
| Is used continually? | Is periodic? |
| Usually uses detailed, specific and descriptive feedback—in a formal or informal report. | Usually uses numbers, scores or marks as part of a formal report. |
| Usually focuses on improvement, compared with the student's own previous performance | Usually compares the student's learning either with other students' learning (norm-referenced) or the standard for a grade |

| | |
|---|---|
| | level (criterion-referenced) |

Source: adapted from Ruth Sutton, unpublished document, 2001, in Alberta Assessment Consortium

**c) Assessment *as* Learning**

Assessment *as* learning means to use assessment to develop and support students' metacognitive skills. This form of assessment is crucial in helping students become lifelong learners. As students engage in peer and self-assessment, they learn to make sense of information, relate it to prior knowledge and use it for new learning. Students develop a sense of efficacy and critical thinking when they use teacher, peer and self-assessment feedback to make adjustments, improvements and changes to what they understand.

| | |
|---|---|
| Garrison, C., & Ehringhaus, M. (2007) | Defining Formative and Summative Assessment<br>http://www.education.vic.gov.au/images/content/studentlearning/forofas.jpg |

**Self Assessment**: 'Formative assessment results in improved teaching learning process.' Comment on the statement and give arguments to support your response.

## 1.4 Characteristics of Classroom Assessment

**1. Effective assessment of student learning begins with educational goals.**

Assessment is not an end in itself but a vehicle for educational improvement. Its effective practice, then, begins with and enacts a vision of the kinds of learning we most value for students and strive to help them achieve. Educational values/ goals should drive not only what we choose to assess but also how we do so. Where questions about educational mission and values are skipped over, assessment threatens to be an exercise in measuring what's easy, rather than a process of improving what we really care about.

**2. Assessment is most effective when it reflects an understanding of learning as multidimensional, integrated, and revealed in performance over time.**

Learning is a complex process. It entails not only what students know but what they can do with what they know; it involves not only knowledge and abilities but values, attitudes, and habits of mind that affect both academic success and performance beyond the classroom. Assessment should reflect these understandings by employing a diverse array of methods, including those that call for actual performance, using them over time

so as to reveal change, growth, and increasing degrees of integration. Such an approach aims for a more complete and accurate picture of learning, and therefore, firm base for improving our students' educational experience.

**3. Assessment works best when it has a clear, explicitly stated purposes.**

Assessment is a goal-oriented process. It entails comparing educational performance with educational purposes and expectations -- those derived from the institution's mission, from faculty intentions in program and course design, and from knowledge of students' own goals. Where program purposes lack specificity or agreement, assessment as a process pushes a campus towards clarity about where to aim and what standards to apply; assessment also prompts attention to where and how program goals will be taught and learned. Clear, shared, implementable goals are the cornerstone for assessment that is focused and useful.

**4. Assessment requires attention to outcomes but also and equally to the experiences that lead to those outcomes.**

Information about outcomes is of high importance; where students "end up" matters greatly. But to improve outcomes, we need to know about student experience along the way -- about the curricula, teaching, and kind of student effort that lead to particular outcomes. Assessment can help us understand which students learn best under what conditions; with such knowledge comes the capacity to improve the whole of their learning.

**5. Assessment works best when it is ongoing not episodic.**

Assessment is a process whose power is cumulative. Though isolated, "one-shot" assessment can be better than none, improvement is best fostered when assessment entails a linked series of activities undertaken over time. This may mean tracking the process of individual students, or of cohorts of students; it may mean collecting the same examples of student performance or using the same instrument semester after semester. The point is to monitor progress towards intended goals in a spirit of continuous improvement. Along the way, the assessment process itself should be evaluated and refined in light of emerging insights.

**6. Assessment is effective when representatives from across the educational community are involved.**

Student education is a campus-wide liability, and assessment is a way of acting out that responsibility. Thus, while assessment attempts may start small, the aim over time is to involve people from across the educational community. Faculty plays an important role, but assessment's questions can't be fully addressed without participation by educators, librarians, administrators, and students. Assessment may also involve individuals from

beyond the campus (alumni/ae, trustees, employers) whose experience can enrich the sense of appropriate aims and standards for learning. Thus understood, assessment is not a task for small groups of experts but a collaborative activity; its aim is wider, better-informed attention to student learning by all parties with a stake in its improvement.

**7. Assessment makes a difference when it begins with issues of use and illuminates questions that people really care about.**

Assessment recognizes the value of information in the process of improvement. But to be useful, information must be connected to issues or questions that people really care about. This implies assessment approaches that produce evidence that relevant parties will find credible, suggestive, and applicable to decisions that need to be made. It means thinking in advance about how the information will be used, and by whom. The point of assessment is not to collect data and return "results"; it is a process that starts with the questions of decision-makers, that involves them in the gathering and interpreting of data, and that informs and helps guide continuous improvement.

**9. Through effective assessment, educators meet responsibilities to students and to the public.**

There is a compelling public stake in education. As educators, we have a responsibility to the public that support or depend on us to provide information about the ways in which our students meet goals and expectations. But that responsibility goes beyond the reporting of such information; our deeper obligation -- to ourselves, our students, and society -- is to improve. Those to whom educators are accountable have a corresponding obligation to support such attempts at improvement. (American Association for Higher Education; 2003)

| | |
|---|---|
| **Activity 1.2:** | Effective assessment involves representatives from across the educational community: Discuss |

## 1.5 Role of Assessment

*"Teaching and learning are reciprocal processes that depend on and affect one another. Thus, the assessment component deals with how well the students are learning and how well the teacher is teaching"* Kellough and Kellough, (1999)

Assessment does more than allocate a grade or degree classification to students – it plays an important role in focusing their attention and, as Sainsbury & Walker (2007) observe, actually drives their learning. Gibbs (2003) states that assessment has 6 main functions:

1. Capturing student time and attention
2. Generating appropriate student learning activity

3. Providing timely feedback which students pay attention to
4. Helping students to internalize the discipline's standards and notions of equality
5. Generating marks or grades which distinguish between students or enable pass/fail decisions to be made.
6. Providing evidence for other outside the course to enable them to judge the appropriateness of standards on the course.

Surgenor (2010) summarized the role of assessment in learning in the following points.

- It fulfills student expectations
- It is used to motivate students
- It provide opportunities to remedy mistakes
- It indicate readiness for progression
- Assessment serves as a diagnostic tool
- Assessment enables grading and degree classification
- Assessment works as a performance indicator for students
- It is used as a performance indicator for teacher
- Assessment is also a performance indicator for institution
- Assessment facilitates learning in the one way or the other.

| **Activity 1.3:** | Enlist different role of formative and summative assessment in teaching learning process. |
|---|---|

## 1.6 Principles of Classroom Assessment

Hamidi (2010) described following principles of classroom assessment.

### 1. Assessment should be formative

Classroom assessment should be carried out regularly in order to inform on-going teaching and learning. It should be formative because it refers to the formation of a concept or process. To be formative, assessment is concerned with the way the student develops, or forms. So it should be *for* learning. In other words, it has a crucial role in "informing the teacher about how much the learners as a group, and how much individuals within that group, have understood about what has been learned or still needs learning as well as the suitability of their classroom activities, thus providing feedback on their teaching and informing planning. Teachers use it to see how far learners have

mastered what they should have learned. So classroom assessment needs fully to reach its formative potential if a teacher is to be truly effective in teaching.

**2. Should determine planning**

Classroom assessment should help teachers plan for future work. First, teachers should identify the purposes for assessment – that is, specify the kinds of decisions teachers want to make as a result of assessment. Second, they should gather information related to the decisions they have made. Next, they interpret the collected information—that is, it must be contextualized before it is meaningful. Finally, they should make the final, or the professional, decisions. The plans present a means for realizing instructional objectives which are put into practice as classroom assessment to achieve the actual outcomes.

**3. Assessment should serve teaching**

Classroom assessment serves teaching through providing feedback on pupils' learning that would make the next teaching event more effective, in a positive, upwards direct. Therefore, assessment must be an integral part of instruction. Assessment seems to drive teaching by forcing teachers to teach what is going to be assessed. Teaching involves assessment; that is, whenever a student responds to a question, offers a comment, or tries out a new word or structure, the teacher subconsciously makes an assessment of the student's performance. So when they are teaching, they are also assessing. A good teacher never ceases to assess students, whether those assessments are incidental or intended.

**4. Assessment should serve learning.**

Classroom assessment is an integral part of learning process as well. The ways in which learners are assessed and evaluated strongly affect the ways they study and learn. It is the process of finding out who the students are, what their abilities are, what they need to know, and how they perceive the learning will affect them. In assessment, the learner is simply informed how well or badly he/she has performed. It can spur learners to set goals for themselves. Assessment and learning are seen as inextricably linked and not separate processes because of their mutually-influenced features. Learning by itself has no meaning without assessment and vice-versa.

**5. Assessment should be curriculum-driven**

Classroom assessment should be the servant, not the master, of the curriculum. Assessment specialists view it as an integral part of the entire curriculum cycle. Therefore, decisions about how to assess students must be considered from the very beginning of curriculum design or course planning.

**6. Assessment should be interactive**

Students should be proactive in selecting the content for assessment. It provides a context for learning as meaning and purpose for learning and engages students in social interaction to develop oral and written language and social skills. Assessment and learning are inextricably linked and not separate processes, Effective assessment is not a process carried out by one person, such as a teacher, on another, a learner, it is seen as a two-way process involving interaction between both parties. Assessment, then, should be viewed as an interactive process that engages both teacher and student in monitoring the student's performance.

**7. Assessment should be student-centered**

Since learner-centered methods of instruction are principally concerned with learner needs, students are encouraged to take more responsibility for their own learning and to choose their own learning goals and projects. Therefore, in learner-centered assessment, they are actively involved in the process of assessment. Involving learners in aspects of classroom assessment minimizes learning anxiety and results in greater student motivation.

**8. Assessment should be diagnostic**

Classroom assessment is diagnostic because teachers use it to find out learners' strengths and weaknesses during the in-progress class instruction. They also identify learning difficulties. If the purpose of assessment is to provide diagnostic feedback, then this feedback needs to be provided in a form – either verbal or written – that is for learners to understand and use.

**9. Assessment should be exposed to learners**

Teachers are supposed to enlighten learners' accurate information about assessment. In other words, it should be transparent to learners. They must know when the assessments occur, what they cover in terms of skills and materials, how much the assessments are worth, and when they can get their results and the results are going to be used. They must also be aware of why they are assessed because they are part of the assessment process. Because the assessment is part of the learning process, it should be done *with* learners, not *to* them. It is also important to provide an assessment schedule before the instruction begins.

**10. Assessment should be non-judgmental**

In the classroom assessment, everything focuses on learning which results from a number of such factors as student needs, student motivation, teaching style, time on task, study intensity, background knowledge, course objectives, etc. So there is no praise or blame for a particular outcome of learning. Teachers should take no stance on determining who

has done better and who has failed to perform well. Assessment should allow students to have reasonable opportunities to demonstrate their expertise without confronting barriers

### 11. Assessment should develop a mutual understanding

Mutual understanding occurs when two people come to a similar feeling of reality. In second language learning, this understanding calls for a linguistic environment in which the teacher and students interact with each other based on the assessment objectives. Therefore, assessment has the ability to create a new world image by having the individuals share their thoughts helpful in learning process. When learning occurs, this is certainly as a result of common understanding between the teacher and students.

### 12. Assessment should lead to learner's autonomy

Autonomy is a principle in which students come to a state of making their own decisions in language learning. They assume a maximum amount of responsibility for what they learn and how they learn it. Autonomous learning occurs when students have made a transition from teacher assessment to self-assessment. This requires that teachers encourage students to reflect on their own learning, to assess their own strengths and weaknesses, and to identify their own goals for learning. Teachers also need to help students develop their self-regulating and met cognitive strategies. Autonomy is a construct to be fostered in students, not taught, by teachers.

### 13. Assessment should involve reflective teaching

Reflective teaching is an approach instruction in which teachers are supposed to develop their understanding of teaching (quality) based on data/information obtained and collected through critical reflection on their teaching experiences. This information can be gathered through formative assessment (i.e., using different methods and tools such as class quizzes, questionnaires, surveys, field notes, feedback from peers, classroom ethnographies, observation notes, etc) and summative assessment (i.e., different types of achievement tests taken at the end of the term).

| | |
|---|---|
| Hamidi, Eameal (2010) | Fundamental Issues in L2 Classroom Assessment Practices. Academic Leadership Online Journal. Volume 8 Issue 2<br>http://www.sisd.net/cms/lib/TX01001452/Centricity/Domain/2073/ALJ_ISSN1533-7812_8_2_444.pdf |

## 1.7 Self Assessment Questions

- Highlight the role of assessment in teaching and learning Process
- Discuss critically the principles of assessment with the help of relevant examples
- Differentiate between assessment for learning and assessment of learning

## 1.8 References/Suggested Reading's


- Catherine Garrison, Dennis Chandler & Michael Ehringhaus, (2009). Effective Classroom Assessment: Linking Assessment with Instruction: NMSA & Measured Progress Publishers

- Kathleen Burke, (2010). How to assess authentic learning. California: Corwin Press

- Charles Hopkins, (2008). Classroom Measurement and Evaluation. Illinois: Peacock

- Carolin Gipps, ( 1994) Beyond Testing : Towards a Theory of Educational Assessment Routledge Publishers

**UNIT–2**

# OBJECTIVES AND ASSESSMENT

**Written by:**
**Prof. Dr. Rehana Masrur**

**Reviewed By:**
**Dr. Naveed Sultana**

# CONTENTS

## LIST OF TABLES



## LIST OF FIGURES

# INTRODUCTION

In this unit you will learn that how important are the objectives and learning outcomes in the process of assessment. A teacher should know that the main advantage of objectives is to guide the teaching-learning activities. In simple words these are the desired outcomes of an effort. Guided by these specific objectives instructional activities are designed and subsequently assessment is carried out through different methods. One of the most common methods to assess the ability of a student in any specific subject is a test. Most tests taken by students are developed by teachers. The goal of this unit is for you to be able to design, construct, and analyze a test for a given set of objectives or content area. Therefore, the objective are key components for developing a test. These are the guiding principles for assessment. For achievement testing cognitive domain is very much emphasized and widely used by educationists. Taxonomy of Educational Objectives developed by Benjamin Bloom (1956) deals with activities like memorizing, interpreting, analyzing and so on. This taxonomy provides a useful way of describing the complexity of an objective by classifying into one of the hierarchical categories from simplest to complex. One of the important task for a teacher while designing a test is the selection and sampling of test items from course contents. The appropriateness of the content of a test is considered at earliest stages of development. Therefore, the process of developing a test should begin with the identification of content domain at first stage and development of table of specification at second stage. In this unit we have focused on what we want students to learn and what content we want our tests to cover.

You will learn that how to work on different stages of assessment.

# OBJECTIVES

After studying this unit, you should be able to;

- describe the role of objectives and outcomes in the assessment of student achievement.
- explain the purpose of a test.
- explain levels of Cognitive Domain.
- develop achievement objectives according to Bloom Taxonomy of Educational objectives.
- identify and describe the major components of a table of specifications.
- identify and describe the factors which determine the appropriate numbers of items for each component in a table of specification.

## 2.1 Purpose of a Test

Assessment of a student in class is inevitable because it is integral part of teaching-learning process. Assessment on one hand provides information to design or redesign instruction and on the other hand it promotes learning. Teachers use different techniques and procedures to assess their students i.e tests, observations, questionnaires, interviews, rating scales, discussion etc. A teacher develops, administers, and marks academic achievement and other types of tests in order to measure the ability of a student in a subject or measures behaviour in class or in school. What are these tests? Does a teacher really need to know that what is test? Yes, it is very important. The teaching-learning process remains incomplete if a teacher does not know that how well her class is doing and to what extent her teaching is effective in terms of achievement of pre defined objectives. There are many technical terms which are related with assessment. Before we go any further, it would be beneficial to define first what is a test.

**What is a Test?**

A test is a device which is used to measure behaviour of a person for a specific purpose. Moreover it is an instrument that typically uses sets of items designed to measure a domain of learning tasks. Tests are systematic method of collecting information that lead to make inferences about the characteristics of people or objects. A teacher must understand that educational test is a measuring device and therefore involves rules (administering, scoring) for assigning numbers that will be used for describing the performance of an individual. You should also keep in mind that it is not possible for a teacher to test all the subject matter of a course that has been taught to the class in a semester or in a year. Therefore, teacher prepares tests while sampling the items from a pool of items in such a way that it represents the whole subject matter. Teacher must also understand that whole content with many topics and concepts that have been taught within a semester or in a year can not be tested in one or two hours. In simple words a test should assess content area in accordance with relative importance a teacher has assigned to them. It is believed most commonly that the meaning of a test is simple paper-and-pencil tests. But now a days other testing procedures have been developed and are practiced in many schools.

Even tests are of many types that can be placed into two main categories. These are:

(i) Subjective type tests

(ii) Objective type tests

At elementary level students do not have much proficiency of writing long essay type answer of a question, therefore, objective type tests are preferred. Objective type tests are also called selective-response tests. In this types of tests responses of an item are provided and the students are required to choose correct response. The objective types of tests that are used at elementary level are:

(i) Multiple choice

(ii) Multiple Binary-choice

(iii) Matching items

You will study about the development process of each of these items in next units. In this unit you have been given just an idea that what does a test mean for a teacher. Definitely after going through this discussion you might be ready to extract yourself from the above mentioned paragraphs that why it is important for a teacher to know about a classroom test. What purpose it serves? The job of a teacher is to teach and to test for the following:

**Purposes of test:**

You have learned that a test is a simple device which measures the achievement level of a student in a particular subject and grade. Therefore we can say that a test is used to serve the following purposes:

**1. Monitoring Student Progress**

Why should teacher assess their students? The simple answer is that it helps teachers to know whether their students are making satisfactory progress. We must realize that the appropriate use of tests and other assessment procedures allows a teacher to monitor the progress of their students. A useful purpose of classroom test is to know whether students are satisfactorily moving towards the instructional goals. After knowing the weaknesses if any, the teacher will modify her/his instructional design. If the progress is adequate there will be no need of instructional changes. The results obtained during the monitoring of students progress can further be utilized for making formative assessment of their instructional procedures. Formative evaluation provides feedback to students as well as to the teachers.

**2. Diagnosing Learning Problems**

Identification of students strength and weaknesses is one of the main purpose of a test. An elementary teacher needs to know that whether a student is comprehending the content that he/she reads. If he/she reads with certain difficulties, then definitely as a teacher you have to address the problem instructionally. Otherwise, it will be wastage of time and energy if students are not comprehending but the teacher is moving forward. Thus by measuring students current status teacher can determine:

(i) How to improve students weaknesses through instructional changes?

(ii) How to instructionally avoid already mastered skills and knowledge?

The diagnosis taken before instruction is usually referred as pre-testing or pre-assessment. It provides the teacher that what is the level of previous knowledge the students possess at the beginning of instruction.

### 3. Assigning Grades

A teacher assigns grade after scoring the test. The best way to assign grades is to collect objective information related to student achievements and other academic accomplishments. Different institutions have different criteria for assigning the grades. Mostly alphabets 'A, B, C, D, or F are assigned on the bases of numerical evidence.

### 4. Classification and Selection of Students

A teacher makes different decisions regarding the classification, selection and placement of students. Though these terms are used interchangeably, but technically they have different meanings. On the bases of test scores students are classified in to high ability, average ability and low ability groups. Or test can be used to classify students having learning disabilities, emotionally disturbed children, or some other category of disability (speech handicap etc). On the basis of test score students are selected or rejected for admission in schools, colleges and or in other institutions. As contrary to selection, while making placement decisions no one is rejected rather all students are placed in various categories of educational levels, for example regular, remedial, or honors.

### 5. Evaluating Instruction

Students' performance on tests helps the teacher to evaluate her/his own instructional effectiveness or to know that how effective their teaching have been. A teacher teaches a topic for two weeks. After the completion of topic the teacher gives a test. The score obtained by students show that they learned the skills and knowledge that was expected to learn. But if the obtained score is poor, then the teacher will decide to retain, alter or totally discard their current instructional activities.

| | |
|---|---|
| **Activity-2.1:** | Visit some schools of your area and perform the following:<br><br>Conduct an interview of at least 10 teachers and ask the teachers why do they administer the tests to their students. Match their responses with the purposes of test (1-5) given in section 2.3. |

## 2.2 Objectives and Educational Outcomes

### 1. Definition of Objectives

Education is, without any doubt, a purposeful activity. Every step of this activity has and should definitely have a particular purpose. Therefore learning objectives are a prime and integral part of teaching learning process.

A learning objective refers to the statement of what students will obtain through instruction of certain content. In other words 'an *objective* is a description of a performance you want learners to be able to exhibit before you consider them competent. An objective describes an intended *result* of instruction, rather than the *process* of instruction itself.' (Mager, p. 5)

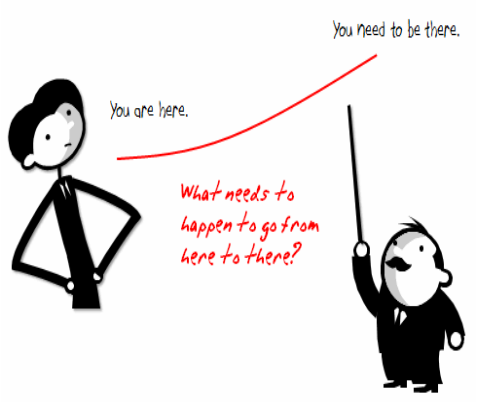



**Figure 2.1 Defining objectives**

In teaching learning process, learning objectives have a unique importance. The role learning objectives play includes but is not limited to the following three: firstly, they guide and direct for the selection of instructional content and procedures. Secondly, they facilitate the appropriate evaluation of the instruction. Thirdly, learning objectives help the students to organize their efforts to accomplish the intent of the instruction.

## 2 Characteristics/ Attributes of the Objectives

Good objectives have three essential characteristics:

- *Behaviour* - Firstly, an objective must explain the competency to be learned, the intended change in the behaviour of the learners. For this purpose it is necessary to use the verb in the statement of the objective which identifies an observable behaviour of the learner.
- *Criterion* - Secondly, an objective must clarify the intended degree of performance. In other words objective should not only indicate the change in the behaviour of the students but also the level or degree of that change as well. For this purpose the statement of the objective must indicate a degree of accuracy, a quantity or proportion of correct responses or the like.
- *Conditions* - Thirdly, an objective should describe the conditions under which the learning will occur. In other words, under what circumstances the learner will develop the competency? What will the learner be given or already be expected to know to accomplish the learning? For example, a condition could be stated as, told a case study, shown a diagram, given a map, after listening a lecture or observing a demonstration, after through reading, etc

Though all the three characteristics are essential for stating clear objectives, in some cases one or two of these elements are easily implied by a simple statement.

## 3 Taxonomy of Educational Objectives

Following the 1948 Convention of the American Psychological Association, a group of college examiners considered the need for a system of classifying educational goals for the evaluation of student performance. Years later and as a result of this effort, Benjamin Bloom formulated a classification of "the goals of the educational process". Eventually, Bloom established a hierarchy of educational objectives for categorizing level of abstraction of questions that commonly occur in educational settings (Bloom, 1965). This classification is generally referred to as Bloom's Taxonomy. Taxonomy means 'a set of classification principles', or 'structure'. The followings are six levels in this taxonomy: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. The detail is given below:

**Cognitive domain**: The cognitive domain (Bloom, 1956) involves the development of intellectual skills. This includes the recall or recognition of specific facts, procedural patterns, and concepts that serve in the development of intellectual abilities and skills. There are six levels of this domain starting from the simplest cognitive behaviour to the most complex. The levels can be thought of as degrees of difficulties. That is, the first ones must normally be mastered before the next ones can take place.

**Affective domain**: The affective domain is related to the manner in which we deal with things emotionally, such as feelings, values, appreciation, enthusiasms, motivations, and attitudes. The five levels of this domain include: receiving, responding, valuing, organization, and characterizing by value.

**Psychomotor domain**: Focus is on physical and kinesthetic skills. The psychomotor domain includes physical movement, coordination, and use of the motor-skill areas. Development of these skills requires practice and is measured in terms of speed, precision, distance, procedures, or techniques in execution. There are seven levels of this domain from the simplest behaviour to the most complex. Domain levels include: Perception, set, guided response, mechanism, complex or overt response, adaptation.

- http://www.nwlink.com/~donclark/hrd/bloom.html
- http://www.learningandteaching.info/learning/bloomtax.htm

Over all Bloom's taxonomy is related to the three Hs of education process that are Head, Heart and Hand.

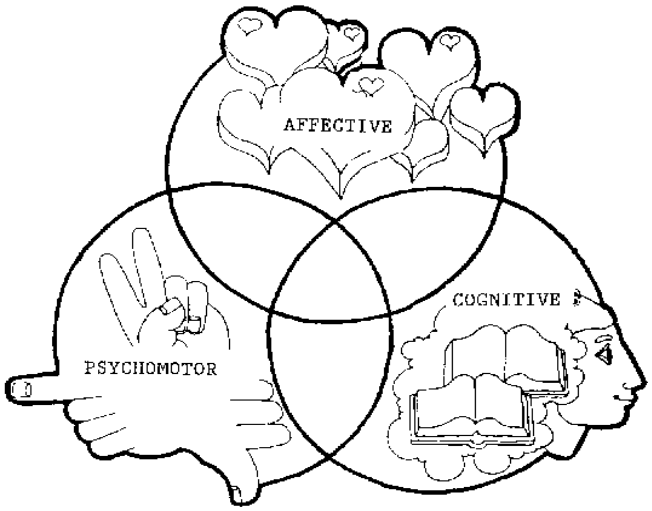



**Figure -2.2 Taxonomy of Educational Objectives**

Note: In each of the three domains Bloom's Taxonomy is based on the premise that the categories are ordered in degree of difficulty. **An important premise of Bloom's Taxonomy is that each 'level' must be mastered before progressing to the next**. As such the levels within each domain are levels of learning development, and these levels increase in difficulty.

## 2.3 Writing Cognitive Domain Objectives

In teaching learning process, cognitive domain of Blooms taxonomy is of prime focus. So let's discuss this domain in detail and learn to write objectives of this domain.

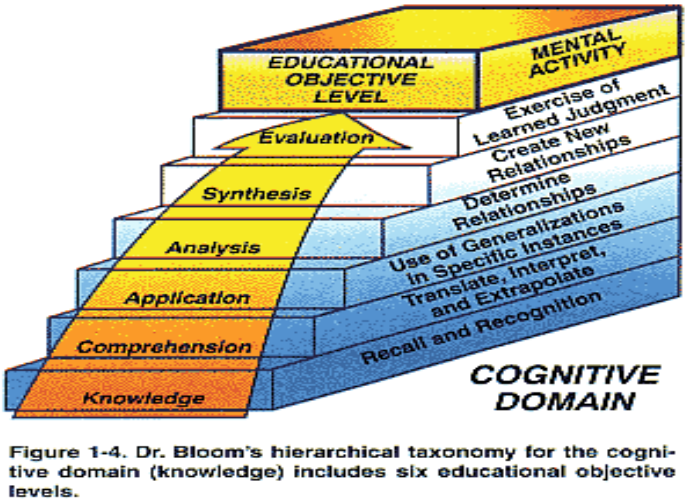


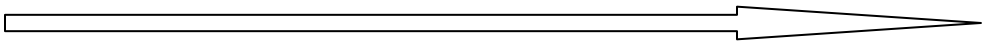

Figure 1-4. Dr. Bloom's hierarchical taxonomy for the cognitive domain (knowledge) includes six educational objective levels.

**Figure -2.3 Bloom's Hierarchical Taxonomy of Educational Objectives**

Cognitive abilities in this taxonomy are arranged on continuum ranging from the lower to the higher

Lower

Higher



Knowledge Comprehension Application Analysis Synthesis Evaluation

An analogy depicting the taxonomy of learning objectives can be thought as assembling blocks in building a pyramid. The knowledge level creates the basis for the foundation from which the higher- level skills are built.

When writing educational objectives, a teacher must know that for a good objective it is necessary to ***use the clear verb*** that clearly indicates the type of observable behaviour. The following table will not only help you to understand the level of cognitive domain but will guide you what action verbs can be used to state objectives of that particular level.

**Table 2.1 Learning Objectives and Action Verbs**

| Learning Objective/ Level | Description | Action Verbs to be used to state objectives |
|---|---|---|
| **Knowledge** | The first level of learning is knowledge.<br>Knowledge can be characterized as awareness of specifics and of the ways and means of dealing with specifics. The knowledge level focuses on memory or recall where the learner recognizes information, ideas, principles in the approximate form in which they were learned. | To arrange, to define, to describe, to identify, to list, to label, to name, to order, to recognize, to recall, to relate, to repeat, to reproduce, to state, to underline. |

| | | |
|---|---|---|
| **Comprehension** | Comprehension is the next level of learning and encompasses understanding. Has the knowledge been internalized or understood? The student should be able to translate, comprehend, or interpret information based on the knowledge. | To choose, to compare, to classify, to describe, to demonstrate, to determine, to discuss, to discriminate, to explain, to express, to identify, to indicate, to interpret, to label, to locate, to pick, to recognize, to relate, to report, to respond, to restate, to review, to select, to tell, to translate |
| **Application** | Application is the use of knowledge. Can the student use the knowledge in a new situation? It can also be the application of theory to solve a real world problem. The student selects, transfers, and uses data and principles to complete solve a problem. | To apply, to classify, to demonstrate, to develop,<br>to dramatize, to employ, to generalize, to illustrate, to interpret, to initiate, to operate, to organize, to practice, to relate, to restructure, to rewrite, to schedule, to sketch, to solve, to use, to utilize, to transfer |
| **Analysis** | Analysis involves taking apart a piece of knowledge, the investigation of parts of a concept. It can only occur if the student has obtained knowledge of and comprehends a concept. The student examines, classifies, hypothesizes, collects data, and draws conclusions. | To analyze, to appraise, to calculate, to categorize, compare, conclude, contrast, or criticize; to detect, to debate, to determine, to develop, distinguish, or deduce; to diagram, to diagnose, differentiate, or discriminate; to estimate, to examine, to evaluate, to experiment, to inventory, to inspect, to relate, solve, or test; to question |
| **Synthesis** | Synthesis is the creative act. It's the taking of knowledge and the creation of something new. It is an inductive process—one of building rather than one of breaking down. The student originates, integrates, and combines ideas into something that is new to him/her. | To arrange, to assemble, to collect, to compose, to construct, to constitute, to create, to design, to develop, to device, to document, to<br>formulate, to manage, to modify, to originate, to organize, to plan, to prepare, to predict, to produce, to propose, to relate, to reconstruct,<br>to set up, to specify, to synthesize, to systematize, to tell, to transmit |

| | | |
|---|---|---|
| **Evaluation** | Evaluation is judgment or decision making. The student appraises, assesses or criticizes on a basis of specific standards and criteria. | To appraise, argue, or assess; to attach, to choose, to contrast, to consider, to critique, to decide, to defend, to estimate, to evaluate, to<br>judge, to measure, to predict, to rate, to revise, to score, to select, to support, to standardize,<br>to validate, to value, to test |

***Source:*** *Jolly T. Holden: A Guide To Developing Cognitive Learning Objectives. Retrieved From*

*http://gates.govdl.org/docs/A%20Guide%20to%20Developing%20Cogntive%20Learning%20Objectives.pdf*

**Activity-2.2:** Develop two objectives of comprehension level for this unit by using appropriate action verbs**.**

Bloom's Taxonomy underpins the classical '**Knowledge, Attitude, Skills**' structure of learning. It is such a simple, clear and effective model, both for explanation and application of learning objectives, teaching and training methods, and measurement of learning outcomes.

Bloom's Taxonomy provides an excellent structure for planning, designing, assessing and evaluating teaching and learning process. The model also serves as a sort of **checklist**, by which you can ensure that instruction is planned to deliver all the necessary development for students.

**Bloom's Revised Taxonomy**

Bloom's former students Lorin Anderson and David Krathwohl revised Bloom's Taxonomy in 1990. - Bloom's Revised Taxonomy was published in 2001. Key to this is the use of verbs rather than nouns for each of the categories and a rearrangement of the sequence within the taxonomy. They are arranged below in increasing order, from **Lower Order Thinking Skills (LOTS)** to **Higher Order Thinking Skills (HOTS).**

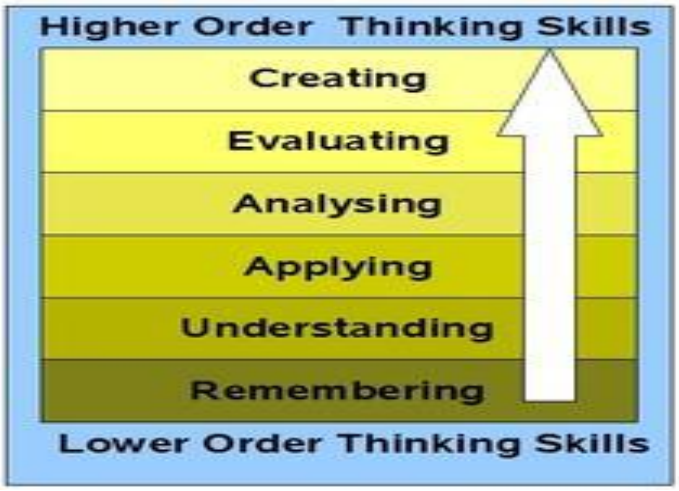



**Figure-2.4 Order of Thinking Skills**

| | |
|---|---|
| **Activity-2.3:** | Identify the differences in original and revised Blooms Taxonomy and discuss wether these changes are desirable? If yes why. |

## 2.4 Defining Learning Outcomes

Learning outcomes are the statements indicating what a student is expected to be able to do as a result of a learning activity. Major difference between learning objectives and out comes is that objectives are focused upon the instruction, what will be given to the students and the outcomes are focused upon the students what behaviour change they are being expected to show as the result of the instruction.

### 1. Different Definitions of Learning Outcomes

Adam, 2004 defines learning outcomes as:

A learning outcome is a written statement of what the successful student/learner is expected to be able to do at the end of the module/course unit, or qualification.

The Credit Common Accord for Wales defines learning outcomes as:

Statements of what a learner can be expected to know, understand and/or do as a result of a learning experience. (QCA /LSC, 2004, p. 12)

University of Exeter (2007) defines:

Learning Outcome: An expression of what a student will demonstrate on the successful completion of a module. Learning outcomes:

- are related to the level of the learning;
- indicate the intended gain in knowledge and skills that a typical student will achieve;
- should be capable of being assessed.

## 2. Difference between Learning Outcomes and Objectives

Learning outcomes and objectives' are often used synonymously, although they are not the same. In simple words, objectives are concerned with teaching and the teacher's intentions whereas learning outcomes are concerned with students learning.

However, objectives and learning outcomes are usually written in same terms. For further detail check the following website.

http://www.qualityresearchinternational.com/glossary/learningoutcomes.htm

## 3. Importance of Learning Outcomes

Learning outcomes facilitate teachers more precisely to tell students what is expected of them. Clearly stated learning outcomes:

- help students to learn more effectively. They know where they stand and the curriculum is made more open to them.
- make it clear what students can hope to gain from a particular course or lecture.
- help instructors select the appropriate teaching strategy, for example lecture, seminar, student self-paced, or laboratory class. It obviously makes sense to match the intended outcome to the teaching strategy.
- help instructors more precisely to tell their colleagues what a particular activity is designed to achieve.
- assist in setting examinations based on the content delivered.
- Help in the selection of appropriate assessment strategies.

| **Activity-2.4** | Differentiate between learning Objective and Outcome with the help of relevant examples |
|---|---|

## 4. SOLO Taxonomy

The SOLO taxonomy stands for:

Structure of
Observed
Learning
Outcomes

SOLO taxonomy was developed by Biggs and Collis (1982) which is further explained by Biggs and Tang (2007). This taxonomy is used by Punjab for the assessment.

It describes level of increasing complexity in a student's understanding of a subject through five stages, and it is claimed to be applicable to any subject area. Not all students get through all five stages, of course, and indeed not all teaching.

**1** **Pre-structural:** here students are simply acquiring bits of unconnected information, which have no organisation and make no sense.

**2** **Unistructural:** simple and obvious connections are made, but their significance is not grasped.

**3** **Multistructural:** a number of connections may be made, but the meta-connections between them are missed, as is their significance for the whole.

**4** **Relational** level: the student is now able to appreciate the significance of the parts in relation to the whole.

**5** At the **extended abstract** level, the student is making connections not only within the given subject area, but also beyond it, able to generalise and transfer the principles and ideas underlying the specific instance.

SOLO taxonomy
http://www.learningandteaching.info/learning/solo.htm#ixzz1nwXTmNn9

## 2.5 Preparation of Content Outline

First you must understand that what is content. In this regard content refers to the major matter that will be included in a measuring device. For example, the test of General Science he diagrams, pictures of different plants, insects or animal or living or non-living things that constitute the test. For a psychomotor test such as conducting an experiment in laboratory might require setting up of apparatus for the experiment. For an effective device, the content might consist of the series of statement to which the students might choose correct or best answer. Most tests taken by students are developed by teachers who are already teaching the subject for which they have to develop the test. Therefore selection of test content might not be the problem for them. Selection and preparation of content also depends on the type of decisions a teacher has to make about the students. If the purpose of a test is to evaluate the instruction, then the content of a test must reflect the age appropriateness. If test is made for making decisions regarding selection then the

content might of predictive nature. This type of test domain will provide information that how well the student will perform in the program.

A teacher should know that items selected for the test come from instructional material which a teacher has covered during teaching. You may heard about students reaction during examination that ‘ test was out of course’. It indicates that teacher while developing the test items has not considered the content that was taught to the student. The items included in the test might have been not covered during the instruction period.

Look at following these diagrams:

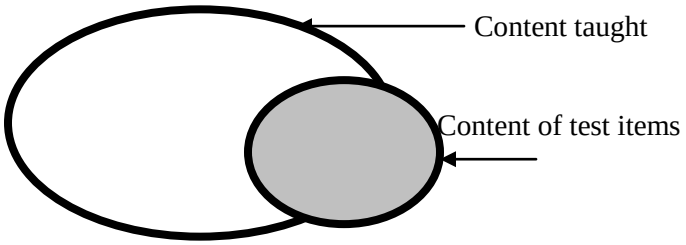



**Figure- 2.5 Poor representativeness**

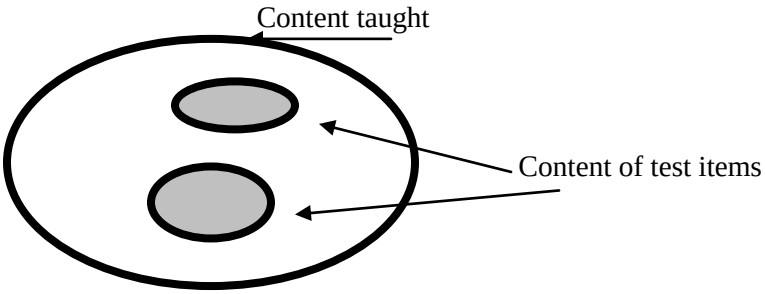



**Figure- 2.6 Inadequate representativeness**

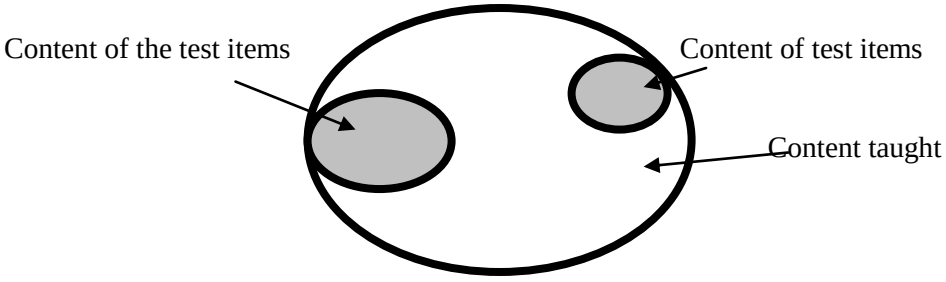


**Figure-2.7 Inadequate representativeness**

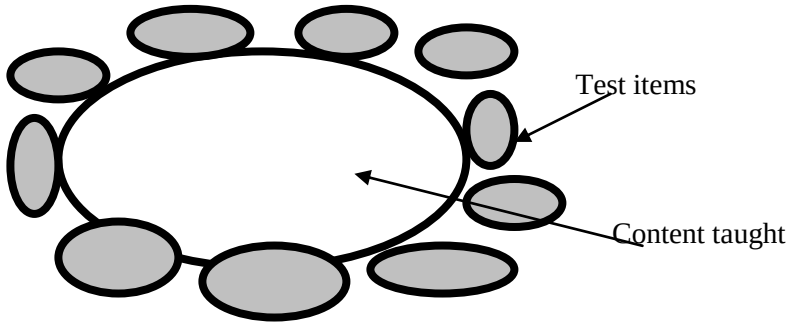



**Figure-2.8 Completely inadequate representativeness**

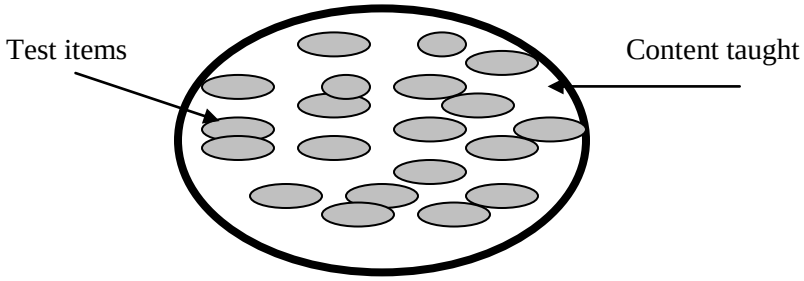



**Figure-2.9 Adequate representativeness**

In figures 2.5 to 2.9 the shaded area represents the test items which cover the content of subject matter whereas un-shaded area is the subject matter (learning domain) which the teacher has taught in the class in the subject of social studies.

Figures 2.5-2.8 show the poor or inadequate representativeness of content of test items. For example in figure-2.5 test covers a small portion (shaded area) of taught content domain, rest of the items do not coincide with the taught domain. In figure 2.5 & 2.6 most of the test items/questions have been taken from a specific part of taught domain, therefore, the representation of taught content domain is inadequate. Though, the test items have been taken from the same content domain. The content of test items in figure 2.7 give very poor picture of a test. None of the parts of taught domain have

been assessed, therefore test shows zero representativeness. None of the test items in figure 2.8 have been taken from the taught content domain. Contrary to this look at figure 2.9, the test items effectively sample the full range of taught content.

It implies that the content from which the test item have to be taken should be well defined and structured. With out setting the boundary of knowledge, behaviour, or skills to be measured, the test development task will become difficult and complex. As a result the assessment will produce unreliable results. Therefore a good test represents the taught content up to maximum extent. A test which is representative of the entire content domain is actually is a good test. Therefore it is imperative for a teacher to prepare outline of the content that will be covered during the instruction. The next step is the selection of subject matter and designing of instructional activities. All these steps are guided by the objectives. One must consider objectives of the unit before selection of content domain and subsequently designing of a test. It is clear from above discussion that the outline of the test content should based on the following principles:

1. Purpose of the test (diagnostic test, classification, placement, or job employment)
2. Representative sample of the knowledge, behaviour, or skill domain being measured.
3. Relevancy of the topic with the content of the subject
4. Language of the content should be according to the age and grade level of the students.
5. Developing table of specification.

A test, which meets the criteria stated in above principles, will provide reliable and valid information for correct decision regarding the individual. Now keeping in view these principles go on the following activity.

**Activity-2.5:**

Visit elementary school of your area and collect question papers/tests of sixth class of any subject developed by the school teachers. Now perform the following:

(1)
   a. How many items are related with the content?
   b. How many items (what percentage) are not related with the content covered for the testing period?
   c. Is the test representative of the entire content domain?
   d. Does the test fulfill the criteria of test construction? Explain.

(2) Share your results electronically with your classmates, and get their opinion on the clarification of concept discussed in unit-2

## 2.6 Preparation of Table of Specification

It has been discussed earlier that the educational objectives play a significant role in the development of classroom tests. The reason is that the preparation of classroom test is closely related to the curriculum and educational objectives. And we have also explained that a test should measure what was taught. For ensuring that there is similarity between classroom instruction and test content is the development and application of **table of specification,** which is also called **a test blue print.** As the name implies, it specifies the content of a test. It is a two-way framework which ensures the congruence between classroom instruction and test content. This is one of the most popular procedures used by test developers for defining the content-domain. One dimension of the test reflects the content to be covered and other dimension describes the kinds of student cognitive behaviour to be assessed. Table 2.2 Provides the example of table of specification.

**Table 2.2 General Table of Specification**

**Number of Test Items for Each Cognitive Level**

| **Topics** | **Knowledge** | **Comprehension** | **Application** | **Analysis** | **Total** |
|---|---|---|---|---|---|
| Topic 1 | 5 | 2 | 2 | 3 | **12** |
| Topic 2 | 3 | 3 | 4 | 2 | **12** |
| Topic 3 | 2 | 2 | 3 | 2 | **9** |
| Topic 4 | 3 | 3 | 1 | 1 | **8** |
| Topic 5 | 1 | 2 | 1 | 1 | **5** |
| Topic 6 | 2 | 2 | 0 | 0 | **4** |
| **Total** | **16** | **14** | **11** | **9** | **50** |

Look at table 2.2, the top of each column of the table represent the level of cognitive domain, the extreme left column represent the categories of the content (topics) or assessment domains. The numerals in the cells of two way table show the numbers of items to be included in the test. You can readily see that how the fifty items in this table have been allocated to the content topics and the levels of cognitive behaviour. The teacher may add some more dimensions. The table of specification represents four level of cognitive domain. It is not necessary for teacher to develop a test that completely coincides with the content of taught domain. The teacher is required to adequately sample the content of the assessment domain. The important consideration here for teachers is that they must make a careful effort on conceptualizing the assessment domain. An appropriate representativeness must be ensured. Unfortunately, many teachers develop tests without figuring out what domains of knowledge, skills, or attitude should be promoted and consequently, formally be assessed. A classroom test should measure what was taught. In simple words a test must emphasize what was emphasized in the

class. Now look at table 2.3. The table of specification shows the illustration of assessment domain of unit-2 of this book:

**Table 2.3 Table of Specification of Unit-2**

**Number of test Items for Each Cognitive Level**

| **Topics** | **Knowledge** | **Comprehension** | **Application** | **Analysis** | **Total** |
|---|---|---|---|---|---|
| Purpose of a test: | 2 | 1 | 1 | | **4** |
| Objectives and Educational outcomes | 2 | 2 | 2 | 1 | **7** |
| Preparation of content outline | 2 | 2 | 2 | | **6** |
| Preparation of table of Specification | 2 | 3 | 2 | 1 | **8** |
| **Total** | **8** | **8** | **7** | **2** | **25** |

Table 2.3 is a very simple table of specification. It is possible to add more dimensions of the content. You may further distribute the table in subtopics for each main topic. Lets have another look on a very specific table of the following:

**Table 2.4 Specific Table of Specification**
**Number of Test Items for following Cognitive Level**
**Knowledge Comprehension Application Analysis**

| Level of Cognitive domain / Topics | Knows symbols & terms | Knows specific facts | Understands effects of factors | Solves equation | Interprets results | **Total** | **Total** |
|---|---|---|---|---|---|---|---|
| Speed & Velocity | 2 | 2 | 2 | 3 | 4 | **13** | **26%** |
| Potential Energy and Kinetic | 4 | 2 | 2 | 4 | 4 | **16** | **32%** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Energy | | | | | | | |
| Law of Motion | 4 | 4 | 4 | 5 | 4 | **21** | **42%** |
| **Total** | **10** | **8** | **8** | **12** | **12** | **50** | **100 %** |
| **Total %** | **20 %** | **16%** | **16%** | **24%** | **24%** | **100 %** | |

A table of specification helps teachers to review the curriculum content on one hand and on the other hand it helps teachers to be careful in overlooking important concepts or including unimportant and irrelevant concepts. On the similar patterns a teacher can develop table of specification for affective and psychomotor domain.

**Activity 2.6:** Prepare table of specification for unit-2, you have just studied.

## 2.7 Self- Assessment Questions:

(1) Explain with examples the purpose a classroom test.

(2) How do you define an objective and a outcome? Differentiate between objectives and outcomes with the help of examples.

(3) What is your understanding on the importance of learning outcomes?

(4) What is cognitive domain? Explain all levels with examples.

(5) Develop two objectives for measuring recall level, two objectives for measuring application level and two for evaluation level for $5^{th}$ class from English text book,

(6) Prepare a table of specification of 50 items for General Science subject for $6^{th}$ class.

## 2.8 References Suggested Readings:


Anderson, L.W., Krathwohl, D.R. (Eds.), (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. New York: Longman

Adam, S., 2004, *Using Learning Outcomes: A Consideration of the Nature, role, Application and Implications for European Education of Employing 'Learning Outcomes' at the Local, National and International Levels.* United Kingdom Bologna

Seminar 1–2 July 2004, Heriot-Watt University (Edinburgh Conference Centre) Edinburgh. Scotland.

Gronlund, N. E. (2006). *Assessment of Student Achievement. (Eighth Edition).* USA: Pearson Education.

Popham, W.J. (2005). *Classroom Assessment: What Teachers Need to Know*. USA: Pearson Education.


**Web References**


SOLO taxonomy
http://www.learningandteaching.info/learning/solo.htm#ixzz1nwXTmNn9

http://www.nwlink.com/~donclark/hrd/bloom.html
http://www.learningandteaching.info/learning/bloomtax.html

*http://gates.govdl.org/docs/A%20Guide%20to%20Developing%20Cogntive%20Learning%20Objectives.pdf*

http://www.qualityresearchinternational.com/glossary/learningoutcomes.htm

**UNIT–3**

# TYPES OF ASSESSMENT TESTS AND TECHNIQUES

**Written By:**
**Dr. Naveed Sultana**

**Reviewed By:**
**Dr. Muhammad Tanveer Afzal**

## CONTENT

# INTRODUCTION

Educational reformers are seeking answers to two fundamental questions: (1) How well are students learning? And (2) how effectively are teachers teaching? Classroom assessment responds directly concern about better learning and more effective teaching. Classroom assessment, involves students and teachers in the continuous monitoring of students' learning. It provides faculty with feedback about their effectiveness as teachers, and it gives students a measure of their progress as learners. Most important, because classroom assessments are created, administered, and analyzed by teachers themselves on questions of teaching and learning that are important to them, the likelihood that instructors will apply the results of the assessment to their own teaching is greatly enhances. The classroom assessment process assumes that students need to receive feedback early and often, that they need to evaluate the quality of their own learning, and that they can help the teacher improve the strength of instruction. Assessment is integral to the teaching–learning process, facilitating student learning and improving instruction, and can take a variety of forms. Classroom assessment is generally divided into three types: assessment for learning, assessment of learning and assessment as learning. Classroom assessment is the process of collecting information from your students about their experience as learners in your class. There are many different ways of collecting information, depending on what you are teaching and what kind of information teacher need.

All types of assessment are based on the principle that the more clearly and specifically to understand how students are learning, the more effectively teacher can teach them. When assessing the classroom, some issues to consider are how to allow all students to contribute, how to respond to the student feedback, and how often to collect feedback. For this purpose teacher uses different modes such as test and techniques for assessing (a) course-related knowledge and skills; (b) learner attitudes, values, and self-awareness; and (c) learner reactions to teachers and teaching. Classroom assessment test and techniques are formative evaluation methods that serve two purposes. They can help you to assess the degree to which your students understand the course content and they can provide information about the effectiveness of teaching learning process. So this unit addresses the different types of tests and techniques and their application for assessing the degree to which students understand the course contents and they can provide information about the effectiveness of teaching learning process.

# OBJECTIVES

After studying this unit, prospective teachers will be able to:

1. understand and describe the different types of tests and techniques.
2. examine the purposes and characteristics of tests and techniques.
3. describe the role of tests and techniques for improving the teaching learning process.
4. analyze the advantages and disadvantages of each type of test and technique.

## 3.1 Tests

### 3.1.1 Achievement Tests

Achievement tests are widely used throughout education as a method of assessing and comparing student performance. Achievement tests may assess any or all of reading, math, and written language as well as subject areas such as science and social studies. These tests are available to assess all grade levels and through adulthood. The test procedures are highly structured so that the testing process is the same for all students who take them.

It is developed to measure skills and knowledge learned in a given grade level, usually through planned instruction, such as training or classroom instruction. Achievement tests are often contrasted with tests that measure aptitude, a more general and stable cognitive trait.

Achievement test scores are often used in an educational system to determine what level of instruction for which a student is prepared. High achievement scores usually indicate a mastery of grade-level material, and the readiness for advanced instruction. Low achievement scores can indicate the need for remediation or repeating a course grade.

Teachers evaluate students by: observing them in the classroom, evaluating their day-to-day class work, grading their homework assignments, and administrating unit tests. These classroom assessments show the teacher how well a student is mastering grade level learning goals and provide information to the teacher that can be used to improve instruction. Overall achievement testing serves following purposes:

- Assess level of competence
- Diagnose strength and weaknesses
- Assign Grades
- Achieve Certification or Promotion
- Advanced Placement/College Credit Exams
- Curriculum Evaluation
- Accountability
- Informational Purposes

**(i) Types of Achievement Tests**

**(a) Summative Evaluation**:

Testing is done at the end of the instructional unit. The test score is seen as the summation of all knowledge learned during a particular subject unit.

**(a) Formative Evaluation**:

Testing occurs constantly with learning so that teachers can evaluate the effectiveness of teaching methods along with the assessment of students' abilities.

**(ii) Advantages of Achievement Test:**

- One of the main advantages of testing is that it is able to provide assessments that are psychometrically valid and reliable, as well as results which are generalized and replicable.
- Another advantage is aggregation. A well designed test provides an assessment of an individual's mastery of a domain of knowledge or skill which at some level of aggregation will provide useful information. That is, while individual assessments may not be accurate enough for practical purposes, the mean scores of classes, schools, branches of a company, or other groups may well provide useful information because of the reduction of error accomplished by increasing the sample size.

**(iii) Designing the Test**

**Step 1:** The first step in constructing an effective achievement test is to identify what you want students to learn from a unit of instruction. Consider the relative importance of the objectives and include more questions about the most important learning objectives.

**Writing the questions:**

**Step2:** Once you have defined the important learning objectives and have, in the light of these objectives, determined which types of questions and what form of test to use, you are ready to begin the second step in constructing an effective achievement test. This step is writing the questions.

**Step3:** Finally, review the test. Are the instructions straightforward? Are the selected learning objectives represented in appropriate proportions? Are the questions carefully and clearly worded? Special care must be taken not to provide clues to the test-wise student. Poorly constructed questions may actually measure not knowledge, but test-taking ability.

**(iv) General Principles:**

While the different types of questions--multiple choice, fill-in-the-blank or short answer, true-false, matching, and essay--are constructed differently, the following principles apply to construct questions and tests in general.

- Make the instructions for each type of question simple and brief.
- Use simple and clear language in the questions. If the language is difficult, students who understand the material but who do not have strong language skills may find it difficult to demonstrate their knowledge. If the language is ambiguous, even a student with strong language skills may answer incorrectly if his or her interpretation of the question differs from the instructor's intended meaning.
- Write items that require specific understanding or ability developed in that course, not just general intelligence or test-wiseness.

- Do not suggest the answer to one question in the body of another question. This makes the test less useful, as the test-wise student will have an advantage over the student who has an equal grasp of the material, but who has less skill at taking tests.
- Do not write questions in the negative. If you must use negatives, highlight them, as they may mislead students into answering incorrectly.
- Specify the units and precision of answers. For example, will you accept numerical answers that are rounded to the nearest integer?

**(v) Interpreting the Test Results:**

If you have carefully constructed an achievement test using the above principles, you can be confident that the test will provide useful information about the students' knowledge of the learning objectives. Considering the questions relating to the various learning objectives as separate subtests, you can develop a profile of each student's knowledge of or skill in the objectives. The scores of the subtests can be a useful supplement to the overall test score, as they can help you identify specific areas which may need attention. A carefully-constructed achievement test can, by helping you know what your students are learning, help you to teach more effectively and, ultimately, help the students to master more of the objectives.

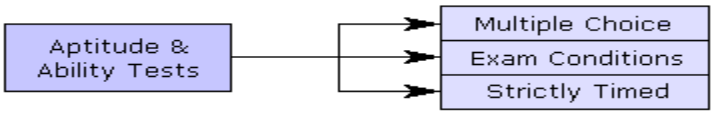
| **Activity 3.1:** | Prepare the achievement test on content to be taught of any subject while focusing its steps and discuss with your course mates. |
|---|---|

### 3.1.2 Aptitude Tests

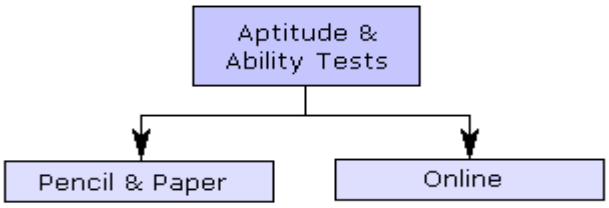
Aptitude tests assume that individuals have inherent strengths and weaknesses, and are naturally inclined toward success or failure in certain areas based on their inherent characteristics.

Aptitude tests determine a person's ability to learn a given set of information. They do not test a person's knowledge of existing information. The best way to prepare for aptitude tests is to take practice tests.

Aptitude and ability tests are designed to assess logical reasoning or thinking performance. They consist of multiple choice questions and are administered under exam conditions. They are strictly timed and a typical test might allow 30 minutes for 30 or so questions. Test result will be compared to that of a control group so that judgments can be made about your abilities.



You may be asked to answer the questions either on paper or online. The advantages of online testing include immediate availability of results and the fact that the test can be taken at employment agency premises or even at home. This makes online testing particularly suitable for initial screening as it is obviously very cost-effective.




**(i) Types of Aptitude Test**

The following is a list of the different types of aptitude test that are used for assessment process.

**(a) Critical Thinking**

Critical thinking is defined as a form of reflective reasoning which analyses and evaluates information and arguments by applying a range of intellectual skills in order to reach clear, logical and coherent judgments within a given context. Critical thinking tests force candidates to analyse and evaluate short passages of written information and make deductions to form answers.

**(b) Numerical Reasoning Tests**

**Numerical tests**, sometimes known as **numerical reasoning**, are used during the application process at all major investment banks and accountancy & professional services firms. Test can be either written or taken online. The tests are usually provided by a third party.

Perceptual Speed Tests
**Perceptual speed** is the ability to quickly and accurately compare letters, numbers, objects, pictures, or patterns. In tests of perceptual speed the things to be compared may be presented at the same time or one after the other. Candidates may also be asked to compare a presented object with a remembered object.

**(c) Spatial Visualization Tests**

Spatial visualization ability or Visual-spatial ability refers to the ability to mentally manipulate 2-dimensional and 3-dimensional figures. It is typically measured with simple

cognitive tests and is predictive of user performance with some kinds of user interfaces

### (d) Logical Reasoning Tests

**Logical reasoning** aptitude tests (also known as **Critical Reasoning** Tests) may be either verbal (word based, e.g. "Verbal Logical Reasoning"), numerical (number based, e.g. "Numerical Logical Reasoning") or diagrammatic (picture based, see diagrammatic tests for more information).

### (e) Verbal Reasoning Tests

Verbal reasoning tests are a form of aptitude test used by interviewers to find out how well a candidate can assess verbal logic. In a verbal reasoning test, you are typically provided with a passage, or several passages, of information and required to evaluate a set of statements by selecting one of the following possible answers.

### (f) Perceptual Speed Tests:

Perceptual speed is the ability to quickly and accurately compare letters, members, objects, pictures, or patterns. In tests of perceptual speed the things to be compared may presented at the same time or one after the other. Candidates may also be asked to compare a presented object with a remembered object.

### (ii) Value of Aptitude Tests

Aptitude tests tell us what a student brings to the task regardless of the specific curriculum that the student has already experienced. The difference between aptitude and achievement tests is sometimes a matter of degree. Some aptitude and achievement tests look a lot alike. In fact, the higher a student goes in levels of education, the more the content of aptitude tests resembles achievement tests. This is because the knowledge that a student has already accumulated is a good predictor of success at advanced levels.

In addition, group aptitude tests--usually given as part of a group achievement battery of tests--can be given quickly and inexpensively to large numbers of children. Children who obtain extreme scores can be easily identified to receive further specialized attention. Aptitude tests are valuable in making program and curricula decisions.

- They are excellent predictors of future scholastic achievement.
- They provide ways of comparing a child's performance with that of other children in the same situation.
- They provide a profile of strengths and weaknesses.
- They assess differences among individuals.
- They have uncovered hidden talents in some children, thus improving their educational opportunities.

- They are valuable tools for working with handicapped children.

**(iii) How can we use aptitude test results?**

In general, aptitude test results have three major uses:

**(a) Instructional**

Teachers can use aptitude test results to adapt their curricula to match the level of their students, or to design assignments for students who differ widely. Aptitude test scores can also help teachers form realistic expectations of students. Knowing something about the aptitude level of students in a given class can help a teacher identify which students are not learning as much as could be predicted on the basis of aptitude scores. For instance, if a whole class were performing less well than would be predicted from aptitude test results, then curriculum, objectives, teaching methods, or student characteristics might be investigated.

**(b) Administrative**

Aptitude test scores can identify the general aptitude level of a high school, for example. This can be helpful in determining how much emphasis should be given to college preparatory programs. Aptitude tests can be used to help identify students to be accelerated or given extra attention, for grouping, and in predicting job training performance.

**(c) Guidance**

Guidance counselors use aptitude tests to help parents develop realistic expectations for their child's school performance and to help students understand their own strengths and weaknesses.

| **Activity:** 3.2 | Discuss with your course mate about their aptitudes towards teaching profession and analyze their opinions. |
|---|---|

### 3.1.3 Attitude

Attitude is a posture, action or disposition of a figure or a statue. A mental and neural state of readiness, organized through experience, exerting a directive or dynamic influence upon the individual's response to all objects and situations with which it is related.

Attitude is the state of mind with which you approach a task, a challenge, a person, love, life in general. The definition of attitude is “a complex mental state involving beliefs and feelings and values and dispositions to act in certain ways”. These beliefs and feelings are different due to various interpretations of the same events by various people and these differences occur due to the earlier mentioned inherited characteristics’.

**(i) Components of Attitude**

**1. Cognitive Component:**

It refers that's part of attitude which is related in general know how of a person, for example, he says smoking is injurious to health. Such type of idea of a person is called cognitive component of attitude.

**2. Effective Component:**

This part of attitude is related to the statement which affects another person. For example, in an organization a personal report is given to the general manager. In report he points out that the sale staff is not performing their due responsibilities. The general manager forwards a written notice to the marketing manager to negotiate with the sale staff.

**3. Behavioral Component:**

The behavioral component refers to that part of attitude which reflects the intension of a person in short run or long run. For example, before the production and launching process the product. Report is prepared by the production department which consists of the intention in near future and long run and this report is handed over to top management for the decision.

**(ii) List of Attitude:**

In the broader sense of the word there are only three attitudes, a positive attitude, a negative attitude, and a neutral attitude. But in general sense, an attitude is what it is expressed through. Given below is a list of attitudes that are expressed by people, and are more than personality traits which you may have heard of, know of, or might be even carrying them:

- Acceptance
- Confidence
- Seriousness
- Optimism
- Interest
- Cooperative
- Happiness
- Respectful
- Authority
- Sincerity
- Honest

- Sincere

| | |
|---|---|
| **Activity:** | Develop an attitude scale for analyzing the factors motivating the prospective teachers to join teaching profession. |

### 3.1.4 Intelligence Tests

Intelligence involves the ability to think, solve problems, analyze situations, and understand social values, customs, and norms. Two main forms of intelligence are involved in most intelligence assessments:

- Verbal Intelligence is the ability to comprehend and solve language-based problems; and
- Nonverbal Intelligence is the ability to understand and solve visual and spatial problems.

Intelligence is sometimes referred to as intelligence quotient (IQ), cognitive functioning, intellectual ability, aptitude, thinking skills and general ability.

While intelligence tests are psychological tests that are designed to measure a variety of mental functions, such as reasoning, comprehension, and judgment.

**Intelligence test** is often defined as a measure of general mental ability. Of the standardized intelligence tests, those developed by David Wechsler are among those most widely used. Wechsler defined intelligence as "the global capacity to act purposefully, to think rationally, and to deal effectively with the environment." While psychologists generally agree with this definition, they don't agree on the **operational definition** of intelligence (that is, a statement of the procedures to be used to precisely define the variable to be measured) or how to accomplish its measurement.

The goal of intelligence tests is to obtain an idea of the person's intellectual potential. The tests center around a set of stimuli designed to yield a score based on the test maker's model of what makes up intelligence. Intelligence tests are often given as a part of a battery of tests.

#### (i) Types of Intelligence Tests

Intelligence tests (also called instruments) are published in several forms:

**(a)** **Group Intelligence tests** usually consist of a paper test booklet and scanned scoring sheets. Group achievement tests, which assess academic areas, sometimes include a cognitive measure. In general, group tests are not recommended for the purpose of identifying a child with a disability. In some cases, however, they can be helpful as a screening measure to consider whether further testing is needed and can provide good background information on a child's academic history.

**(b)** **Individual intelligence tests** may include several types of tasks and may involve easel test books for pointing responses, puzzle and game-like tasks, and question and answer sessions. Some tasks are timed.

**(c)** **Computerized tests** are becoming more widely available, but as with all tests, examiners must consider the needs of the child before choosing this format.

**(d)** **Verbal tests** evaluate your ability to spell words correctly, use correct grammar, understand analogies and analyze detailed written information. Because they depend on understanding the precise meaning of words, idioms and the structure of the language they discriminate very strongly towards native speakers of the language in which the test has been developed. If you speak English as a second language, even if this is at a high standard, you will be significantly disadvantaged in these tests. There are two distinct types of verbal ability questions, those dealing with spelling, grammar and word meanings, and those that try to measure your comprehension and reasoning abilities. Questions about spelling, grammar and word meanings are speed tests in that they don't require very much reasoning ability. You either know the answer or you don't.

**(e)** **Non-verbal tests** are comprised of a variety of item types, including series completion, codes and analogies. However, unlike verbal reasoning tests, none of the question types requires learned knowledge for its solution. In an educational context, these tests are typically used as an indication of a pupil's ability to understand and assimilate novel information independently of language skills. Scores on these tests can indicate a pupil's ability to learn new material in a wide range of school subjects based on their current levels of functioning.

**(ii) Advantages**

In general, intelligence tests measure a wide variety of human behaviours better than any other measure that has been developed. They allow professionals to have a uniform way of comparing a person's performance with that of other people who are similar in age. These tests also provide information on cultural and biological differences among people.

Intelligence tests are excellent predictors of academic achievement and provide an outline of a person's mental strengths and weaknesses. Many times the scores have revealed talents in many people, which have led to an improvement in their educational opportunities. Teachers, parents, and psychologists are able to devise individual curricula that matches a person's level of development and expectations.

**(iii) Disadvantages**

Some researchers argue that intelligence tests have serious shortcomings. For example, many intelligence tests produce a single intelligence score. This single score is often inadequate in explaining the multidimensional.

Another problem with a single score is the fact that individuals with similar intelligence test scores can vary greatly in their expression of these talents. It is important to know the person's performance on the various subtests that make up the overall intelligence test score. Knowing the performance on these various scales can influence the understanding of a person's abilities and how these abilities are expressed. For example, two people have identical scores on intelligence tests. Although both people have the same test score, one person may have obtained the score because of strong verbal skills while the other may have obtained the score because of strong skills in perceiving and organizing various tasks.

Furthermore, intelligence tests only measure a sample of behaviors or situations in which intelligent behavior is revealed. For instance, some intelligence tests do not measure a person's everyday functioning, social knowledge, mechanical skills, and/or creativity. Along with this, the formats of many intelligence tests do not capture the complexity and immediacy of real-life situations. Therefore, intelligence tests have been criticized for their limited ability to predict non-test or nonacademic intellectual abilities. Since intelligence test scores can be influenced by a variety of different experiences and behaviors, they should not be considered a perfect indicator of a person's intellectual potential.

**Activity 3.4*:***

Discuss with your course mate about the intelligence testing and identify the methods used to measure intelligence, and make a list of problems in measuring intelligence

### 3.1.5 Personality Tests

Your personality is what makes you who you are. It's that organized set of unique traits and characteristics that makes you different from every other person in the world. Not only does your personality make you special, it makes you!?

> *"The particular pattern of behavior and thinking that prevails across time and contexts, and differentiates one person from another."*

The goal of psychologists is to understand the causes of individual differences in behavior. In order to do this one must firstly identify personality characteristics (often called personality traits), and then determine the variables that produce and control them. A personality trait is assumed to be some enduring characteristic that is relatively constant as opposed to the present temperament of that person which is not necessarily a stable characteristic. Consequently, trait theories are specifically focused on explaining the more permanent personality characteristics that differentiate one individual from another. For example, things like being; dependable, trustworthy, friendly, cheerful, etc.

A personality test is completed to yield a description of an individual's distinct personality traits. In most instances, your personality will influence relationships with your family, friends, and classmates and contribute to your health and well being. Teachers can administer a personality test in class to help your children discover their strengths and developmental needs. The driving force behind administering a personality test is to open up lines of communication and bring students together to have a higher appreciation for one another. A personality test can provide guidance to teachers of what teaching strategies will be the most effective for their students. Briefly p*ersonality test can benefit your students by:*

- Increasing productivity
- Get along better with classmates
- Help students realize their full potential
- Identify teaching strategies for students
- Help students appreciate other personality types.

### (i) Types of Personality Tests

Personality tests are used to determine your type of personality, your values, interests and your skills. They can be used to simply assess what type of person you are or, more specifically, to determine your aptitude for a certain type of occupation or career.

There are many different types of personality tests such as self report inventory, Likert scale and projective tests.

### (a) Self-report Inventory

A self-report inventory is a type of psychological test often used in personality assessment. This type of test is often presented in a paper-and-pencil format or may even be administered on a computer. A typical self report inventory presents a number of questions or statements that may or may not describe certain qualities or characteristics of the test subject.

Chances are good that you have taken a self-report inventory at some time the past. Such questionnaires are often seen in doctors' offices, in on-line personality tests and in market research surveys. This type of survey can be used to look at your current behaviors, past behaviors and possible behaviors in hypothetical situations.

### (i) Strengths and Weaknesses of Self-Report Inventories

Self-report inventories are often good solution when researchers need to administer a large number of tests in relatively short space of time. Many self report inventories can be

completed very quickly, often in as little as 15 minutes. This type of questionnaire is an affordable option for researchers faced with tight budgets.

Another strength is that the results of self report inventories are generally much more reliable and valid. Scoring of the tests a standardized and based on norms that have been previously established.

However, self report inventories do have their weaknesses. Such as people are able to exercise deception while taking self report tests (Anastasi & Urbina, 1997).

Another weakness is that some tests are very long and tedious. For example, the MMPI takes approximately 3 hours to complete. In some cases, test respondents may simply lose interest and not answer questions accurately. Additionally, people are sometimes not the best judges of their own behavior. Some individuals may try to hide their own feelings, thoughts and attitudes.

**(ii) Types of Self Reports**

- **Myers-Briggs Inventory**
  First designed to help suite people's personality to jobs
  identifies 'type' of person not 'traits' in people
- **MMPI & MMPI-2**
  used to assess personality and mental health
- **16 Personality Factor Questionnaire**
  identifies a person's traits
- **The Big Five**
  identifies on a scale of five traits where a person sits

**(b) Likert Scale**

A Likert Scale is a type of psychometric scale frequently used in psychology questionnaires. It was developed by and named after organizational psychologist Rensis Likert. A Likert item is simply a statement which the respondent is asked to evaluate according to any kind of subjective or objective criteria; generally the level of agreement or disagreement is measured. It is considered symmetric or "balanced" because there are equal amounts of positive and negative positions. Often five ordered response levels are used, although many psychometricians advocate using seven or nine levels.

The format of a typical five-level Likert item, for example, could be:

1. Strongly disagree
2. Disagree
3. Uncertain
4. Agree

5. Strongly Agree

Likert scaling is a bipolar scaling method, measuring either positive or negative response to a statement. Sometimes an even-point scale is used, where the middle option of "Neither agree nor disagree" is not available. This is sometimes called a "forced choice" method, since the neutral option is removed. The neutral option can be seen as an easy option to take when a respondent is unsure, and so whether it is a true neutral option is questionable. It has been shown that when comparing between a 4-point and a 5-point Likert scale, where the former has the neutral option unavailable, the overall difference in the response is negligible.

**(c) Projective tests**

A **projective test** is a personality test designed to let a person respond to ambiguous stimuli, presumably revealing hidden emotions and internal conflicts. In psychology, a projective test is a type of personality test in which the individual offers responses to ambiguous scenes, words or images. This type of test emerged from the psychoanalytic school of thought, which suggested that people have unconscious thoughts or urges. These projective tests were intended to uncover such unconscious desires that are hidden from conscious awareness.

**(i) How Do Projective Test Work?**

In many projective tests, the participant is shown an ambiguous image and then asked to give the first response that comes to mind. The key to projective tests is the ambiguity of the stimuli. According to the theory behind such tests, clearly defined questions result in answers that are carefully crafted by the conscious mind. By providing the participant with a question or stimulus that is not clear, the underlying and unconscious motivations or attitudes are revealed.

**(ii) Types of Projective Tests**

There are a number of different types of projective tests. The following are just a few examples of some of the best-known projective tests.

**(a)** The Rorschach Inkblot Test

The Rorschach Inkblot was one of the first projective tests and continues to be one of the best-known. Developed by Swiss psychiatrist Hermann Rorschach in 1921, the test consists of 10 different cards that depict an ambiguous inkblot. The participant is shown one card at a time and asked to describe what he or she sees in the image. The responses are recorded verbatim by the tester. Gestures, tone of voice and other reactions are also noted. The results of the test can vary depending on which of the many existing scoring systems the examiner uses.

**(b) The Thematic Apperception Test (TAT)**

In the Thematic Apperception Test, an individual is asked to look at a series of ambiguous scenes. The participant is then asked to tell a story describing the scene, including what is happening, how the characters are feeling and how the story will end. The examiner then scores the test based on the needs, motivations and anxieties of the main character as well as how the story eventually turns out.

**(iii) Strengths and Weaknesses of Projective Tests**

- Projective tests are most frequently used in therapeutic settings. In many cases, therapists use these tests to learn qualitative information about a client. Some therapists may use projective tests as a sort of icebreaker to encourage the client to discuss issues or examine thoughts and emotions.
- While projective tests have some benefits, they also have a number of weaknesses and limitations. For example, the respondent's answers can be heavily influenced by the examiner's attitudes or the test setting. Scoring projective tests is also highly subjective, so interpretations of answers can vary dramatically from one examiner to the next.

| | |
|---|---|
| **Activity 3.5:** | Apply the projective tests to any class and analyze the traits of students which differ them with each other. |

### 3.1.6 Norm-referenced Tests and Criterion-Referenced Tests

Tests can be categorized into two major groups: norm-referenced tests and criterion-referenced tests. These two tests differ in their intended purposes, the way in which content is selected, and the scoring process which defines how the test results must be interpreted.

**(a) Definition of Norm-Referenced Test**

Norm-referenced tests are made with compare test takers to each other. On an NRT driving test, test-takers would be compared as to who knew most or least about driving rules or who drove better or worse. Scores would be reported as a percentage rank with half scoring above and half below the mid-point.

This type of test determines a student's placement on a normal distribution curve. Students compete against each other on this type of assessment. This is what is being referred to with the phrase, 'grading on a curve'.

**(b) *Definition of Criterion-Referenced Tests***

*Criterion-referenced tests* are intended to measure how well a person has learned a specific body of knowledge and skills.

Criterion-referenced test is a term which is used daily in classes. These tests assess specific skills covered in class.

Criterion-referenced tests measure specific skills and concepts. Typically, they are designed with 100 total points possible. Students are earned points for items completed correctly. The students' scores are typically expressed as a percentage. Criterion-referenced tests are the most common type of test teacher's use in daily classroom work.

**(c) Norm- Reference V.S Criterion-Referenced Testing**

Norm-referenced tests compare an examinee's performance to that of other examinees. Standardized examinations such as the SAT are norm-referenced tests. The goal is to rank the set of examinees so that decisions about their opportunity for success can be made.

Criterion-referenced tests differ in that each examinee's performance is compared to a pre-defined set of criteria or a standard. The goal with these tests is to determine whether or not the candidate has the demonstrated mastery of a certain skill or set of skills. These results are usually "pass" or "fail" and are used in making decisions about job entry, certification, or licensure. A national board medical exam is an example of a Criterion Reference Test. Either the examinee has the skills to practice the profession, in which case he or she is licensed, or does not.

**(i) Purposes of Criterion and Norm – Reference testing**

The major reason for using a norm-referenced test is to classify students. Norm Reference Tests are designed to highlight achievement differences between and among students to produce a dependable rank order of students across a continuum of achievement from high achievers to low achievers. School systems might want to classify students in this way so that they can be properly placed in remedial or gifted programs. These types of tests are also used to help teachers select students for different ability level reading or mathematics instructional groups.

With norm-referenced tests, a representative group of students is given the test prior to its availability to the public. The scores of the students who take the test after publication are then compared to those of the norm group.

While norm-referenced tests ascertains the rank of students, criterion-referenced tests determine what test takers can do and what they know, not how they compare to others

Criterion Reference Tests report how well students are doing relative to a pre-determined performance level on a specified set of educational goals or outcomes included in the school, district, or state curriculum.

Educators or policy makers may choose to use a Criterion Reference Test when they wish to see how well students have learned the knowledge and skills which they are expected to have mastered. This information may be used as one piece of information to determine

how well the student is learning the desired curriculum and how well the school is teaching that curriculum.

Both Norm Reference Tests and Criterion Reference Tests can be standardized. The U.S. Congress, Office of Technology Assessment defines a standardized test as one that uses uniform procedures for administration and scoring in order to assure that the results from different people are comparable. Any kind of test--from multiple choices to essays or oral examinations--can be standardized if uniform scoring and administration are used. This means that the comparison of student scores is possible. Thus, it can be assumed that two students who receive the identical scores on the same standardized test demonstrate corresponding levels of performance. Most national, state and district tests are standardized so that every score can be interpreted in a uniform manner for all students and schools.

**(ii) Comparison of CRT/NRT Characteristics**

| **Criterion-Referenced Tests** | **Norm-Referenced Tests** |
|---|---|
| • To determine whether each student has achieved specific skills or concepts based on standards.<br>• Measures specific skills which make up a designated curriculum. These skills are identified by teachers and curriculum experts<br>• Each individual is compared with a preset standard for acceptable achievement. The performance of other examinees is irrelevant.<br>• Student's score is usually expressed as a percentage. Student achievement is reported for individual skills. | • To rank each student with respect to the achievement of others in order to discriminate between high and low achievers.<br>• Measures broad skill areas sampled from a variety of textbooks, syllabi, and the judgments of curriculum experts.<br>• Each individual is compared with other examinees and assigned a score--usually expressed as a percentile. Student achievement is reported for broad skill areas, although some norm-referenced tests do report student achievement for individual skills |

**(iii) Advantage of Criterion Referenced Test**

Following are the major advantages of criterion referenced tests:

First, students are only tested on their knowledge of specific goals or standards. For example, if you had taught a lesson on adding fractions, you will give the student a test on adding fractions. If he or she scores 85% that means that that particular student has

learned 85% of that goal. If a student does not score particularly well, then the teacher can adjust their instruction accordingly.

Another benefit is that if students do not seem to master a particular standard, the teacher will be able to go back and teach that standard again until the student performs better.

### (iv) Disadvantages of Criterion-Referenced Tests

Criterion-referenced tests have some built-in disadvantages. Creating tests that are both valid and reliable requires fairly extensive and expensive time and effort. In addition, results cannot be generalized beyond the specific course or program. Such tests may also be compromised by students gaining access to test questions prior to exams. Criterion-referenced tests are specific to a program and cannot be used to measure the performance of large groups.

### (v) Advantages of Norm reference Test

The advantage of a norm-referenced test is that it shows us how our student is doing related to other students across the country. They are good for using the placement of students at the beginning and then again four or six months later, or at the end of the year. This will show growth over the period of the time.

Norm-referenced tests along with informal observational evaluation are useful for showing student growth over time. They aren't to be used for grading though they can be one element in a total grade. One must remember we can't expect great growth, if any, over short periods of times, particularly as shown on a norm-referenced test.

### (v) Disadvantage of Norm Reference test

An obvious disadvantage of norm-referenced tests is that it cannot measure progress of the population as a whole, only where individuals fall within the whole. Thus, only measuring against a fixed goal can be used to measure the success of an educational reform program which seeks to raise the achievement of all students against new standards which seek to assess skills beyond choosing among multiple choices. However, while this is attractive in theory, in practice the bar has often been moved in the face of excessive failure rates, and improvement sometimes occurs simply because of familiarity with and teaching to the same test.

| | |
|---|---|
| **Activity 3.6:** | Discuss with your course mate about characteristics of norm and criterion referenced tests and prepare a report about their usability. |

## 3.2 Techniques

### 3.2.1 Questionnaire

A questionnaire is a research instrument consisting of a series of questions and other prompts for the purpose of gathering information from respondents. Although they are often designed for statistical analysis of the responses, this is not always the case.

A questionnaire is a list of written questions that can be completed in one of two basic ways

**Firstly,** respondents could be asked to complete the questionnaire with the researcher not present. This is a postal questionnaire and (loosely) refers to any questionnaire that a respondent completes without the aid of the researcher.

**Secondly,** respondents could be asked to complete the questionnaire by verbally responding to questions in the presence of the researcher. This variation is called a structured interview.

Although the two variations are similar (a postal questionnaire and a structured interview could contain exactly the same questions), the difference between them is important. If, for example, we are concerned with protecting the respondent's anonymity then it might be more appropriate to use a postal questionnaire than a structured interview.

### (i) Different Types of Questions in Questionnaire Design

The following is a list of the different types of questions in questionnaire design:

### 1. Open Format Questions

Open format questions are those questions that give your audience an opportunity to express their opinions. In these types of questions, there are no predetermined set of responses and the person is free to answer however he/she chooses. By including open format questions in your questionnaire, you can get true, insightful and even unexpected suggestions. Qualitative questions fall under the category of open format questions. An ideal questionnaire would include an open format question at the end of the questionnaire that would ask the respondent about suggestions for changes or improvements.

**Example of an Open Format Question**

| |
|---|
| *State your opinion about the quality of teaching during workshop*. |
| …….. |
| …….. |

### 2. Closed Format Questions

Closed format questions are questions that include multiple choice answers. Multiple choice questions fall under the category of closed format questions. These multiple choices could either be in even numbers or in odd numbers. By including closed format questions in your questionnaire design, you can easily calculate statistical data and percentages. Preliminary analysis can also be performed with ease. Closed format questions can be asked to different groups at different intervals. This can enable you to efficiently track opinion over time.

**Example of an Open Format Question**

| |
|---|
| *Which are the elements necessary for classroom teaching?* |
| *Circle those elements:* |
| **(a)** *Teacher (b) Library (c) Lesson planning (d) Laptop* |

### 3. Leading Questions

Leading questions are questions that force your audience for a particular type of answer. In a leading question, all the answers would be equally likely. An example of a leading question would be a question that would have choices such as, fair, good, great, poor, superb, excellent etc. By asking a question and then giving answers such as these, you will be able to get an opinion from your audience.

**Example of an Open Format Question**

| |
|---|
| How would you rate lecture method? |
| (i) Fair (ii) Good (iii) Excellent (iv) Superb |

### 4. Importance Questions

In importance questions, the respondents are usually asked to rate the importance of a particular issue, on a rating scale of 1-5. These questions can help you grasp what are the things that hold importance to your respondents. Importance questions can also help you make business critical decisions.

**Example of an Open Format Question**

| |
|---|
| Students' involvement in classroom is: |
| *(i) Extremely Important (ii) Very Important (iii) Somewhat Important*<br>*(ii) Not very Important (v) Not at all Important* |

### 5. Likert Questions

Likert questions can help you ascertain how strongly your respondent agrees with a particular statement. Likert questions can also help you assess how your customers feel towards a certain issue, product or service.

**Example of an Open Format Question**

| |
|---|
| Lecture method is suitable for all subjects<br>(1)Strongly Agree (2) Agree (3) Undecided (4) Disagree (5) Strongly Disagree |

### 6. Dichotomous Questions

Dichotomous questions are simple questions that ask respondents to just answer yes or no. One major drawback of a dichotomous question is that it cannot analyze any of the answers between yes and no.

**Example of an Open Format Question**

| |
|---|
| Do you like smoking?<br>Yes No |

### 7. Bipolar Questions

Bipolar questions are questions that have two extreme answers. The respondent is asked to mark his/her responses between the two opposite ends of the scale.

**Example of an Open Format Question**

| |
|---|
| How would you describe the services of Allama Iqbal Open University?<br>Efficient…… ……. ……. ……. ……. …….. Inefficient<br>Fast …….. …….. ……. …… ……. …….. Slow |

### 8. Rating Scale Questions

In rating scale questions, the respondent is asked to rate a particular issue on a scale that ranges from poor to good. Rating scale questions usually have an even number of choices, so that respondents are not given the choice of a middle option.

**Example of an Open Format Question**

| How would you rate the quality of lecture method? |
|---|
| Good Fair Poor Very poor |

**9. Buying Propensity Questions**

Buying propensity questions are questions that try to assess the future intentions of customers. These questions ask respondents if they want to buy a particular product, what requirements they want to be addressed and whether they would buy such a product in the future.

**Example of an Open Format Question**

| Pakistani products have the good quality, would you prefer to buy it? |
|---|
| Definitely Probably Not Probably Not Sure Definitely Not |

**(ii) FORMATTING THE QUESTIONNAIRE**

As the questions are determined, a series of decisions must be made about the questionnaire format: its appearance, length, and order of questions. The questionnaire must be pleasing to look at and easy to complete.

The following guidelines may help in formatting the questionnaire.

- Begin with an introduction which includes the questionnaire's purpose, who is conducting it, to what use the information will go, and confidentiality. In mailed questionnaires, reinforce points that were made in the cover letter.

- Make the first questions interesting. Make them clearly related and useful to the topic of the questionnaire. The beginning questions should not be open-ended or questions with a long list of answer choices.

- Put the more important questions at the beginning.

- Arrange the order of questions to achieve continuity and a natural flow. Try to keep all questions on one subject together. Put the more general questions first, followed by a more specific question. For example, if you want to find out about a person's knowledge of insurance, start with questions about types of insurance, purpose of the different types, followed by questions about costs of these various types.

- Try to use the same type of question/responses throughout a particular train of thought. It breaks the attention span to have a multiple choice question following a YES/NO question, then an open-ended question.

- Place demographic questions (age, gender, race/ethnicity, etc.) in the beginning of the questionnaire.

- Use quality print in an easy-to-read type face. Allow sufficient open space to let the respondent feel it is not crowded and hard to read.

- Keep the whole question and its answers on the same page. Don't cause respondents to turn a page in the middle of a question or between the question and its answers.

- Be sure that the question is distinguishable from the instructions and the answers. May be put the instructions in boldface or italics.

- Try to arrange questions and answers in a vertical flow. This way, the respondent moves easily down the page, instead of side to side.
-
- Give directions on how to answer. Specific instructions may include: (Circle the number of your choice.) (Circle only one.) (Check all that apply.) (Please fill in the blank.) (Enter whole numbers.) (Please do not use decimals or fractions.)

**(iii) Advantages of the Questionnaires**

The main advantage of using questionnaires is that a large number of people can be reached relatively easily and economically. A standard questionnaire provides quantifiable answers for a research topic. These answers are relatively easy to analyze.

Questionnaires can be designed to target a certain "audience even if they are geographically spread." Depending on the design of questionnaires, the data collected may be either quantitative or qualitative. Quantitative data is in numerical form and can be used to find answers about a particular problem such as: customers' perceptions about certain products, feelings about services being offered by "Call Centers", and so on. Another good thing about questionnaires is that they "reduce bias".

Effective questionnaires may be designed in such a way that the questions are "short and focused" and have at least less than "12 words" (Marshall, 2004, p. 132).

**(iv) Disadvantages**

Questionnaires are not always the best way to gather information. For example, if there is little previous information on a problem, a questionnaire may only provide limited additional insight. On one hand, the investigators may not have asked the right questions which allow new insight in the research topic. On the other hand, questions often only allow a limited choice of responses. If the right response is not among the choice of answers, the investigators will obtain little or no valid information.

Another setback of questionnaires is the varying responses to questions. Respondents sometimes misunderstand or misinterpret questions. If this is the case, it will be very hard to correct these mistakes and collect missing data in a second round.

| **Activity 3.7:** | Prepare a five point scale questionnaire to rank the problems of elementary school teachers of rural areas. |
|---|---|

### 3.2.2 Observation

An observation is information about objects, events, moves, attitudes and phenomena using directly one or more senses. Observation can be defined as the visual study of something or someone in order to gain information or learn about behaviour, trends, or changes. This then allows us to make informed decisions, adjustments, and allowances based on what has been studied. Observation is a basic but important aspect of learning from and interacting with our environment. Observation is an important part of learning how to teach. Much of what beginner teachers need to be aware of cannot be learned solely in the class. Therefore classroom observation presents an opportunity to see real-life teachers in real-life teaching situations. In their reflections, many of our teacher friends mention their observations and how these observations influence the way they plan and teach. Teachers are forever reflecting and making decisions, and when they see someone else in action, in as much as they are seeing someone else, they are almost simultaneously seeing themselves. This means that observation is important at every stage of a teacher's career. Overall classroom observation is form of ongoing assessment. Most teachers can "read" their students; observing when they are bored, frustrated, excited, motivated, etc. As a teacher picks up these cues, he/she can adjust the instruction accordingly. It is also beneficial for teachers to make observational notes (referred to as anecdotal notes). These notes serve to document and describe student learning relative to concept development, reading, social interaction, and communication skill.

**(a) Classroom Observation Guidelines**

- To make useful observations in a child care program, the observer needs to be respectful of the program's needs to operate effectively. meeting the following guidelines will help:
- Observers should not interfere with the child care program's activities in any way while making observations.
- Observers may sit in a chair so a standing adult observer does not intimidate the children. Do not sit on other furniture such as shelves, tables, the children's chairs near an activity table or on play equipment.
- Refrain from talking with other observers, with the caregivers or the children while in the child care area. Take notes on a pad to help

remember what you have seen and frame questions you can ask of the director later.

- Acknowledge children if they approach you, but do not otherwise take part in the activities of the children. You can tell them you are watching them play today, or that you have to finish your work.
- Keep your personal possessions with you at all times unless you are given a safe place to leave them in the facility. Do not allow children to have access to your things.
- Treat all you see and hear as confidential. Do not repeat anything about the adults, children or facility that could be traced back to your observation

**(b) Purposes of Classroom Observation**

Classroom observation has many valid and important educational purposes. This three important purposes or areas where systematic classroom observation has been widely used:

- Description of instructional practices.
- Investigation of instructional inequities for different groups of students.
- Improvement of teachers' classroom instruction based on feedback from individual classroom or school profiles.

**(c ) Advantage and Disadvantage of Observation**

**(i) Advantage:**

- Data gathered can be highly reliable.
- The analyst is able to see what is being done.
- Observation is less expensive compared to other technique.
- It is useful when the subject cannot provide information.
- It helps to make appropriate decision about students personality.

**(ii) Disadvantages:**

- People feel uncomfortable being watched, they may perform differently when being observed.
- The work being observed may not involved the level of difficulty or volume normally experienced during that time period.

- Some activities may take place at odd times, it might be inconvenience for the system analyst.
- The task being observed is subjected to types of interruptions.
- Some task may not be in the manner in which they are observed.

Sometimes people act temporarily and perform their job correctly when they are being observed they might actually violate the standard of manner.

| **Activity 3.8:** | Prepare and conduct a classroom observation focusing on different teaching competencies of your classroom teacher, after collecting the data to analyze the teachers performance in different subjects. |
|---|---|

### 3.2.3 Interview

A conversation in which one person (the interviewer) elicits information from another person (the subject or interviewee). A transcript or account of such a conversation is also called an interview.

#### (a) Objectives of Interview

1. Collecting the data – both extensively and intensively.
2. Exchanging the data and also the experience

#### (b) Importance of Interview

Interview is important for the interviewer and the interviewee. Its importance may be analyzed through following points:

- An interview first helps the interviewer to analyze the communication skill of the candidate.
- Through oral interview the applicants' communication standards can be assessed. The oral response of the candidate also helps the interviewer to analyze the social behavior of the candidate. Additional information's can also be collected through interviews. The candidate's attitude and mind can be assessed only by such oral interviews.
- An interview helps the interviewer to assess the knowledge of the applicant. quires related with the job requirements; education and technical aspects will assist the interviewer to take a decision on the candidate upon his subject and technical knowledge.
- The expectation of the interviewer and the candidate can be freely discussed only through interviews.

- Interview is very important in helping the interviewer to choose the right candidate for their organization.
- An interview gives you insight on what the person you are interviewing thinks, or appears to be thinking.

Hence every interview should be taken seriously and all things that went unattended during the interview must be corrected. Because interview helps to collect different information. Sensitive topics which people may feel uncomfortable discussing in a focus group that can be taken through interview.

**(c) Types of Interview**

**1. Structured Interview**

Here, every single detail of the interview is decided in advance. The questions to be asked, the order in which the questions will be asked, the time given to each candidate, the information to be collected from each candidate, etc. is all decided in advance. Structured interview is also called Standardized, Patterned, Directed or Guided interview. Structured interviews are preplanned. They are accurate and precise. All the interviews will be uniform (same). Therefore, there will be consistency and minimum bias in structured interviews.

**2. Unstructured Interview**

This interview is not planned in detail. Hence it is also called as **Non-Directed** interview. The question to be asked, the information to be collected from the candidates, etc. are not decided in advance. These interviews are non-planned and therefore, more flexible. Candidates are more relaxed in such interviews. They are encouraged to express themselves about different subjects, based on their expectations, motivations, background, interests, etc. Here the interviewer can make a better judgment of the candidate's personality, potentials, strengths and weaknesses. However, if the interviewer is not efficient then the discussions will lose direction and the interview will be a waste of time and effort.

**3. Group Interview**

Here, all the candidates or small groups of candidates are interviewed together. The time of the interviewer is saved. A group interview is similar to a group discussion. A topic is given to the group, and they are asked to discuss it. The interviewer carefully watches the candidates. He tries to find out which candidate influences others, who clarifies issues, who summarizes the discussion, who speaks effectively, etc. He tries to judge the behaviour of each candidate in a group situation.

**4. Exit Interview**

When an employee leaves the company, he is interviewed either by his immediate superior or by the Human Resource Development (HRD) manager. This interview is called an exit interview. Exit interview is taken to find out why the employee is leaving the company. Sometimes, the employee may be asked to withdraw his resignation by providing some incentives. Exit interviews are taken to create a good image of the company in the minds of the employees who are leaving the company. They help the company to make proper Human Resource Development (HRD) policies, to create a favourable work environment, to create employee loyalty and to reduce labour turnover.

### 5. Depth Interview

This is a semi-structured interview. The candidate has to give detailed information about his background, special interest, etc. He also has to give detailed information about his subject. Depth interview tries to find out if the candidate is an expert in his subject or not. Here, the interviewer must have a good understanding of human behaviour.

### 6. Stress Interview

The purpose of this interview is to find out how the candidate behaves in a stressful situation. That is, whether the candidate gets angry or gets confused or gets frightened or gets nervous or remains cool in a stressful situation. The candidate who keeps his cool in a stressful situation is selected for the stressful job. Here, the interviewer tries to create a stressful situation during the interview. This is done purposely by asking the candidate rapid questions, criticizing his answers, interrupting him repeatedly, etc. Then the behviour of the interviewee is observed and future educational planning based on his/her stress levels and handling of stress.

### 7. Individual Interview

This is a 'One-To-One' Interview. It is a verbal and visual interaction between two people, the interviewer and the candidate, for a particular purpose. The purpose of this interview is to match the candidate with the job. It is a two way communication.

### 8. Informal Interview

Informal interview is an oral interview which can be arranged at any place. Different questions are asked to collect the required information from the candidate. Specific rigid procedure is not followed. It is a friendly interview.

### 9. Formal Interview

Formal interview is held in a more formal atmosphere. The interviewer asks pre-planned questions. Formal interview is also called **planned** interview.

### 10. Panel Interview

Panel means a selection committee or interview committee that is appointed for interviewing the candidates. The panel may include three or five members. They ask questions to the candidates about different aspects. They give marks to each candidate. The final decision will be taken by all members collectively by rating the candidates. Panel interview is always better than an interview by one interviewer because in a panel interview, collective judgment is used for selecting suitable candidates.

### 11. Behavioral Interview

In a behavioural interview, the interviewer will ask you questions based on common situations of the job you are applying for. The logic behind the behavioral interview is that your future performance will be based on a past performance of a similar situation. You should expect questions that inquire about what you did when you were in some situation and how did you dealt with it. In a behavioral interview, the interviewer wants to see how you deal with certain problems and what you do to solve them.

### 12. Phone Interview

A phone interview may be for a position where the candidate is not local or for an initial prescreening call to see if they want to invite you in for an in-person interview. You may be asked typical questions or behavioural questions.

Most of the time you will schedule an appointment for a phone interview. If the interviewer calls unexpectedly, it's ok to ask them politely to schedule an appointment. On a phone interview, make sure your call waiting is turned off, you are in a quiet room, and you are not eating, drinking or chewing gum.

### (d) Advantages of Interview

- Very good technique for getting the information about the complex, emotionally laden subjects.
- Can be easily adapted to the ability of the person being interviewed.
- Yields a good percentage of returns.
- Yields perfect sample of the general population.
- Data collected by this method is likely to be more correct as compared to the other methods that are used to investigate issues in an in depth way for the data collection
- Discover how individuals think and feel about a topic and why they hold certain opinions
- Investigate the use, effectiveness and usefulness of particular library collections and services
- Inform decision making, strategic planning and resource allocation

- Sensitive topics which people may feel uncomfortable discussing in a focus group
- Add a human dimension to impersonal data
- Deepen understanding and explain statistical data.

**Disadvantages of Interview**

- Time consuming process.
- Involves high cost.
- Requires highly skilled interviewer.
- Requires more energy.
- May sometimes involve systematic errors.
- More confusing and a very complicated method.
- Different interviewers may understand and transcribe interviews in different ways.

| **Activity 3.9:** | Conduct an interview with your teachers regarding their jobs and find out the problems of teachers during their jobs. |
|---|---|

### 3.2.4 Rating Scale

A rating scale is a tool used for assessing the performance of tasks, skill levels, procedures, processes, qualities, quantities, or end products, such as reports, drawings, and computer programs. These are judged at a defined level within a stated range. Rating scales are similar to checklists except that they indicate the degree of accomplishment rather than just *yes* or *no*. Hence rating scale used to determine the degree to which the child exhibits a behaviour or the quality of that behavior; each trait is rated on a continuum, the observer decides where the child fits on the scale overall rating scale focuses on:

- Make a qualitative judgment about the extent to which a behavior is present
- Consist of a set of characteristics or qualities to be judged by using a systematic procedure
- **Numerical** and **graphic rating scales** are used most frequently

**Types of Rating Scales**

**Numerical Rating Scales:**

A sequence of numbers is assigned to descriptive Categories; the rater marks a number to indicate the degree to which a characteristic is present

**Graphic Rating Scales:**

A set of categories described at certain points along the line of a continuum; the rater can mark his or her judgment at any location on the line.

**(a) Advantages of Rating Scales:**

- Used for behaviours not easily measured by other means
- Quick and easy to complete
- User can apply knowledge about the child from other times
- Minimum of training required
- Easy to design using consistent descriptors (e.g., always, sometimes, rarely, or never)
- Can describe the child's steps toward understanding or mastery

**(b) Disadvantages**

- Highly subjective (rater error and bias are a common problem).
- Raters may rate a child on the basis of their previous interactions or on an emotional, rather than an objective, basis.
- Ambiguous terms make them unreliable: raters are likely to mark characteristics by using different interpretations of the ratings (e.g., do they all agree on what "sometimes" means?).

| **Activity 3.10:** | Prepare a rating scale on attributes of good teaching and administer it in your classroom for evaluating the performance of your teachers of different subjects. |
|---|---|

## 3.3 Standardized Testing

Standardized tests are tools designed to allow measure of student performance relative to all others taking the same test. A standardized **test** is a test that is administered and scored in a consistent, or "standard", manner. Standardized tests are designed in such a way that the questions, conditions for administering, scoring procedures, and interpretations are consistent and are administered and scored in a predetermined, standard manner. Any test in which the same test is given in the same manner to all test takers is a standardized test. Standardized tests need not be high-stakes tests, time-limited tests, or multiple-choice tests. The opposite of a standardized test is a *non-standardized test*. Non-standardized testing gives significantly different tests to different test takers, or

gives the same test under significantly different conditions (e.g., one group is permitted far less time to complete the test than the next group), or evaluates them differently (e.g., the same answer is counted right for one student, but wrong for another student).

Standardized testing has received criticism from psychologists, educators and parents. Criticism of academic testing often focuses on linguistic biases against minorities, the testing methods that may not work for all types of students and negative reinforcement of lower performing students.

### (a) Types of Standardized Testing

There are two types of standardized tests: Norm-referenced and Criterion referenced. Norm-referenced testing measures performance relative to all other students taking the same test. It lets you know how well a student did compare to the rest of the testing population. For example, if a student is ranked in the 86th percentile, that means he/she did better than 86 percent of others who took the test. This type of testing is the most common found among standardized testing. Criterion referenced testing measures factual knowledge of a defined body of material. Multiple-choice tests that people take to get their license or a test in fractions are both examples of this type of testing.

In addition to the two main categories of standardized tests, these tests can be divided even further into performance tests or aptitude tests. Performance tests are assessments of what learning has already occurred in a particular subject area, while aptitude tests are assessments of abilities or skills considered important to future success in school.

Intelligence tests are also standardized tests that aim to determine how a person can handle problem solving using higher level cognitive thinking. Often just called an IQ test for common use, a typical IQ test asks problems involving pattern recognition and logical reasoning. It then takes into account the time needed and how many questions the person completes correctly, with penalties for guessing. Specific tests and how the results are used change from district to district but intelligence testing is common during the early years of schooling.

### (b) Advantages

- It can be obtained easily and available on researcher's convenience.
- It can be adopted and implemented quickly.
- It reduces or eliminates faculty time demands in instrument development and grading.
- It helps to score objectively.
- It can provide the external validity of test.
- It helps to provide reference group measures.
- It can make longitudinal comparisons.

• It can test large numbers of students.

**(c) Disadvantages**

- It measures relatively superficial knowledge or learning.
- Norm-referenced data may be less useful than criterion-referenced.
- It may be cost prohibitive to administer as a pre- and post-test.
- It is more summative than formative (may be difficult to isolate what changes are needed).
- It may be difficult to receive results in a timely manner.

**(d) Recommendations**

- It must be selected carefully based on faculty review and determination of match between test content and curriculum content.
- Request technical manual and information on reliability and validity from publisher.
- Check with other users.
- If possible, purchase data disk for creation of customized reports.
- If possible, select tests that also provide criterion-referenced results.
- Check results against those obtained from other assessment methods.
- Embedding the test as part of a course's requirements may improve student motivation.

**Activity 3.11:** Download a standardized test for measuring the achievements of elementary students in English language and administer it in your school. After administering it to analyze and interpret the score and explore the deficiency of students in different aspects of English language.

## 3.4 Summary

Classroom assessment test and techniques are a series of tools and practices designed to give teachers accurate information about the quality of student learning. Information gathered isn't used for grading or teacher evaluation. Instead, it's used to facilitate dialogue between students and teacher on the quality of the learning process, and how to improve it. For this purpose there are many different types and techniques of testing that can be done during an evaluation. They can be done by our school system or independently. Keeping in view the learning domains or aspects different tests such as

achievement tests, aptitude tests, attitude scale, intelligence tests, personality tests, norm and criterion tests and assessment techniques such as questionnaire, interview, observation, rating scale and standardized testing were discussed.

## 3.5 Self Assessment Questions

1. Discuss the nature of tests and techniques. Also highlight their characteristics in teaching learning process.
2. Categories the functions of different tests and techniques. To what extend these functions are fulfilled in our schools? Discuss.
3. Enlist the different types of tests and their role in education system.
4. Enlist the different types of techniques and their role in education system.
5. Enlist the advantages and disadvantages of different tests and techniques. Also give suggestions for their improvements.

## 3.6 References/Suggested Readings


- Airasian, P. (1994) "Classroom Assessment," Second Edition, NY" McGraw-Hill.
- American Psychological Association. (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.
- Anastasi, A. (1988). *Psychological Testing* (6th ed.). New York, NY: MacMillan Publishing Company.
- Cangelosi, J. (1990) "Designing Tests for Evaluating Student Achievement." NY: Addison-Wesley.
- Cunningham, G.K. (1998). Assessment in the Classroom. Bristol, PA: Falmer Press.
  Ward, A.W., & Murray-Ward, M. (1999). Assessment in the Classroom. Belmont, CA: Wadsworth Publishing Co.
- Gronlund, N. (1993) "How to Make Achievement Tests and Assessments," 5th Edition, NY: Allyn and Bacon.
- Gronlund, N. E. & Linn, R. L. (1995). *Measurement and Assessment in Teaching*. New Delhi: Baba Barkha Nath Printers.
- Haladyna, T.M. & Downing, S.M. (1989) Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. "Applied Measurement in Education," 2(1), 51-78.
- Monahan, T. (1998) The Rise of Standardized Educational Testing in the U.S. – A Bibliographic Overview.
- Ravitch, Diane, "The Uses and Misuses of Tests", in *The Schools We Deserve* (New York: Basic Books, 1985), pp. 172–181.
- Thissen, D., & Wainer, H. (2001). Test Scoring. Mahwah, NJ: Erlbaum.
- Wilson, N. (1997) Educational Standards and the Problem of Error. Education Policy Analysis Archives, Vol 6 No 10
- http://www.medterms.com/script/main/art.asp?articlekey=11519
- http://www.minddisorders.com/Flu-Inv/Intelligence-tests.html
- http://learningdisabilities.about.com/od/glossar1/a/intelligencetes.htm
- http://www.cliffsnotes.com/study_guide/Intelligence-Tests.topicArticleId-25438,articleId-25413.html

**UNIT: 4**

# TYPES OF TESTS

**Written By:**
**Dr. Naveed Sultana**

**Reviewed By:**
**Dr. Muhammad Tanveer Afzal**

## CONTENT

***Sr. No*** ***Topic*** ***Page No***

# INTRODUCTION

Classroom tests play a central role in the assessment of student learning. Tests provide relevant measures of many important learning outcomes and indirect evidence concerning others. They make expected learning outcomes explicit to students and parents and show what types of performance are valued. The validity of the information they provide, however, depends on the care that goes into the planning and preparation of tests. The main goal of classroom testing is to obtain valid, reliable and useful information concerning assessment. This requires determining what is to be measured and then defining it precisely so that tasks that evoke the desired performance can be constructed. In a standard based approach to education and training, informed by Constructivist theory, assessment informed instruction is the expectation as is continuous improvement. One of the most widely used tools in assessment and evaluation is the traditional or classic classroom achievement test, whether the classroom is on- or offline. These measures are often fraught with reliability and validity problems as the process for constructing such tests is often not followed or misunderstood, thereby introducing significant measurement error into the measurement process. Poor measurement frequently leads to inaccurate data-based inferences, which in turn leads to bad decision-making. Moreover classroom tests and assessment can be used for a variety of instructional purposes such examining the quality of teaching learning process, students achievement individually and success of institution overall. So in this unit we will examine the test item type and item format, writing select response items (multiple-choice, true/false, matching, completion and short-answer) and supply response items (brief and extended response). Each type of test item has its own unique characteristics, uses, advantages, limitations and rules for construction, which will be elaborated in this unit.

# OBJECTIVES

After reading this unit, you will be able to:

- define the nature of selection and supply type time items.
- examine the role, advantages and disadvantages of different types of objective and subjective type tests for measuring the students' achievement.
- describe the learning outcomes that are best measured with selection and supply test items.
- differentiate the characteristics of all types of selection and supply categories of items concentrating to measure the higher level of thinking of students.

## 4.1 Selection Type Items (objective type)

There are four types of test items in selection category of test which are in common use today. They are multiple-choice, matching, true-false, and completion items.

### 4.1.1 Multiple Choice Questions

Multiple-choice test items consist of a stem or a question and three or more alternative answers (options) with the correct answer sometimes called the keyed response and the incorrect answers called distracters. This form is generally better than the incomplete stem because it is simpler and more natural.

Grounlund (1995) writes that the multiple choice question is probably the most popular as well as the most widely applicable and effective type of objective test. Student selects a single response from a list of options. It can be used effectively for any level of course outcome. It consists of two parts: the stem, which states the problem and a list of three to five alternatives, one of which is the correct (key) answer and the others are distracters (incorrect options that draw the less knowledgeable pupil away from the correct response). Multiple choice questions consist of three obligatory parts:

1. The question ("body of the question")
2. The correct answer ("the key of the question")
3. Several incorrect alternatives (the so called "distracters") and optional (and especially valuable in self-assessment)
4. Feedback comment on the student's answer.

The stem may be stated as a direct question or as an incomplete statement. For example:

**Direct question**

Which is the capital city of Pakistan? ---------- (Stem)

A. Paris. ---------- (Distracter)
B. Lisbon. ---------- (Distracter)
C. Islamabad. ---------- **(Key)**
D. Rome. ---------- (Distracter)

**Incomplete Statement**

The capital city of Pakistan is

A. Paris.
B. Lisbon.
C. Islamabad.
D. Rome.

Multiple choice questions are composed of one question with multiple possible answers (options), including the correct answer and several incorrect answers (distracters). Typically, students select the correct answer by circling the associated number or letter, or filling in the associated circle on the machine-readable response sheet. Students can generally respond to these types of questions quite quickly. As a result, they are often used to test student's knowledge of a broad range of content. Creating these questions can be time consuming because it is often difficult to generate several plausible distracters. However, they can be marked very quickly.

**Multiple Choice Questions Good for**:

- Application, synthesis, analysis, and evaluation levels

RULES FOR WRITING MULTIPLE-CHOICE QUESTIONS

There are several rules we can follow to improve the quality of this type of written examination.

**1. Examine only the Important Facts!**

Make sure that every question examines only the important knowledge. Avoid detailed questions - each question has to be relevant for the previously set instructional goals of the course.

**2. Use Simple Language!**

Use simple language, taking care of spelling and grammar. Spelling and grammar mistakes (unless you are testing spelling or grammar) only confuse students. Remember that you are examining knowledge about your subject and not language skills.

**3. Make the Questions Brief and Clear!**

Clear the text of the body of the question from all superfluous words and irrelevant content. It helps students to understand exactly what is expected of them. It is desirable to formulate a question in such way that the main part of the text is in the body of the question, without being repeated in the answers.

**4. Form the Questions Correctly!**

Be careful that the formulation of the question does not (indirectly) hide the key to the correct answer. Student (adept at solving tests) will be able to recognize it easily and will find the right answer because of the word combination, grammar etc, and not because of their real knowledge.

**5. Take into Consideration the Independence of Questions!**

Be careful not to repeat content and terms related to the same theme, since the answer to one question can become the key to solve another.

**6. Offer Uniform Answers!**

All offered answers should be unified, clear and realistic. For example, unlikely realisation of an answer or uneven text quantity of different answers can point to the right answer. Such a question does not test real knowledge. The position of the key should be random. If the answers are numbers, they should be listed in an ascending order.

**7. Avoid Asking Negative Questions!**

If you use negative questions, negation must be emphasized by using CAPITAL letters, e.g. "Which of the following IS NOT correct..." or "All of the following statements are true, EXCEPT...".

**8. Avoid Distracters in the Form of "All the answers are correct" or "None of the Answers is Correct"!**

Teachers use these statements most frequently when they run out of ideas for distracters. Students, knowing what is behind such questions, are rarely misled by it. Therefore, if you do use such statements, sometimes use them as the key answer. Furthermore, if a student recognizes that there are two correct answers (out of 5 options), they will be able to conclude that the key answer is the statement "all the answers are correct", without knowing the accuracy of the other distracters.

**9. Distracters must be Significantly Different from the Right Answer (key)!**

Distracters which only slightly differ from the key answer are bad distracters. Good or strong distracters are statements which themselves seem correct, but are not the correct answer to a particular question.

**10. Offer an Appropriate Numbers of Distracters.**

The greater the number of distracters, the lesser the possibility that a student could guess the right answer (key). In higher education tests questions with 5 answers are used most often (1 key + 4 distracters). That means that a student is 20% likely to guess the right answer.

**Advantages:**

Multiple-choice test items are not a panacea. They have advantages and advantages just as any other type of test item. Teachers need to be aware of these characteristics in order to use multiple-choice items effectively.

**Advantages**

**Versatility**

Multiple-choice test items are appropriate for use in many different subject-matter areas, and can be used to measure a great variety of educational objectives. They are adaptable to various levels of learning outcomes, from simple recall of knowledge to more complex levels, such as the student's ability to:

- Analyze phenomena
- Apply principles to new situations
- Comprehend concepts and principles
- Discriminate between fact and opinion
- Interpret cause-and-effect relationships
- Interpret charts and graphs
- Judge the relevance of information
- Make inferences from given data
- Solve problems

The difficulty of multiple-choice items can be controlled by changing the alternatives, since the more homogeneous the alternatives, the finer the distinction the students must make in order to identify the correct answer. Multiple-choice items are amenable to item analysis, which enables the teacher to improve the item by replacing distracters that are not functioning properly. In addition, the distracters chosen by the student may be used to diagnose misconceptions of the student or weaknesses in the teacher's instruction.

**Validity**

In general, it takes much longer to respond to an essay test question than it does to respond to a multiple-choice test item, since the composing and recording of an essay answer is such a slow process. A student is therefore able to answer many multiple-choice items in time it would take to answer a single essay question. This feature enables the teacher using multiple-choice items to test a broader sample of course contents in a given amount of testing time. Consequently, the test scores will likely be more representative of the students' overall achievement in the course.

**Reliability**

Well-written multiple-choice test items compare favourably with other test item types on the issue of reliability. They are less susceptible to guessing than are true-false test items, and therefore capable of producing more reliable scores. Their scoring is more clear-cut than short answer test item scoring because there are no misspelled or partial answers to deal with. Since multiple-choice items are objectively scored, they are not affected by scorer inconsistencies as are essay questions, and they are essentially immune to the influence of bluffing and writing ability factors, both of which can lower the reliability of essay test scores.

**Efficiency**

Multiple-choice items are amenable to rapid scoring, which is often done by scoring machines. This expedites the reporting of test results to the student so that any follow-up clarification of instruction may be done before the course has proceeded much further. Essay questions, on the other hand, must be graded manually, one at a time. Overall multiple choice tests are:

- Very effective
- Versatile at all levels
- Minimum of writing for student
- Guessing reduced
- Can cover broad range of content

**Disadvantages**

**Versatility**

Since the student selects a response from a list of alternatives rather than supplying or constructing a response, multiple-choice test items are not adaptable to measuring certain learning outcomes, such as the student's ability to:

- Articulate explanations
- Display thought processes
- Furnish information
- Organize personal thoughts.
- Perform a specific task
- Produce original ideas
- Provide examples

Such learning outcomes are better measured by short answer or essay questions, or by performance tests.

**Reliability**

Although they are less susceptible to guessing than are true false-test items, multiple-choice items are still affected to a certain extent. This guessing factor reduces the reliability of multiple-choice item scores somewhat, but increasing the number of items on the test offsets this reduction in reliability.

**Difficulty of Construction**

Good multiple-choice test items are generally more difficult and time-consuming to write than other types of test items. Coming up with plausible distracters requires a certain amount of skill. This skill, however, may be increased through study, practice, and experience.

Gronlund (1995) writes that multiple-choice items are difficult to construct. Suitable distracters are often hard to come by and the teacher is tempted to fill the void with a "junk" response. The effect of narrowing the range of options will available to the test wise student. They are also exceedingly time consuming to fashion, one hour per question being by no means the exception. Finally multiple-choice items generally take student longer to complete (especially items containing fine discrimination) than do other types of objective question.

- Difficult to construct good test items.
- Difficult to come up with plausible distracters/alternative responses.

| | |
|---|---|
| **Activity 4.1:** | Construct two items of direct question and two items of incomplete statement while following the rules of multiple items. |

### 4.1.2 True/False Questions

A True-False test item requires the student to determine whether a statement is true or false. The chief disadvantage of this type is the opportunity for successful guessing.

According to Gronlund (1995) the alternative response test items that consists of a declaration statement that the pupil is asked to mark true or false, right or wrong, correct or incorrect, yes or no, fact or opinion, agree or disagree and the like. In each case there are only two possible answers. Because the true-false option is the most common, this type is mostly refers to true-false type. Students make a designation about the validity of the statement. Also known as a "binary-choice" item because there are only two options to select from. These types of items are more effective for assessing knowledge, comprehension, and application outcomes as defined in the cognitive domain of Blooms' Taxonomy of educational objectives.

**Example**

Directions: Circle the correct response to the following statements.

1. Allama Iqbal is the founder of Pakistan. T/F
2. Democracy system is for the people. T/F
3. Quaid-e-Azam was the first Prime Minister of Pakistan. T/F

**Good for:**

- Knowledge level content
- Evaluating student understanding of popular misconceptions
- Concepts with two logical responses

**Advantages:**

- Easily assess verbal knowledge
- Each item contains only two possible answers
- Easy to construct for the teacher
- Easy to score for the examiner
- Helpful for poor students
- Can test large amounts of content
- Students can answer 3-4 questions per minute

**Disadvantages:**

- They are easy to construct.
- It is difficult to discriminate between students that know the material and students who don't know.
- Students have a 50-50 chance of getting the right answer by guessing.
- Need a large number of items for high reliability.
- Fifty percent guessing factor.
- Assess lower order thinking skills.
- Poor representative of students learning achievement.

**Tips for Writing Good True/False items**:

- Avoid double negatives.
- Avoid long/complex sentences.
- Use specific determinants with caution: never, only, all, none, always, could, might, can, may, sometimes, generally, some, few.

- Use only one central idea in each item.
- Don't emphasize the trivial.
- Use exact quantitative language
- Don't lift items straight from the book.
- Make more false than true (60/40). (Students are more likely to answer true.)
- The desired method of marking true or false should be clearly explained before students begin the test.
- Construct statements that are definitely true or definitely false, without additional qualifications. If opinion is used, attribute it to some source.

**Avoid the following:**

a. verbal clauses, absolutes, and complex sentences;

b. broad general statements that are usually not true or false without further qualifications;

c. terms denoting indefinite degree (e.g., large, long time, or regularly) or absolutes (e.g., never, only, or always).

d. placing items in a systematic order (e.g., TTFF, TFTF, and so on);

e. taking statements directly from the text and presenting them out of context.

| **Activity 4.2:** | Enlist five items by indicating them T/F (True & False) |
|---|---|

### 4.1.3 Matching items

According to Cunningham (1998), the matching items consist of two parallel columns. The column on the left contains the questions to be answered, termed premises; the column on the right, the answers, termed responses. The student is asked to associate each premise with a response to form a matching pair.

**For example;**

| Column "A" Capital City | Column "B" Country |
|---|---|
| Islamabad | Iran |
| Tehran | Spain |
| Istanbul | Portugal |
| Madrid | Pakistan |
| Jaddah | Netherlands |
| | Turkey |
| | West Germany |

Matching test items are used to test a student's ability to recognize relationships and to make associations between terms, parts, words, phrases, clauses, or symbols in one column with related alternatives in another column. When using this form of test item, it is a good practice to provide alternatives in the response column that are used more than once, or not at all, to preclude guessing by elimination. Matching test items may have either an equal or unequal number of selections in each column.

Matching-Equal Columns. When using this form, providing for some items in the response column to be used more than once, or not at all, can preclude guessing by elimination.

**Good for:**

- Knowledge level
- Some comprehension level, if appropriately constructed

**Types:**

- Terms with definitions
- Phrases with other phrases
- Causes with effects
- Parts with larger units
- Problems with solutions

**Advantages:**

The chief advantage of matching exercises is that a good deal of factual information can be tested in minimal time, making the tests compact and efficient. They are especially well suited to who, what, when and where types of subject matter. Further students frequently find the tests fun to take because they have puzzle qualities to them.

- Maximum coverage at knowledge level in a minimum amount of space/prep time
- Valuable in content areas that have a lot of facts

**Disadvantages:**

The principal difficulty with matching exercises is that teachers often find that the subject matter is insufficient in quantity or not well suited for matching terms. An exercise should be confined to homogeneous items containing one type of subject matter (for instance, authors-novels; inventions inventors; major events-dates terms – definitions; rules examples and the like). Where unlike clusters of questions are used to adopt but poorly informed student can often recognize the ill-fitting items by their irrelevant and extraneous nature (for instance, in a list of authors the inclusion of the names of capital cities).

Student identifies connected items from two lists. It is useful for assessing the ability to discriminate, categorize, and association amongst similar concepts.

- Time consuming for students
- Not good for higher levels of learning

**Tips for Writing Good Matching items:**

Here are some suggestions for writing matching items:

- Keep both the list of descriptions and the list of options fairly short and homogeneous – they should both fit on the same page. Title the lists to ensure homogeneity and arrange the descriptions and options in some logical order. If this is impossible you're probably including too wide a variety in the exercise. Try constructing two or more exercises.
- Make sure that all the options are plausible distracters for each description to ensure homogeneity of lists.
- The list of descriptions on the left side should contain the longer phrases or statements, whereas the options on the right side should consist of short phrases, words or symbols.
- Each description in the list should be numbered (each is an item), and the list of options should be identified by letter.

- Include more options than descriptions. If the option list is longer than the description list, it is harder for students to eliminate options. If the option list is shorter, some options must be used more than once. Always include some options that do not match any of the descriptions, or some that match more than one, or both.
- In the directions, specify the basis for matching and whether options can be used more than once.
- Need 15 items or less.
- Give good directions on basis for matching.
- Use items in response column more than once (reduces the effects of guessing).
- Make all responses plausible.
- Put all items on a single page.
- Put response in some logical order (chronological, alphabetical, etc.).

| | |
|---|---|
| **Activity 4.3:** | Keeping in view the nature of matching items, construct at least five items of matching case about any topic. |

### 4.1.4 Completion Items

Like true-false items, completion items are relatively easy to write. Perhaps the first tests classroom teachers' construct and students take completion tests. Like items of all other formats, though, there are good and poor completion items. Student fills in one or more blanks in a statement. These are also known as "Gap-Fillers." Most effective for assessing knowledge and comprehension learning outcomes but can be written for higher level outcomes. e.g.

The capital city of Pakistan is ----------.

**Suggestions for Writing Completion or Supply Items**

Here are our suggestions for writing completion or supply items:

I. If at all possible, items should require a single-word answer or a brief and definite statement. Avoid statements that are so indefinite that they may be logically answered by several terms.

**a. Poor item:**

World War II ended in _____.

**b. Better item:**

World War II ended in the year _____.

II. Be sure the question or statement poses a problem to the examinee. A direct question is often more desirable than an incomplete statement because it provides more structure.

III. Be sure the answer that the student is required to produce is factually correct. Be sure the language used in the question is precise and accurate in relation to the subject matter area being tested.

IV. Omit only key words; don't eliminate so many elements that the sense of the content is impaired.

**a. Poor item:**

The _____ type of test item is usually more _____ than the _____ type.

**b. Better item:**

The supply type of test item is usually graded less objectively than the _____ type.

I. Word the statement such that the blank is near the end of the sentence rather than near the beginning. This will prevent awkward sentences.

II. If the problem requires a numerical answer, indicate the units in which it is to be expressed.

**Activity 4.3:** Construct five fill in the blanks about Pakistan.

## 4.2 Supply Type Items

The aviation instructor is able to determine the students' level of generalized knowledge of a subject through the use of supply-type questions. There are four types of test items in supply type category of test. Commonly these are completion items, short answers, restricted response and extended response (essay type comprises the restricted and extended responses).

### 4.2.1 Short Answer

Student supplies a response to a question that might consistent of a single word or phrase. Most effective for assessing knowledge and comprehension learning outcomes but can be written for higher level outcomes. Short answer items are of two types.

- Simple direct questions
  Who was the first president of the Pakistan?
- Completion items

The name of the first president of Pakistan is _____.

The items can be answered by a work, phrase, number or symbol. Short-answer tests are a cross between essay and objective tests. The student must supply the answer as with an essay question but in a highly abbreviated form as with an objective question.

**Good for:**

- Application, synthesis, analysis, and evaluation levels

**Advantages:**

- Easy to construct
- Good for "who," what," where," "when" content
- Minimizes guessing
- Encourages more intensive study-student must know the answer vs. recognizing the answer.

  Gronlund (1995) writes that short-answer items have a number of advantages.

- They reduce the likelihood that a student will guess the correct answer
- They are relatively easy for a teacher to construct.
- They are will adapted to mathematics, the sciences, and foreign languages where specific types of knowledge are tested (The formula for ordinary table salt is _____).
- They are consistent with the Socratic question and answer format frequently employed in the elementary grades in teaching basic skills.

**Disadvantages:**

- May overemphasize memorization of facts
- Take care - questions may have more than one correct answer
- Scoring is laborious

According to Grounlund (1995) there are also a number of disadvantages with short-answer items.

- They are limited to content areas in which a student's knowledge can be adequately portrayed by one or two words.
- They are more difficult to score than other types of objective-item tests since students invariably come up with unanticipated answers that are totally or partially correct.
- Short answer items usually provide little opportunity for students to synthesize, evaluate and apply information.

**Tips for Writing Good Short Answer Items:**

- When using with definitions: supply term, not the definition-for a better judge of student knowledge.
- For numbers, indicate the degree of precision/units expected.
- Use direct questions, not an incomplete statement.
- If you do use incomplete statements, don't use more than 2 blanks within an item.
- Arrange blanks to make scoring easy.
- Try to phrase question so there is only one answer possible.

**Activity 4.5:** Develop a test of short answers on democracy in Pakistan.

### 4.2.3 Essay

Essay questions are supply or constructed response type questions and 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. The student has to formulate and write a response, which may be detailed and lengthy. The accuracy and quality of the response are judged by the teacher.

Essay questions provide a complex prompt that requires written responses, which can vary in length from a couple of paragraphs to many pages. Like short answer questions, they provide students with an opportunity to explain their understanding and demonstrate creativity, but make it hard for students to arrive at an acceptable answer by bluffing. They can be constructed reasonably quickly and easily but marking these questions can be time-consuming and grade agreement can be difficult.

Essay questions differ from short answer questions in that the essay questions are less structured. This openness allows students to demonstrate that they can integrate the course material in creative ways. As a result, essays are a favoured approach to test higher levels of cognition including analysis, synthesis and evaluation. However, the requirement that the students provide most of the structure increases the amount of work required to respond effectively. Students often take longer time to compose a five paragraph essay than they would take to compose paragraph answer to short answer questions.

Essay items can vary from very lengthy, open ended end of semester term papers or take home tests that have flexible page limits (e.g. 10-12 pages, no more than 30 pages etc.) to essays with responses limited or restricted to one page or less. Essay questions are used both as formative assessments (in classrooms) and summative assessments (on standardized tests). There are 2 major categories of essay questions -- *short response* (also referred to as *restricted* or *brief*) and *extended response*.

- Restricted Response: more consistent scoring, outlines parameters of responses

- Extended Response Essay Items: synthesis and evaluation levels; a lot of freedom in answers

## A. Restricted Response Essay Items

An essay item that poses a specific problem for which a student must recall proper information, organize it in a suitable manner, derive a defensible conclusion, and express it within the limits of posed problem, or within a page or time limit, is called a restricted response essay type item. The statement of the problem specifies response limitations that guide the student in responding and provide evaluation criteria for scoring.

**Example 1:**

List the major similarities and differences in the lives of people living in Islamabad and Faisalabad.

**Example 2:**

Compare advantages and disadvantages of lecture teaching method and demonstration teaching method.

### When Should Restricted Response Essay Items be used?

Restricted Response Essay Items are usually used to:-

- Analyze relationship
- Compare and contrast positions
- State necessary assumptions
- Identify appropriate conclusions
- Explain cause-effect relationship
- Organize data to support a viewpoint
- Evaluate the quality and worth of an item or action
- Integrate data from several sources

## B. Extended Response Essay Type Items

An essay type item that allows the student to determine the length and complexity of response is called an extended-response essay item. This type of essay is most useful at the synthesis or evaluation levels of cognitive domain. We are interested in determining whether students can organize, integrate, express, and evaluate information, ideas, or pieces of knowledge the extended response items are used.

**Example:**

Identify as many different ways to generate electricity in Pakistan as you can? Give advantages and disadvantages of each. Your response will be graded on its accuracy, comprehension and practical ability. Your response should be 8-10 pages in length and it will be evaluated according to the RUBRIC (scoring criteria) already provided.

Over all Essay type items (both types restricted response and extended response) are

**Good for:**

- Application, synthesis and evaluation levels

**Types:**

- Extended response: synthesis and evaluation levels; a lot of freedom in answers
- Restricted response: more consistent scoring, outlines parameters of responses

**Advantages:**

- Students less likely to guess
- Easy to construct
- Stimulates more study
- Allows students to demonstrate ability to organize knowledge, express opinions, show originality.

**Disadvantages:**

- Can limit amount of material tested, therefore has decreased validity.
- Subjective, potentially unreliable scoring.
- Time consuming to score.

**Tips for Writing Good Essay Items:**

- Provide reasonable time limits for thinking and writing.
- Avoid letting them to answer a choice of questions (You won't get a good idea of the broadness of student achievement when they only answer a set of questions.)
- Give definitive task to student-compare, analyze, evaluate, etc.
- Use checklist point system to score with a model answer: write outline, determine how many points to assign to each part
- Score one question at a time-all at the same time.

| | |
|---|---|
| **Activity 4.6:** | Develop an essay type test on this unit while covering the levels of knowledge, application and analysis. |

## 4.3 Self Assessment Questions:

1. In an area in which you are teaching or plan to teach, identify several learning outcomes that can be best measured with objective and subjective types questions.
2. Criticize the different types of selection and supply categories. In your opinion which type is more appropriate for measuring the achievement level of elementary students?
3. What factors should be considered in deciding whether subjective or objective type questions should be included in a classroom tests?
4. Compare the functions of selection and supply types items.

## 4.4 References/Suggested Readings


Airasian, P. (1994) "Classroom Assessment," Second Edition, NY" McGraw-Hill.

American Psychological Association. (1985). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

Anastasi, A. (1988). Psychological Testing (6th ed.). New York, NY: MacMillan Publishing Company.

Cangelosi, J. (1990) "Designing Tests for Evaluating Student Achievement." NY: Addison-Wesley.

Cunningham, G.K. (1998). Assessment in the Classroom. Bristol, PA: Falmer Press. Ward, A.W., & Murray-Ward, M. (1999). Assessment in the Classroom. Belmont, CA: Wadsworth Publishing Co.

Gronlund, N. (1993) "How to Make Achievement Tests and Assessments," 5th Edition, NY: Allyn and Bacon.

Gronlund, N. E. & Linn, R. L. (1995). Measurement and Assessment in Teaching. New Delhi: Baba Barkha Nath Printers.

Haladyna, T.M. & Downing, S.M. (1989) Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. "Applied Measurement in Education," 2(1), 51-78.

Monahan, T. (1998) The Rise of Standardized Educational Testing in the U.S. – A Bibliographic Overview.

Ravitch, Diane, "The Uses and Misuses of Tests", in The Schools We Deserve (New York: Basic Books, 1985), pp. 172–181.

Thissen, D., & Wainer, H. (2001). Test Scoring. Mahwah, NJ: Erlbaum.

Wilson, N. (1997) Educational standards and the problem of error. Education Policy Analysis Archives, Vol 6 No 10

**UNIT–5**

# RELIABILITY OF THE ASSESSMENT TOOLS

**Written by:**
**Dr. Muhammad Tanveer Afzal**
**Reviewed by:**
**Prof. Dr. Rehana Masrur**

## CONTENT

***Sr. No*** ***Topic*** ***Page No***

# INTRODUCTION

Assessment is an integral part of teaching-learning process which allows teachers to evaluate their student's achievement during an educational course. Many teachers feel deficiency in preparing and grading exams, and most students are fearful of taking them. Yet test is a significant educational tool. Therefore, this tool must be reliable and valid in such a way that everyone has credibility on its results.

Every classroom assessment measure must be appropriately reliable and valid, whether, it is the routine classroom achievement test, attitudinal measure, or performance assessment. A measure must first be reliable before it can be valid.

Teachers have been designing achievement tests since decades. But before preparing a test a teacher or external exam designer must be aware of the qualities of an achievement test. A measure that ignores the basic principles of developing a test may produce such results that may be unacceptable for the students, and will not be measuring the actual performance.

Therefore this particular unit is meant for the prospective teachers addressing the concept and meaning of the reliability, its types, factors affecting reliability of the tests and the usability of the tests.

# OBJECTIVES

After studying this unit, prospective teachers will be able to:

- define reliability in their own words.
- apply the different methods of assuring reliability on the tests.
- identify the factors affecting reliability.
- construct a test and check how much reliable it is.
- identify measures for reducing the problems in conducting the tests.

## 5.1. Reliability

What does the term reliability mean? Reliability means Trustworthy. A test score is called reliable when we have reasons for believing the test score to be stable and objective. For example if the same test is given to two classes and is marked by different teachers even then it produced the similar results, it may be considered as reliable. Stability and trustworthiness depends upon the degree to which score is free of chance error. We must first build a conceptual bridge between the question asked by the individual (i.e. are my scores reliable?) and how reliability is measured scientifically. This bridge is not as simple as it may first appear. When a person thinks of reliability, many things may come into his mind – my friend is very reliable, my car is very reliable, my internet bill-paying process is very reliable, my client's performance is very reliable, and so on. The characteristics being addressed are the concepts such as consistency, dependability, predictability, variability etc. Note that implicit, reliability statements, is the behaviour, machine performance, data processes, and work performance may sometimes not reliable. The question is "how much the scores of tests vary over different observations?"

### 5.1.1 Some Definitions of Reliability:

**According to Merriam Webster Dictionary:**

"Reliability is the extent to which an experiment, test, or measuring procedure yields the same results on repeated trials."

**According to Hopkins & Antes (2000):**

"Reliability is the consistency of observations yielded over repeated recordings either for one subject or a set of subjects."

**Joppe (2000) defines reliability as:**

"…The extent to which results are consistent over time and an accurate representation of the total population under study is referred to as reliability and if the results of a study can be reproduced under a similar methodology, then the research instrument is considered to be reliable." (p. 1)

The more general definition of the reliability is: The degree to which a score is stable and consistent when measured at different times (test-retest reliability), in different ways (parallel-forms and alternate-forms), or with different items within the same scale (internal consistency).

## 5.2 Types of Reliability

Reliability is one of the most important elements of test quality. It has to do with the consistency, or reproducibility, of an examinee's performance in the test. It's not possible

to calculate reliability exactly. Instead, we have to estimate reliability, and this is always an imperfect attempt. Here, we introduce the major reliability estimators and talk about their strengths and weaknesses.

There are six *general classes of reliability estimates*, each of which estimates reliability in a different way. They are:

**i) Inter-Rater or Inter-Observer Reliability**

To assess the degree to which different raters/observers give consistent estimates of the same phenomenon. That is if two teachers mark same test and the results are similar, so it indicates the inter-rater or inter-observer reliability.

**ii) Test-Retest Reliability:**

To assess the consistency of a measure from one time to another, when a same test is administered twice and the results of both administrations are similar, this constitutes the test-retest reliability. Students may remember and may be mature after the first administration creates a problem for test-retest reliability.

**iii) Parallel-Form Reliability:**

To assess the consistency of the results of two tests constructed in the same way from the same content domain. Here the test designer tries to develop two tests of the similar kinds and after administration the results are similar then it will indicate the parallel form reliability.

**iv) Internal Consistency Reliability:**

To assess the consistency of results across items within a test, it is correlation of the individual items score with the entire test.

**v) Split half Reliability:**

To assess the consistency of results comparing two halves of single test, these halves may be even odd items on the single test.

**vi) Kuder-Richardson Reliability:**

To assess the consistency of the results using all the possible split halves of a test.

Let's discuss each of these in turn.

### 5.2.1. Inter-Rater or Inter-Observer Reliability

Whenever we observe or activities of humans, we have to think about the procedure for reliable and consistent results. For this two or more than two observers are assigned to observe the students or teachers. So how do we determine whether two observers are being consistent in their observations? We probably should establish inter-rater reliability by considering the similarity of the scores awarded by the two observers. After all, if we use data to establish reliability, and we find that reliability is low. We should have to focus upon the criteria established for the observation. And if it is tried first in the actual situation then it may help to develop the reasonable criteria for the observation, and may be more objective.

There are two major ways to actually estimate inter-rater reliability. If your measurement consists of categories -- the raters are checking off which category each observation falls in -- you can calculate the percent of agreement between the raters. For instance, let's say you had 100 observations that were being rated by two raters. For each observation, the rater could check one of three categories. Imagine that on 86 of the 100 observations, the raters checked the same category. In this case, the percent of agreement would be 86%. OK, it's a crude measure, but it does give an idea of how much agreement exists, and it works no matter how many categories are used for each observation.

The other major way to estimate inter-rater reliability is appropriate when the measure is a continuous one. There, all you need to do is calculate the correlation between the ratings of the two observers. For instance, they might be rating the overall level of activity in a classroom on a 1-to-7 scale. You could have them give their rating at regular time intervals (e.g., every 30 seconds). The correlation between these ratings would give you an estimate of the reliability or consistency between the raters.

One might think of this type of reliability as "calibrating" the observers. There are other things one could do to encourage reliability between observers, even without estimating it. For instance, in a psychiatric unit where every morning a nurse had to do a ten-item rating of each patient on the unit. Of course, it’s difficult to count on the same nurse being present every day, so there is a need to find a way to assure that any of the nurses would give comparable ratings. The way we did, it was to hold weekly "calibration" meetings where we would have all of the nurses ratings for several patients and discuss why they chose the specific values they did. If there were disagreements, the nurses would discuss them and attempt to come up with rules for deciding when they would give a "3" or a "4" for a rating on a specific item. Although this was not an estimate of reliability, it probably went a long way towards improving the reliability between raters.

| | |
|---|---|
| **Activity 5.1:** | Develop an essay type test for any class, administer it, get it marked from two raters and then compare the marks given by the two raters for each question. |

### 5.2.2. Test-Retest Reliability

Test-retest is a statistical method used to determine a test's reliability. The test is performed twice; in the case of a questionnaire, this would mean giving a group of participants the same questionnaire on two different occasions.

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

We estimate test-retest reliability when we administer the same test to the same sample on two different occasions. This approach assumes that there is no substantial change in the construct being measured between the two occasions. The amount of time allowed between measures is critical. We know that if we measure the same thing twice that the correlation between the two observations will depend in part by how much time elapses between the two measurement occasions. The shorter the time gap, the higher the correlation; the longer the time gap, the lower the correlation. This is because the two observations are related over time -- the closer in time we get the more similar the factors that contribute to error. Since this correlation is the test-retest estimate of reliability, you can obtain considerably different estimates depending on the interval.

| | |
|---|---|
| **Activity 5.2:** | Develop a test of English for sixth grade students, administer it twice with a gap of six weeks, find the relationship between the scores of students between 1st and 2nd administration. |

### 5.2.3. Split-Half Reliability

Suppose you have to develop a test of 30 items and you want to know that how reliable the test is? What you have to do is to administer the test, mark it and divide it in to two parts, in such a way that place all the even numbered items (2,4,6……) in one half and the odd numbered items (1,3,5……..) in the second. Calculate the reliability by using the Spearman-Brown prophecy formula given below.

Actually in split-half reliability we randomly divide all items that claim to measure the same contents into two sets. We administer the entire instrument to a sample of students and calculate the total score for each randomly divided half. The split-half reliability estimate is simply the correlation between these two total scores.

Normally a single test is used to make two shorter alternate forms. This method has the advantage that only one test administration is required, and therefore memory and the practice and maturation effects are not involved. Furthermore, it does not require two tests. So it has many advantages over parallel form and test-retest methods, therefore it is the most frequently used method of finding internal consistency of the classroom tests. The formula used for the reliability of the full test is Spearman-Brown prophecy formula as given below.

$$\text{Reliability of the Full Test} = \frac{2(\text{reliability of the half test})}{1+ (\text{reliability of the half test})}$$

### 5.2.4 Parallel-Form Reliability

In parallel form reliability we have to create two different tests from the same contents to measure the same learning outcomes. The easiest way to accomplish this is to write a large set of questions that address the same contents and then randomly divide the questions into two sets. Now it's time to administer both instruments to the same students at the same time. The correlation between the two parallel forms is the estimate of reliability. One major problem with this approach is that you have to be able to write lots of items that reflect the same contents. This is often no easy to do job. Furthermore, this approach makes the assumption that the randomly divided halves are parallel or equivalent. Even by chance, this will sometimes not be the case. The parallel forms approach is very similar to the split-half reliability described earlier. The major difference is that parallel forms are constructed so that the two forms can be used independent of each other and considered equivalent measures. For instance, we might be concerned about a testing threat to internal validity. If we use Form A for the pretest and Form B for the posttest, we minimize that problem. It would even be better if we randomly assign individuals to receive Form A or B on the pretest and then switch them on the posttest. With split-half reliability we have an instrument that we wish to use as a single measurement instrument and only develop randomly split halves for purposes of estimating reliability.

| | |
|---|---|
| **Activity 5.3:** | Make two tests of mathematics and compare its reliability through Parallel-Forms Reliability method. |

### 5.2.5. Internal Consistency Reliability

In internal consistency reliability estimation, we use our single test. The test is administered to a group of students on one occasion to estimate reliability. In effect we judge the reliability of the instrument by estimating how well the items that reflect the same content give similar results. We are looking at how consistent the results are for different items for the same construct within the measure. There are a wide variety of internal consistency measures that can be used.

### 5.2.6. Kuder Richardson Reliability

The estimates of internal consistency of the test are commonly calculated by using Kuder-Richardson methods. These measures to extent to which items within one form of

the test have as much in common with one another as do the items in that one form with corresponding items in an equivalent form. The strength of this estimate of reliability depends upon the context to which the entire test represents a single, fairly consistent measure of a concept.

Normally these estimates are lower than the split halves but estimates higher than the test-retest and parallel form estimates. These techniques are also called item total correlations. There are different techniques to estimate the internal consistency of the test using K-R procedures, but two of them are more frequently used by the measurement experts. The first KR-20 is difficult to calculate as it is based on the information of the percentages of the students passing each item on the test. However, it gives more accurate results (Kubiszyn and Borich, 2003). The KR-20 formula is given below.

KR20 Formula

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

Where “pq” provides a test score error variance for an "average" person, we know that the sampled people vary, i.e., the variance of their raw scores is greater than zero. Persons with high or low scores have less score error variance than those with scores near fifty percent correct where the score error variance is maximum. Since the "average" person variance used in the KR20 formula is always larger than the lower score error variance of persons with extreme scores, it must always overestimate their score error variances.

The second formula, which is easier to calculate but slightly less accurate is called KR21. It requires only the information about the number of items, the mean of the test score and the standard deviation. The formula KR21 is as under.

$$r_1 = \frac{n\sigma^2 m(n-m)}{\sigma^2(n-1)}$$

Studies indicated that this formula provide good results even when the item difficulties are not consistent.

### 5.3 Factors Affecting Reliability

Reliability of the test is an important characteristic as we use the test results for the future decisions about the students’ educational advances and for the job selection and many more. The methods to assure the reliability of the tests have been discussed. Many examples have been provided in order to in-depth understanding of the concepts. Here we shall focus upon the different factors that may affect the reliability of the test. The degree of the affect of each factor varies from the situation to situation. Controlling the factor may improve the reliability and otherwise it may lower the consistency of production of scores. Some of the factors that directly or indirectly affect the test reliability are given as under.

### 5.3.1. Test Length

As a rule, adding more homogeneous questions to a test will increase the test's reliability. The more observations there are of a specific trait, the more accurate the measure is likely to be. Adding more questions to a psychological test is similar to adding finer distinctions on a measuring tape.

### 5.3.2. Method Used to Estimate Reliability

The reliability coefficient is an estimate that can change depending on the method used to calculate it. The method chosen to estimate the reliability should fit the way in which the test will be used.

### 5.3.3 Heterogeneity of Scores

Heterogeneity is referred as the differences among the scores obtained from class. You may say that there are some students who got high scores and some students who got low scores or intelligent students who got high scores and other one got low scores or the difference could be due to any reason may be income level, intelligence of the students, parents qualification etc. Whichever is the reason for the variability of the scores the greater the variability (range) of test scores, the higher the reliability. Increasing the heterogeneity of the examinee sample increases variability (individual differences) thus reliability increases.

### 5.3.4 Difficulty

A test that is too difficult or too easy reduces the reliability (e.g., fewer test-takers get the answers correctly or vice-versa). A moderate level of difficulty increases test reliability.

### 5.3.5 Errors that Can Increase or Decrease Individual Scores:

There might be some errors committed by the test developers that also affect the reliability of the tests developed by teachers. These errors initially affect the students' scores, mean deviate the scores from the true ability of the students, and therefore affect the reliability. A careful consideration of these factors may help to measure the true ability of the students.

- The test itself: the overall look of the test may affect the students score. Normally a test is written in well readable font size and style, the language of the test should be simple and understandable.
- The test administration: After the development of the test, the test developer may have to prepare the manual of the test administration, the time, environment, invigilation, and the anxiety also affects students' performance while attempting

the test. Therefore the uniform administration of the test leads to the increased reliability.

- The test scoring: Marking of the test is another factor towards the variation in the scores of the students. Normally there are many raters to rate the students' responses/answers on the test. Objective type test items and the marking rubric for essay type/ supply type test items help to get the consistent scores.

**Ensuring the Reliability of Test:**

The most straightforward ways to improve a test's reliability are`

**First**, calculate the item-test correlations and rewrite or reject any that are too low. Any item that does not correlate with the total test at least (point-biserial) r = .25, should be reconsidered.

**Second**, look at the items that did correlate well and write more like them. The longer the test, the higher the reliability will be.

## 5.4 Usability of Assessment Tools

Another important feature of a good assessment tool (Classroom test) is its usability. Classroom teachers are well familiar with issues related to the usability and practicality of the tests, but they need to think of how practical matters relate to testing. Usability refers to the extent to which a test can be used by students and teachers to achieve specified goals in an effective and efficient manner. It also refers to facilities available to test developers regarding both administration and scoring procedures of a test. As far as administration is concerned, test developers should be attentive to the possibilities of giving a test under reasonably acceptable conditions. For example, suppose a team of experts decide on giving a listening comprehension test to large groups of examinees. In this case, test developers should make sure those facilities such as audio equipments and/or suitable acoustic rooms are available. Otherwise, no matter how reliable and valid the test may be, it will not be practical.

Regarding the scoring procedures of a test, one should pay attention to the problem of ease of scoring as well as ease of interpretation of scores. For instance, assume that composition tests are excellent indicators of language ability. Would it be possible to use it in large scale administrations? How would the compositions be scored? How long would it take to score them? All these questions relate to the usability of the test in terms of scoring. Therefore, test developers should be very careful in selecting and administering a test. The test should be practical, i.e., it should be easy to administer, easy to score, and easy to interpret the scores in other words easy to use.

A good classroom test should be "teacher-friendly". A teacher should be able to develop, administer and mark it within the available time and with available resources. Classroom tests are only valuable to students when they are returned promptly and when the feedback from assessment is understood by the student. In this way, students can benefit

from the test-taking process. The issues regarding usability of the test include cost of test development and maintenance, time (for development and test length), resources (everything from computer access, copying facilities, AV equipment to storage space), ease of marking, availability of suitable/trained markers and administrative logistics.

The following are two very important aspects that contribute towards the usability of the test.

### Transparency

In simple words transparency is a process which requires from teachers to maintain objectivity and the honesty for developing, administering, marking and reporting the test results. Transparency refers to the availability of clear, accurate information to students about testing. Such information should include outcomes to be evaluated, formats used, weighting of items and sections, time allowed to complete the test, and grading criteria. Transparency makes students part of the testing process. No one could doubt any aspect of the testing process. It also requires setting rules and keeping record of the testing process.

### Security

Most teachers feel that security is an issue only in large-scale, high-stakes testing. However, security is part of both reliability and validity. If a teacher invests time and energy in developing good tests that accurately reflect the course outcomes, then it is desirable to be able to recycle the tests or similar materials. This is especially important if analyses show that the items, distracters and test sections are valid and discriminating. In some parts of the world, cultural attitudes towards “collaborative test-taking” are a threat to test security and thus to reliability and validity. As a result, there is a trade-off between letting tests into the public domain and giving students adequate information about tests.

## 5.5 Summary

This unit dealt with the reliability and usability of a good test. First, the concepts were defined, and then the methods of estimating and assuring reliability and the factors affecting was discussed in detail. Finally, the concept of practicality was explained.

The procedures for test construction may seem tedious. However, regardless of the complexity of the tasks in determining the reliability and usability of a test, these concepts are essential parts of test construction. It means that in order to have an acceptable and applicable test, upon which reasonably sound decisions can be made, test developers should go through planning, preparing, reviewing, and pretesting processes.

Without determining these parameters, nobody is ethically allowed to use a test for practical purposes. Otherwise, the test users are bound to make inexcusable mistakes, unreasonable decisions and unrealistic appraisals.

## 5.6 Self Assessment Questions

### 5.6.1 Essay Type

1. Define the term reliability and elaborate the importance and scope of reliability of a test.
2. State different types of reliability and explain each type with examples.
3. Give the limitations of test retest, split half and parallel form reliability methods.
4. Identify different factors affecting reliability a test also suggest measures to control the impact of these factors.
5. Discuss the problems encountered by teachers and students while using the tests.

### 5.6.2 Objective Type

**I Mark the following statements as true or false.**

- Assessment is an integral part of teaching learning process.
- If a test measures for what it is designed to measure then it is a reliable test.
- If the scores of the two administration of a test are consistent then it is called the test-retest reliability of the test.
- Administering two different forms of the test at a time is method of split half reliability.
- If item does not correlate with the total test scores it should be reconsidered.

## 5.7 Reference/ Suggested Readings:


Anastasi, A. (1982). Psychological Testing. New York: Macmillan.

Babour, R. S. (1998). Mixing Qualitative Methods: Quality Assurance or Qualitative quagmire? *Qualitative Health Research, 8(3), 352-361*.

Bazovsky, I. (1961). Reliability Theory and Practice. Prentice-Hall Report.

Bogdan, R. C. & Biklen, S. K. (1998). Qualitative Research in Education: An Introduction to Theory and Methods (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Cohen, R. J., Swerdlik, M. E., & Phillips, S. M. (1996). Psychological Testing and Measurement: An Introduction to Tests and Measurement. Mountain View, CA: Mayfield Publishing Company.

Crooks, T. J. (1988). The Impact of Classroom Evaluation Practices on Students. Review of Educational Research, 58(4): 438-481.

Hopkins, C.D. & Antes, R.L. (2000). Classroom Measurement and Evaluation, (3rd Ed). F.E. Peacock Publishers, Int. ITASCA, ILLIONS.

Kubiszyn, T. & Borich, G. (2003). Educational Testing and Measurement: Classroom Application and Practice. New York, Johan Wiley and Sons, Inc.

Joppe, M. (2000). *The Research Process*. Retrieved December 16, 2006, from http://www.ryerson.ca/~mjoppe/rp.htm

**UNIT–6**

# VALIDITY OF THE ASSESSMENT TOOLS

**Written by:**
**Dr. Muhammad Tanveer Afzal**
**Revised by:**
**Prof. Dr. Rehana Masrur**

## CONTENT

# INTRODUCTION

Assessment is a process by which information is obtained relative to some known objective or goal. Assessment is a broad term that includes measurement, testing and valuing the worth. Most of the times the teachers use assessment to make the educational decisions on the basis of tests. If we desire to uncover the truths about the educational advances of the students we focus on the assessment procedures and the final assessments made by the teachers both during the instructional process and at the end of the instruction. Therefore it is necessary to make the valid and reliable assessments during and after the teaching learning process. According to Boud (1995) students may (with difficulty) escape from the effects of poor teaching, but they cannot (by definition if they want to graduate) escape the effects of poor assessment. This highlights the importance of getting our assessment practices right for our students.

Rowntree (1987) states that assessment procedures offer answers to the following questions:

- What student qualities and achievements are actively valued and rewarded by the system?
- How are its purposes and intentions realized?

Two major purposes of the assessment has been identified by the experts of measurement, the first is to assist learning and second to determine the effectiveness of the educational process these can only be achieved when the teachers are sure about the tools for example tests, they use for the assessment that test are valid and reliable. When a teachers or instructor has to choose among the two or more tests, all of which are available from the well reputable sources, it impose some difficulty for the teacher/instructor. Therefore it is essential to check the local conditions and the contents of the instructions, that which one is closely aligned with the contents. On the other hand, we can say that we have to focus upon the objectives of the instructional process. The alignment of test items with the learning outcomes, this characteristic of the assessment tools is called the validity of the test.

In order to assure the validity, we must ask these questions to make sure that our assessment matches our educational purposes. As a teacher we should find the most appropriate assessment method for assessing the desired learning outcomes. When considering the assessment tasks we should consider the strengths and weaknesses of the test items and the arrangement of the items in the tests.

In the previous unit you have learnt about the reliability of the assessment tools, that refers to the consistency, here in this unit the prime consideration is the validity, which may be referred as the credibility of the assessment tool. Therefore different definitions of validity, methods of assuring validity of the assessment tools and the factors affecting the validity of the assessment tools have been discussed in this unit.

# OBJECTIVES

After studying this unit, prospective teachers will be able to:

- define and explain the term validity.
- differentiate among the different forms of establishing validity of the assessment tools.
- establish construct validity of the assessment tools.
- assure concurrent validity of the assessment tools
- establish predictive validity of the assessment tools.
- assure criterion validity of the assessment tools.
- identify the factors affecting validity of the assessment tools.
- construct valid and reliable assessment tools.

## 6.1 Nature of Validity

The validity of an assessment tool is the degree to which it measures for what it is designed to measure. For example if a test is designed to measure the skill of addition of three digit in mathematics but the problems are presented in difficult language that is not according to the ability level of the students then it may not measure the addition skill of three digits, consequently will not be a valid test. Many experts of measurement had defined this term, some of the definitions are given as under.

According to Business Dictionary the "Validity is the degree to which an instrument, selection process, statistical technique, or test measures what it is supposed to measure."

Cook and Campbell (1979) define validity as the appropriateness or correctness of inferences, decisions, or descriptions made about individuals, groups, or institutions from test results.

According to APA (American Psychological association) standards document the validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself.

Howell's (1992) view of validity of the test is; a valid test must measure specifically what it is intended to measure.

According to Messick the validity is a matter of degree, not absolutely valid or absolutely invalid. He advocates that, over time, validity evidence will continue to gather, either enhancing or contradicting previous findings.

Overall we can say that in terms of assessment, validity refers to the extent to which a test's content is representative of the actual skills learned and whether the test can allow accurate conclusions concerning achievement. Therefore validity is the extent to which a test measures what it claims to measure. It is vital for a test to be valid in order for the results to be accurately applied and interpreted.

Let's consider the following examples.

**Examples:**

1. Say you are assigned to observe the effect of strict attendance policies on class participation. After observing two or three weeks you reported that class participation did increase after the policy was established.
2. Say you are intended to measure the intelligence and if math and vocabulary truly represent intelligence then a math and vocabulary test might be said to have high validity when used as a measure of intelligence.

A test has validity evidence, if we can demonstrate that it measures what it says to measure. For instance, if it is supposed to be a test for fifth grade arithmetic ability, it should measure fifth grade arithmetic ability and not the reading ability.

### 6.1.1 Test Validity and Test Validation

Tests can take the form of written responses to a series of questions, such as the paper-and-pencil tests, or of judgments by experts about behaviour in the classroom/school, or for a work performance appraisal. The form of written test results also vary from pass/fail, to holistic judgments, to a complex series of numbers meant to convey minute differences in behaviour.

Regardless of the form a test takes, its most important aspect is how the results are used and the way those results impact individual persons and society as a whole. Tests used for admission to schools or programs or for educational diagnosis not only affect individuals, but also assign value to the content being tested. A test that is perfectly appropriate and useful in one situation may be inappropriate or insufficient in another. For example, a test that may be sufficient for use in educational diagnosis may be completely insufficient for use in determining graduation from high school.

Test validity, or the validation of a test, explicitly means validating the use of a test in a specific context, such as college admission or placement into a course. Therefore, when determining the validity of a test, it is important to study the test results in the setting in which they are used. In the previous example, in order to use the same test for educational diagnosis as for high school graduation, each use would need to be validated separately, even though the same test is used for both purposes.

### 6.1.2 Purpose of Measuring Validity

Most, but not all, tests are designed to measure skills, abilities, or traits that are and are not directly observable. For example, scores on the Scholastic Aptitude Test (SAT) measure developed critical reading, writing and mathematical ability. The score on the SAT that an examinee obtains when he/she takes the test is not a direct measure of critical reading ability, such as degrees centigrade is a direct measure of the heat of an object. The amount of an examinee's developed critical reading ability must be inferred from the examinee's SAT critical reading score.

The process of using a test score as a sample of behaviour in order to draw conclusions about a larger domain of behaviours is characteristic of most educational and psychological tests. Responsible test developers and publishers must be able to demonstrate that it is possible to use the sample of behaviours measured by a test to make valid inferences about an examinee's ability to perform tasks that represent the larger domain of interest.

### 6.1.3 Validity versus Reliability

A test can be reliable but may not be valid. If test scores are to be used to make accurate inferences about an examinee's ability, they must be both reliable and valid. Reliability is a prerequisite for validity and refers to the ability of a test to measure a particular trait or skill consistently. In simple words we can say that same test administered to same students may yield same score. However, tests can be highly reliable and still not be valid for a particular purpose. Consider the example of a thermometer if there is a systematic error and it measures five degrees higher. When the repeated readings has been taken under the same conditions the thermometer will yield consistent (reliable) measurements, but the inference about the temperature is faulty.

This analogy makes it clear that determining the reliability of a test is an important first step, but not the defining step, in determining the validity of a test.

There are different methods of assuring the validity of the assessment tools. Some of the important methods namely, content, construct, predictive, and criterion validity are discussed in section 6.4.

## 6.2 Methods of Measuring Validity

Validity is the appropriateness of a particular uses of the test scores, test validation is then the process of collecting evidence to justify the intended use of the scores. In order to collect the evidence of validity there are many types of validity methods that provide usefulness of the assessment tools. Some of them are listed below.

### 6.2.1 Content Validity

The evidence of the content validity is judgmental process and may be formal or informal. The formal process has systematic procedure which arrives at a judgment. The important components are the identification of behavioural objectives and construction of table of specification. Content validity evidence involves the degree to which the content of the test matches a content domain associated with the construct. For example, a test of the ability to add two numbers, should include a range of combinations of digits. A test with only one-digit numbers, or only even numbers, would not have good coverage of the content domain. Content related evidence typically involves Subject Matter Experts (SME's) evaluating test items against the test specifications.

It is a non-statistical type of validity that involves “the systematic examination of the test content to determine whether it covers a representative sample of the behaviour domain to be measured” (Anastasi & Urbina, 1997). For example, does an IQ questionnaire have items covering all areas of intelligence discussed in the scientific literature?

A test has content validity built into it by careful selection of which items to include (Anastasi & Urbina, 1997). Items are chosen so that they comply with the test specification which is drawn up through a thorough examination of the subject domain. Foxcraft et al. (2004, p. 49) note that by using a panel of experts to review the test

specifications and the selection of items the content validity of a test can be improved. The experts will be able to review the items and comment on whether the items cover a representative sample of the behaviour domain.

**For Example -** In developing a teaching competency test, experts on the field of teacher training would identify the information and issues required to be an effective teacher and then will choose (or rate) items that represent those areas of information and skills which are expected from a teacher to exhibit in classroom.

**Lawshe (1975)** proposed that each rater should respond to the following question for each item in content validity:

Is the skill or knowledge measured by this item?

- Essential
- Useful but not essential
- Not necessary

With respect to educational achievement tests, a test is considered content valid when the proportion of the material covered in the test approximates the proportion of material covered in the course.

| **Activity 6.1:** | Make a test from any chapter of science book of class 7th and test whether it is valid or not with the reference to its content? |
|---|---|

There are different types of content validity; the major types face validity and the curricular validity are as below.

### 1 Face Validity

Face validity is an estimate of whether a test appears to measure a certain criterion; it does not guarantee that the test actually measures phenomena in that domain. Face validity is very closely related to content validity. While content validity depends on a theoretical basis for assuming if a test is assessing all domains of a certain criterion (e.g. does assessing addition skills yield in a good measure for mathematical skills? - To answer this you have to know, what different kinds of arithmetic skills mathematical skills include ) face validity relates to whether a test appears to be a good measure or not. This judgment is made on the "face" of the test, thus it can also be judged by the amateur.

Face validity is a starting point, but should NEVER be assumed to be provably valid for any given purpose, as the "experts" may be wrong.

**For example**- suppose you were taking an instrument reportedly measuring your attractiveness, but the questions were asking you to identify the correctly spelled word in each list. Not much of a link between the claim of what it is supposed to do and what it actually does.

**Possible Advantage of Face Validity...**

- If the respondent knows what information we are looking for, they can use that "context" to help interpret the questions and provide more useful, accurate answers.

**Possible Disadvantage of Face Validity...**

- If the respondent knows what information we are looking for, they might try to "bend & shape" their answers to what they think we want

| | |
|---|---|
| **Activity 6.2:** | Make an objective type test and discuss its face validity with at three experts of the subject considering the grade levl of the students. |

### 2. Curricular Validity

The extent to which the content of the test matches the objectives of a specific curriculum as it is formally described. Curricular validity takes on particular importance in situations where tests are used for high-stakes decisions, such as Punjab Examination Commission exams for fifth and eight grade students and Boards of Intermediate and Secondary Education Examinations. In these situations, curricular validity means that the content of a test that is used to make a decision about whether a student should be promoted to the next levels should measure the curriculum that the student is taught in schools.

Curricular validity is evaluated by groups of curriculum/content experts. The experts are asked to judge whether the content of the test is parallel to the curriculum objectives and whether the test and curricular emphases are in proper balance. Table of specification may help to improve the validity of the test.

| | |
|---|---|
| **Activity 6.3:** | Curricular validity affects the performance of the examinees, how can you measure the curricular validity of tests, discuss the current practice followed by the secondary level teachers with two or three SST in your town. |

### 6.2.2 Construct Validity

Before defining the construct validity, it seems necessary to elaborate the concept of construct. It is the concept or the characteristic that a test is designed to measure. A construct provides the target that a particular assessment or set of assessments is designed to measure; it is a separate entity from the test itself. According to Howell (1992) Construct validity is a test's ability to measure factors which are relevant to the field of study. Construct validity is thus an assessment of the quality of an instrument or experimental design. It says 'Does it measure the construct it is supposed to measure'. Construct validity is rarely applied in achievement test.

Construct validity refers to the extent to which operationalizations of a construct (e.g. practical tests developed from a theory) do actually measure what the theory says they do. For example, to what extent is an IQ questionnaire actually measuring "intelligence"? Construct validity evidence involves the empirical and theoretical support for the interpretation of the construct. Such lines of evidence include statistical analyses of the internal structure of the test including the relationships between responses to different test items. They also include relationships between the test and measures of other constructs. As currently understood, construct validity is not distinct from the support for the substantive theory of the construct that the test is designed to measure. As such, experiments designed to reveal aspects of the causal role of the construct also contribute to construct validity evidence.

Construct validity occurs when the theoretical constructs of cause and effect accurately represent the real-world situations they are intended to model. This is related to how well the experiment is operationalized. A good experiment turns the theory (constructs) into actual things you can measure. Sometimes just finding out more about the construct (which itself must be valid) can be helpful. The construct validity addresses the construct that are mapped into the test items, it is also assured either by judgmental method or by developing the test specification before the development of the test. The constructs have some essential properties the two of them are listed as under:

1. Are abstract summaries of some regularity in nature?
2. Related with concrete, observable entities.

For Example - Integrity is a construct; it cannot be directly observed, yet it is useful for understanding, describing, and predicting human behaviour.

| | |
|---|---|
| **Activity 6.4:** | Make a tests for a child of class 4th which measures the shyness construct of his personality, and valid this test with reference to its construct validity. |

There are different types of construct validity; the convergent and the discriminant validity are explained as follows.

### 1. Convergent Validity

Convergent validity refers to the degree to which a measure is correlated with other measures that it is theoretically predicted to correlate with. OR

Convergent validity occurs where measures of constructs that are expected to correlate do so. This is similar to concurrent validity (which looks for correlation with other tests).

For example, if scores on a specific mathematics test are similar to students scores on other mathematics tests, then convergent validity is high (there is a positively correlation between the scores from similar tests of mathematics).

### 2. Discriminant Validity

Discriminant validity describes the degree to which the operationalization does not correlate with other operationalizations that it theoretically should not be correlated with.

OR

Discriminant validity occurs where constructs that are expected not to relate with each other, such that it is possible to discriminate between these constructs. For example, if discriminant validity is high, scores on a test designed to assess students skills in mathematics should not be positively correlated with scores from tests designed to assess intelligence.

| |
|---|
| Convergence and discrimination are often demonstrated by correlation of the measures used within constructs. Convergent validity and Discriminant validity together demonstrate construct validity. |

### 6.2.3 Criterion Validity

Criterion validity evidence involves the correlation between the test and a criterion variable (or variables) taken as representative of the construct. In other words, it compares the test with other measures or outcomes (the criteria) already held to be valid. For example, employee selection tests are often validated against measures of job performance (the criterion), and IQ tests are often validated against measures of academic performance (the criterion).

If the test data and criterion data are collected at the same time, this is referred to as concurrent validity evidence. If the test data is collected first in order to predict criterion data collected at a later point in time, then this is referred to as predictive validity evidence.

**For example,** the company psychologist would measure the job performance of the new artists after they have been on-the-job for 6 months. He or she would then correlate scores on each predictor with job performance scores to determine which one is the best predictor.

| | |
|---|---|
| ***Activity 6.5:*** | Administer any test of English to grade $9^{th}$ and predict the performance of the students for future on the basis of that test. Compare its results after a month with their monthly English test to check the criterion validity of that test with reference to the prediction made about his performance on English language. |

### 6.2.4 Concurrent Validity

According to Howell (1992) “concurrent validity is determined using other existing and similar tests which have been known to be valid as comparisons to a test being

developed. There is no other known valid test to measure the range of cultural issues tested for this specific group of subjects".

Concurrent validity refers to the degree to which the scores taken at one point correlates with other measures (test, observation or interview) of the same construct that is measured at the same time. Returning to the selection test example, this would mean that the tests are administered to current employees and then correlated with their scores on performance reviews. This measure the relationship between measures made with existing tests. The existing test is thus the criterion. For example, a measure of creativity should correlate with existing measures of creativity.

**For example:**

To assess the validity of a diagnostic screening test. In this case the predictor (X) is the test and the criterion (Y) is the clinical diagnosis. When the correlation is large this means that the predictor is useful as a diagnostic tool.

### 6.2.5 Predictive Validity

Predictive validity assures how well the test predicts some future behaviour of the examinee. It validity refers to the degree to which the operationalization can predict (or correlate with) other measures of the same construct that are measured at some time in the future. Again, with the selection test example, this would mean that the tests are administered to applicants, all applicants are hired, their performance is reviewed at a later time, and then their scores on the two measures are correlated. This form of the validity evidence is particularly useful and important for the aptitude tests, which attempt to predict how well the test taker will do in some future setting.

This measures the extent to which a future level of a variable can be predicted from a current measurement. This includes correlation with measurements made with different instruments. For example, a political poll intends to measure future voting intent. College entry tests should have a high predictive validity with regard to final exam results. When the two sets of scores are correlated, the coefficient that results is called the predictive validity coefficient.

**Examples:**

1. If higher scores on the Boards Exams are positively correlated with higher G.P.A.'s in the Universities and vice versa, then the Board exams is said to have predictive validity.
2. We might theorize that a measure of math ability should be able to predict how well a person will do in an engineering-based profession.

The predictive validity depends upon the following two steps.

- Obtain test scores from a group of respondents, but do not use the test in making a decision.
- At some later time, obtain a performance measure for those respondents, and correlate these measures with test scores to obtain predictive validity.

## 6.3 Factors Affecting Validity

Validity evidence is an important aspect to consider while thinking of the classroom testing and measurement. There are many factors that tend to make test result invalid for their intended use. A little careful effort by the test developer help to control these factors, but some of them need systematic approach. No teacher would think of measuring knowledge of social studies with an English test. Nor would a teacher consider measuring problem-solving skills in third-grade arithmetic with a test designed for sixth grades. In both instances, the test results would obviously be invalid. The factors influencing validity are of this same general but match more subtle in character. For example, a teacher may overload a social studies test with items concerning historical facts, and thus the scores are less valid as a measure of achievement in social studies. Or a third–grade teacher may select appropriate arithmetic problems for a test but use vocabulary in the problems and directions that only the better readers are able to understand. The arithmetic test then becomes, in part, reading test, which invalidates the result for their intended use. These examples show some of the more subtle factors influencing validity, for which the teacher should be alert, whether constructing classroom tests or selecting published tests. Some other factors that may affect the test validity are discussed as under.

### 1. Instructions to Take A Test:

The instructions with the test should be clear and understandable and it should be in simple language. Unclear instructions may restrict the pupil how to respond to the items, whether it is permissible to guess, and how to record the answers will tend to reduce validity.

### 2. Difficult Language Structure:

Language of the test or instructions to the test that is too complicated for the pupils taking the test will result in the test's measuring reading comprehension and aspects of intelligence, which will distort the meaning of the test results. Therefore it should be simple considering the grade for which the test is meant.

### 3. Inappropriate Level of Difficulty:

In norm-references tests, items that are too easy or too difficult will not provide reliable discriminations among pupils and will therefore lower validity. In criterion-referenced tests, the failure to match the difficulty specified by the learning outcome will lower validity.

**4. Poorly Constructed Test Items**:

There may be some items that provide direction to the answer or test items that unintentionally provide alertness in detecting clues are poor items, these items may harm the validity of the test.

**5. Ambiguity in Items Statements:**

Ambiguous statements in test items contribute to misinterpretations and confusion. Ambiguity sometimes confuses the better pupils more than it does the poor pupils, causing the items to discriminate in a negative direction.

**6. Length of the Test:**

A test is only a Sample of the many questions that might be asked. If a test is too short to provide a representative sample of the performance we are interested in, its validity will suffer accordingly. Similarly a too lengthy test is also a threat to the validity evidence of the test.

**7. Improper Arrangement of Items:**

Test items are typically arranged in order of difficulty, with the easiest items first. Placing difficult items early in the test may cause pupils to spend too much time on these and prevent them from reaching items they could easily answer. Improper arrangement may also influence validity by having a detrimental effect on pupil motivation. The influence is likely to be strongest with young pupils.

**8. Identifiable Pattern of Answers:**

Placing correct answers in some systematic pattern will enable pupils to guess the answers to some items more easily, and this will lower validity.

In short, any defect in the tests construction that prevents the test items from functioning as intended will invalidate the interpretations to be drawn from the results. There may be many other factors that can also affect the validity of the test to some extents. Some of these factors are listed as under.

- Inadequate sample
- Inappropriate selection of constructs or measures.

- Items that do not function as intended
- Improper administration: inadequate time allowed, poorly controlled conditions
- Scoring that is subjective
- Insufficient data collected to make valid conclusions.
- Too great a variation in data (can't see the wood for the trees).
- Inadequate selection of target subjects.
- Complex interaction across constructs.
- Subjects giving biased answers or trying to guess what they should say.

| | |
|---|---|
| ***Activity 6.6:*** | Select a teacher made test for $10^{th}$ grade and discuss it with any teacher for improvement of the validity evidences in light of factors discussed above. |

## 6.4 Relationship between Validity and Reliability

Reliability and validity are two different standards used to gauge the usefulness of a test. Though different, they work together. It would not be beneficial to design a test with good reliability that did not measure what it was intended to measure. The inverse, accurately measuring what we desire to measure with a test that is so flawed that results are not reproducible, is impossible. Reliability is a necessary requirement for validity. This means that you have to have good reliability in order to have validity. Reliability actually puts a cap or limit on validity, and if a test is not reliable, it cannot be valid. Establishing good reliability is only the first part of establishing validity. Validity has to be established separately. Having good reliability does not mean we have good validity, it just means we are measuring something consistently. Now we must establish, what it is that we are measuring consistently. The main point here is reliability is necessary but not sufficient for validity. In short we can say that reliability means noting when the problem is validity.

## 6.5 Summary

The validity of an assessment tools is the degree to which it measures for what it is designed to measure. Lots of terms are used to describe the different types of evidence for claiming the validity of a test result for a particular inference. The terms have been used in different ways over the years by different authors. More important than the terms, is knowing how to look for validity evidence. Does the score correlate with other measures of the same domain? Does the score predict future performance? Does the score correlate with other domains within the same test? Does it negatively correlate with scores that indicate opposite skills? Do the score results make sense when one simply looks at them? What impact on student behaviour has the test had? Each of these

questions relates to different kinds of validity evidence (specifically: content validity, concurrent validity, predictive validity, construct validity, face validity). Content validity evidence involves the degree to which the content of the test matches a content domain associated with the construct. The concurrent validity evidences can be assured by comparing the two tests. There are many factors that can reduce the validity of the test, the teachers or test developers have to consider these factors while constructing and administration of the tests. It better to follow the systematic procedure and this rigorous approach may help to improve the validity and the reliability of the tests.

## 6.6 Self Assessment Questions

1. Define the term validity and elaborate its different types.
2. Develop a table of specification for seventh grade science test so as to assure the content validity.
3. Develop multiple choice test items as per table of specification developed in question#2.
4. Curricular validity affects the performance of the examinees, how can we measure the curricular validity of tests? Explain.
5. Discuss the terms validity and reliability with any of teacher in a nearby high school.
6. Interview the teachers to find that existing practices to control the factors affecting validity of the tests.
7. Which type of validity is more important? Support your statement with arguments

## 6.7 References/Suggested Readings


1. American Educational Research Association, Psychological Association, & National Council on Measurement in Education. (1999). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.
2. Büttner, J (1997). "Diagnostic Validity as a Theoretical Concept and as a Measurable Quantity". Clinica Chimica Acta; International Journal of Clinical Chemistry 260 (2): 131–43.
3. Ogince, M; Hall, T; Robinson, K; Blackmore, AM (2007). "The Diagnostic validity of the Cervical Flexion-rotation Test in C1/2-related Cervicogenic Headache". Manual Therapy 12 (3): 256–62.
4. Kendell, R; Jablensky, A (2003). "Distinguishing Between the Validity and Utility of Psychiatric Diagnoses". The American Journal of Psychiatry 160 (1): 4–12.
5. Kendler, KS (2006). "Reflections on the Relationship between Psychiatric Genetics and Psychiatric Nosology". The American Journal of Psychiatry 163 (7): 1138–46.
6. Cronbach, L. J.; Meehl, P. E. (1955). "Construct Validity in Psychological tests". Psychological Bulletin 52 (4): 281–302.
7. Black, B. (2007). Critical Thinking – A Tangible Construct? Research Matters: A Cambridge Assessment Publication 2, 2-4.
8. Cambridge Assessment (2008) The Cambridge Approach, Cambridge: University of Cambridge Local Examinations Syndicate.
9. Astin, A.W., Banta, T.W. *et al*. (2003). *9 Principles of Good Practice for Assessing Student Leaning*. Available online.
10. Hernon, P.& Dugan, R.E. (2004). *Outcomes Assessment in Higher Education*. Westport, CT: Libraries Unlimited.
11. Knight, P.T. (2002). *The Achilles' Heel of Quality: The Assessment of Student Learning Quality in Higher Education*, 8, 107-115.
12. Taras, M. (2002). Using Assessment for Learning and Learning for Assessment. Assessment & Evaluation in Higher Education, 27, 501-510.
13. Angelo, T. & Cross, P. (1993). *Classroom Assessment Techniques*. San Francisco: Jossey Bass.
14. Suskie, Linda. (2004). *Assessing Student Learning: A Common Sense Guide*. Bolton, MA: Anker Publishing.
15. Walvoord, Barbara. (2004). *Assessment Clear and Simple*. San Francisco: Jossey-Bass.

16. Michelson, E. & Mandell, A. (2004). *Portfolio Development and the Assessment of Prior Learning: Perspectives, Models, and Practices*. Sterling, VA: Stylus

**Web Resources**

17. http://changingminds.org/explanations/research/design/types_validity.htm

18. http://professionals.collegeboard.com/higher-ed/validity/ ces/ handbook/ test-validity

19. http://professionals.collegeboard.com/higher-ed/ validity/ aces/ handbook /evidence

20. http://www.socialresearchmethods.net/kb/measval.php

21. http://www.businessdictionary.com/definition/validity.html

22. http://www.cael.ca/pdf/C6.pdf

23. 15.http://www.cambridgeassessment.org.uk/ca/digitalAssets/171263BB_CT definitionIAEA08.pdf

**UNIT–7**

# Planning and Administering Classroom Tests

**Written By:**
**Muhammad Idrees**
**Reviewed By:**
**Dr. Naveed Sultana**

## CONTENTS

# INTRODUCTION

With the approach for increased accountability in the educational system, it is vital that educators are able to apply a wide range of psychometric skills appropriate to the assessment of the students with different pace of learning and backgrounds. It is equally critical that educators have a comprehensive understanding of current measurement and evaluative trends such as competency testing, performance assessment, curriculum-based assessment and standardized assessment.

There are six major steps in planning and conduction an assessment: defining instructional objectives, outlining course contents, developing a test specification, construction of test items, administration of assessment and interpreting test scores.

There are two types of assessment i.e. assessment of learning & assessment for learning. Assessment of learning is used to assess students learning achievement at terminal stages whereas assessment for learning enhances teaching–learning process. Tests and assessments are an essential part of the instructional process. When properly done, they can not only effectively evaluate but also enhance students' learning and teachers' instruction. When poorly done, they can confuse and alienate students, distort the curriculum, and hinder good instruction. Test scores and grades sometimes affect "high-stakes" decisions about students, prompting intense concern that they be accurate and fair.

This course is designed to provide the students/prospective teachers with the principles and techniques necessary to develop sound student assessment strategies. The primary focus of the course will be on writing instructional objectives, developing different types of test items (selected response & constructed response), utilizing performance based and alternative assessment techniques, administering classroom evaluation procedures and interpreting test score for different purposes. There are eight major steps in planning and conducting an assessment:

- defining instructional objectives,
- outlining course contents,
- developing a test specification,
- selection of appropriate assessment tasks,
- preparation of relevant assessment tasks or construction of test items,
- assembly of assessment tasks,
- administration of assessment
- Interpreting test scores.

These steps will be discussed in detail in the unit. This unit will be of great value for the teachers/prospective teachers in developing and assembling suitable test to assess students learning achievements.

# OBJECTIVES

After intensive study of this unit, the students will be capable to:

- appreciate qualities needed to determine the quality of classroom tests.
- develop different types of test items for selected response test items
- develop different types of test items for constructed response test items
- efficiently administer classroom test
- utilize the techniques of objectively score and grade tests.

## 7.1 Planning a Test

The main objective of classroom assessment is to obtain valid, reliable and useful data regarding student learning achievement. This requires determining what is to be measured and then defining it precisely so that assessments tasks to measure desired performance can be developed. Classroom tests and assessments can be used for the following instructional objectives:

**i. Pre-testing**

Tests and assessments can be given at the beginning of an instructional unit or course to determine:-

- weather the students have the prerequisite skills needed for the instruction (readiness, motivation etc)
- to what extent the students have already achieved the objectives of planned instruction (to determine placement or modification of instruction)

ii. During the Instruction Testing

- provides bases for formative assessment
- monitor learning progress
- detect learning errors
- provide feedback for students and teachers

iii. End of Instruction Testing

- measure intended learning outcomes
- used for formative assessment
- provides bases for grades, promotion etc

Prior to developing an effective test, one needs to determine whether or not a test is the appropriate type of assessment. If the learning objectives are of primarily types of procedural knowledge (how to perform a task) then a written test may not be the best approach. Assessment of procedural knowledge generally calls for a performance demonstration assessed using a rubric. Where demonstration of a procedure is not appropriate, a test can be an effective assessment tool.

The first stage of developing a test is planning the test content and length. Planning the test begins with development of a blueprint or test specifications for the test structured on the learning outcomes or instructional objectives to be assessed by the test instrument. For each learning outcome, a weight should be assigned based on the relative importance of that outcome in the test. The weight will be used to determine the number of items related to each of the learning outcomes.

### 7.1.1 Test Specifications

When an engineer prepares a design to construct a building and choose the materials, he intends to use in construction, he usually know what a building is going to be used for, and therefore designs it to meet the requirements of its planned inhabitants. Similarly, in testing, table of specification is the blueprint of the assessment which specifies

percentages and weightage of test items and measuring constructs. It includes constructs and concepts to be measured, tentative weightage of each construct, specify number of items for each concept, and description of item types to be constructed. It is not surprising that specifications are also referred to as 'blueprints', for they are literally architectural drawings for test construction. Fulcher & Davidson (2009) divided test specifications into the following four elements:

- **Item specifications:** Item specifications describe the items, prompts or tasks, and any other material such as texts, diagrams, and charts which are used as stimuli. Typically, a specification at this sub-level contains two key elements: samples of the tasks to be produced, and guiding language that details all information necessary to produce the task.
- **Presentation Model**: Presentation model provides information how the items and tasks are presented to the test takers.
- **Assembly Model:** Assembly model helps the test developer to combine test items and tasks to develop a test format.
- **Delivery Model:** Delivery Model tells how the actual test is delivered. It includes information regarding test administration, test security/confidentiality and time constraint.

**Table 7.1: Table of Specifications for Social Studies Class VI**

| **Objectives/ Contents** | **Knowledge** | | | **Understanding** | | | **Application** | | | **Percentage** |
|---|---|---|---|---|---|---|---|---|---|---|
| | **LA** | **SA** | **MCQ** | **LA** | **SA** | **MCQ** | **LA** | **SA** | **MCQ** | |
| Climate | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 25% |
| Resources | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 25% |
| Population | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 25% |
| Society | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 25% |
| **Total** | **4** | **8** | **12** | **4** | **8** | **12** | **4** | **8** | **12** | **100%** |

LA: Long Answer, SA: Short Answers, MCQ: Multiple Choice Questions

**Note:** Number of items/questions and percentage may be changed according to the objectives/contents and hierarchy of learning.

## 7.2 General Consideration in Constructing Objective Test Items

The second step in test planning is determining the format and length of the test. The format is based on the different types of items to be included in the test. The construction of valid and good test items is a skill just like effective teaching. Some rules are to be followed and some techniques are to be used to construct good test items. Test items can

be used to assess student's ability to recognize concepts or to recall concepts. Generally there are two types of objective test items:-

i. Select type.
ii. Supply type.

### 7.2.1 Select Type Items

### A. Matching Items

According to W. Wiersma and S.G. Jurs (1990), the matching items consist of two parallel columns. The column on the left contains the questions to be answered, termed premises; the column on the right, the answers, termed responses. The student is asked to associate each premise with a response to form a matching pair. For example

| **Column "A" Capital City** | **Column "B" Country** |
|---|---|
| Islamabad | Iran |
| Tehran | Spain |
| Istanbul | Portugal |
| Madrid | Pakistan |
| Hague | Netherlands |
| | Turkey |
| | West Germany |

According to W. Wiersma and S.G. Jurs (1990) in some matching exercises the number of premises and responses are the same, termed a balanced or perfect matching exercise. In others, the number and responses may be different.

**Advantages**

The chief advantage of matching exercises is that a good deal of factual information can be tested in minimal time, making the tests compact and efficient. They are especially well suited to who, what, when and where types of subject matter. Further students frequently find the tests fun to take because they have puzzle qualities to them.

**Disadvantages**

The principal difficulty with matching exercises is that teachers often find that the subject matter is insufficient in quantity or not well suited for matching terms. An exercise should be confined to homogeneous items containing one type of subject matter (for instance, authors-novels; inventions inventors; major events-dates terms – definitions; rules examples and the like). Where unlike clusters of questions are used to adopt but

poorly informed student can often recognize the ill-fitting items by their irrelevant and extraneous nature (for instance, in a list of authors the inclusion of the names of capital cities).

Student identifies connected items from two lists. It is Useful for assessing the ability to discriminate, categorize, and association amongst similar concepts.

**Suggestions for Writing Matching Items**

Here are some suggestions for writing matching items:

i. Keep both the list of descriptions and the list of options fairly short and homogeneous – they should both fit on the same page. Title the lists to ensure homogeneity and arrange the descriptions and options in some logical order. If this is impossible, you're probably including too wide a variety in the exercise. Try constructing two or more exercises.
ii. Make sure that all the options are plausible distracters for each description to ensure homogeneity of lists.
iii. The list of descriptions on the left side should contain the longer phrases or statements, whereas the options on the right side should consist of short phrases, words or symbols.
iv. Each description in the list should be numbered (each is an item), and the list of options should be identified by letter.
v. Include more options than descriptions. If the option list is longer than the description list, it is harder for students to eliminate options. If the option list is shorter, some options must be used more than once. Always include some options that do not match any of the descriptions, or some that match more than one, or both.
vi. In the directions, specify the basis for matching and whether options can be used more than once.

## B. Multiple Choice Questions (MCQ's)

Norman E. Grounlund (1990) writes that the multiple choice question is probably the most popular as well as the most widely applicable and effective type of objective test. Student selects a single response from a list of options. It can be used effectively for any level of course outcome. It consists of two parts: the stem, which states the problem and a list of three to five alternatives, one of which is the correct (key) answer and the others are distracters ("foils" or incorrect options that draw the less knowledgeable pupil away from the correct response).

The stem may be stated as a direct question or as an incomplete statement. For example:

**Direct question**

Which is the capital city of Pakistan? -------- (Stem)

A. Lahore. ---------- (Distracter)
B. Karachi. ---------- (Distracter)
C. Islamabad. ---------- (Key)
D. Peshawar. ---------- (Distracter)

**Incomplete Statement**

The capital city of Pakistan is

A. Lahore.
B. Karachi.
C. Islamabad.
D. Peshawar.

**RULES FOR WRITING MULTIPLE-CHOICE QUESTIONS**

**1. Use Plausible Distracters (wrong-response options)**

- Only list plausible distracters, even if the number of options per question changes
- Write the options so they are homogeneous in content
- Use answers given in previous open-ended exams to provide realistic distracters

**2. Use a Question Format**

- Experts encourage multiple-choice items to be prepared as questions (rather than incomplete statements)

Incomplete Statement Format:

The capital of AJK is in----------.

Direct Question Format:

In which of the following cities is the capital of AJK?

**3. Emphasize Higher-Level Thinking**

- Use memory-plus application questions. These questions require students to recall principles, rules or facts in a real life context.
- The key to prepare memory-plus application questions is to place the concept in a life situation or context that requires the student to first recall the facts and then apply or transfer the application of those facts into a situation.
- Seek support from others who have experience writing higher-level thinking multiple-choice questions.

**EXAMPLES:**

Memory Only Example (Less Effective)

Which description best characterizes whole foods?

a. orange juice

b. toast

c. bran cereal

d. grapefruit

Memory-Plus Application Example (More Effective)

Sana's breakfast this morning included one glass of orange juice (from Concentrate), one slice of toast, a small bowl of bran cereal and a grapefruit. What "whole food" did Sana eat for breakfast?

a. orange juice

b. toast

c. bran cereal

d. grapefruit

Memory-Plus Application Example

Ability to Interpret Cause-and-Effect Relationships Example

Why does investing money in common stock protect against loss of assets during inflation?

a. It pays higher rates of interest during inflation.

b. It provides a steady but dependable income despite economic conditions.

c. It is protected by the Federal Reserve System.

d. It increases in value as the value of a business increases.

Ability to Justify Methods and Procedures Example

Why is adequate lighting necessary in a balanced aquarium?

a. Fish need light to see their food.

b. Fish take in oxygen in the dark.

c. Plants expel carbon dioxide in the dark.

d. Plants grow too rapidly in the dark.

**4. Keep Option Lengths Similar**

- Avoid making your correct answer the long or short answer

**5. Balance the Placement of the Correct Answer**

- Correct answers are usually the second and third option

**6. Be Grammatically Correct**

- Use simple, precise and unambiguous wording
- Students will be more likely to select the correct answer by finding the grammatically correct option

**7. Avoid Clues to the Correct Answer**

- Avoid answering one question in the test by giving the answer somewhere else in the test
- Have the test reviewed by someone who can find mistakes, clues, grammar and punctuation problems before you administer the exam to students
- Avoid extremes – never, always, only
- Avoid nonsense words and unreasonable statements

**8. Avoid Negative Questions**

- 31 of 35 testing experts recommend avoiding negative questions
- Students may be able to find an incorrect answer without knowing the correct answer

**9. Use Only One Correct Option (Or be sure the best option is clearly the best option)**

- The item should include one and only one correct or clearly best answer
- With one correct answer, alternatives should be mutually exclusive and not overlapping
- Using MC with questions containing more than one right answer lowers discrimination between students

**10. Give Clear Instructions**

Such as:

- Questions 1 - 10 are multiple-choice questions designed to assess your ability to remember or recall basic and foundational pieces of knowledge related to this course.
- Please read each question carefully before reading the answer options. When you have a clear idea of the question, find your answer and mark your selection on the answer sheet. Please do not make any marks on this exam.
- Questions 11 – 20 are multiple-choice questions designed to assess your ability to think critically about the subject.
- Please read each question carefully before reading the answer options.
- Be aware that some questions may seem to have more than one right answer, but you are to look for the one that makes the most sense and is the most correct.
- When you have a clear idea of the question, find your answer and mark your selection on the answer sheet.
- You may justify any answer you choose by writing your justification on the blank paper provided.

**11. Use Only a Single, Clearly-Defined Problem and Include the Main Idea in the Question**

- Students must know what the problem is without having to read the response options

**12. Avoid "All the Above" Option**

- Students merely need to recognize two correct options to get the answer correct

**13. Avoid the "None of the Above" Option**

- You will never know if students know the correct answer

**14. Don't Use MCQ When Other Item Types Are More Appropriate**

- Limited distracters or assessing problem-solving and creativity

**Advantages**

The chief advantage of the multiple-choice question according to N.E. Gronlund (1990) is its versatility. For instance, it is capable of being applied to a wide range of subject areas. In contrast to short answer items limit the writer to those content areas that are capable of being stated in one or two words, multiple choice item necessary bound to homogeneous items containing one type of subject matter as are matching items, and a

multiple choice question greatly reduces the opportunity for a student to guess the correct answer from one choice in two with a true – false items to one in four or five, there by increasing the reliability of the test. Further, since a multiple – choice item contains plausible incorrect or less correct alternative, it permits the test constructor to tine tune the discriminations (the degree or homogeneity of the responses) and control the difficulty level of the test.

**Disadvantages**

N.E. Gronlund (1990) writes that multiple-choice items are difficult to construct. Suitable distracters are often hard to come by and the teacher is tempted to fill the void with a "junk" response. The effect of narrowing the range of options will available to the test wise student. They are also exceedingly time consuming to fashion, one hour per question being by no means the exception. Finally they generally take student longer to complete (especially items containing fine discrimination) than do other types of objective question.

**Suggestions for Writing MCQ's Items**

Here are some guidelines for writing multiple-choice tests:

I. The stem of the item should clearly formulate a problem. Include as much of the item as possible, keeping the response options as short as possible. However, include only the material needed to make the problem clear and specific. Be concise – don't add extraneous information.
II. Be sure that there is one and only one correct or clearly best answer.
III. Be sure wrong answer choices (distracters) are plausible. Eliminate unintentional grammatical clues, and keep the length and form of all the answer choices equal. Rotate the position of the correct answer from item to item randomly.
IV. Use negation questions or statements only if the knowledge being tested requires it. In most cases it is more important for the student to know what a specific item of information is rather than what it is not.
V. Include from three to five options (two to four distracters plus one correct answer) to optimize testing for knowledge rather than encouraging guessing. It is not necessary to provide additional distracters from an item simply to maintain the same number of distracters for each item. This usually leads to poorly constructed distracters that add nothing to test validity and reliability.
VI. To increase the difficulty of a multiple-choice item, increase the similarity of content among the options.
VII. Use the option "none of the above" sparingly and only when the keyed answer can be classified unequivocally as right or wrong.
VII. Avoid using "all of the above". It is usually the correct answer and makes the item too easy for students with partial information.

## II. Supply Type Items

### A. Completion Items

Like true-false items, completion items are relatively easy to write. Perhaps the first tests classroom teachers' construct and students take completion tests. Like items of all other formats, though, there are good and poor completion items. Student fills in one or more blanks in a statement. These are also known as "Gap-Fillers." Most effective for assessing knowledge and comprehension learning outcomes but can be written for higher level outcomes. e.g.

The capital city of Pakistan is ----------.

**Suggestions for Writing Completion or Supply Items**

Here are our suggestions for writing completion or supply items:

I. If at all possible, items should require a single-word answer or a brief and definite statement. Avoid statements that are so indefinite that they may be logically answered by several terms.

**a. Poor item:**

Motorway (M1) opened for traffic in _____.

**b. Better item:**

Motorway (M1) opened for traffic in the year_____.

II. Be sure the question or statement poses a problem to the examinee. A direct question is often more desirable than an incomplete statement because it provides more structure.

III. Be sure the answer that the student is required to produce is factually correct. Be sure the language used in the question is precise and accurate in relation to the subject matter area being tested.

IV. Omit only key words; don't eliminate so many elements that the sense of the content is impaired.

**c. Poor item:**

The _____ type of test item is usually more _____ than the _____ type.

**d. Better item:**

The supply type of test item is usually graded less objectively than the _____ type.

V. Word the statement such that the blank is near the end of the sentence rather than near the beginning. This will prevent awkward sentences.

VI. If the problem requires a numerical answer, indicate the units in which it is to be expressed.

**B. Short Answer**

Student supplies a response to a question that might consistent of a single word or phrase. Most effective for assessing knowledge and comprehension learning outcomes but can be written for higher level outcomes. Short answer items are of two types.

- Simple direct questions
  Who was the first president of the Pakistan?
- Completion items

The name of the first president of Pakistan is _____.

The items can be answered by a work, phrase, number or symbol. Short-answer tests are a cross between essay and objective tests. The student must supply the answer as with an essay question but in a highly abbreviated form as with an objective question.

**Advantages**

Norman E. Gronlund (1990) writes that short-answer items have a number of advantages.

- They reduce the likelihood that a student will guess the correct answer
- They are relatively easy for a teacher to construct.
- They are will adapted to mathematics, the sciences, and foreign languages where specific types of knowledge are tested (The formula for ordinary table salt is _____).
- They are consistent with the Socratic question and answer format frequently employed in the elementary grades in teaching basic skills.

**Disadvantages**

According to Norman E. Grounlund (1990) there are also a number of disadvantages with short-answer items.

- They are limited to content areas in which a student's knowledge can be adequately portrayed by one or two words.
- They are more difficult to score than other types of objective-item tests since students invariably come up with unanticipated answers that are totally or partially correct.
- Short answer items usually provide little opportunity for students to synthesize, evaluate and apply information.

**7.3 General Consideration in Constructing Essay type Test Items**

Robert L. Ebel and David A. Frisbie (1991) in their book, write that "teachers are often as concerned with measuring the ability of students to think about and use knowledge as

they are with measuring the knowledge their students possess. In these instances, tests are needed that permit students some degree of latitude in their responses. Essay tests are adapted to this purpose. Student writes a response to a question that is several paragraphs to several pages long. Essays can be used for higher learning outcomes such as synthesis or evaluation as well as lower level outcomes. They provide items in which students supply rather than select the appropriate answer, usually the students compose a response in one or more sentences. Essay tests allow students to demonstrate their ability to recall, organize, synthesize, relate, analyze and evaluate ideas.

**Types of Essay Tests**

Essay tests may be divided into many types. Monree and Cater (1993) divide essay tests into the many categories like Selective recall-basis given, evaluation recall-basis given, comparison of two things on a single designated basis, comparison of two things in general, Decisions – For or against, cause and effect, explanation of the use or exact meaning of some word, phrase on statement, summary of some unit of the text book or article, analysis, statement of relationships, Illustration or examples, classification, application of rules, laws, or principles to new situation, discussion, statement of an author's purpose in the selection or organization of material, Criticism – as to the adequacy, correctness or relevance of a printed statement or to a class mate's answer to a question on the lesson, reorganization of facts, formulation of new question – problems and question raised, new methods of procedure etc.

**Types of Constructed Response Items**

Essay items can vary from very lengthy, open ended end of semester term papers or take home tests that have flexible page limits (e.g. 10-12 pages, no more than 30 pages etc.) to essays with responses limited or restricted to one page or less. Thus essay type items are of two types:-

- Restricted Response Essay Items
- Extended Response Essay Items

**I. Restricted Response Essay Items**

An essay item that poses a specific problem for which a student must recall proper information, organize it in a suitable manner, derive a defensible conclusion, and express it within the limits of posed problem, or within a page or time limit, is called a restricted response essay type item. The statement of the problem specifies response limitations that guide the student in responding and provide evaluation criteria for scoring.

**Example 1:**

List the major similarities and differences in the lives of people living in Islamabad and Faisalabad.

**Example 2:**

Compare advantages and disadvantages of lecture teaching method and demonstration teaching method.

**When Should Restricted Response Essay Items be used?**

Restricted Response Essay Items are usually used to:-

- Analyze relationship
- Compare and contrast positions
- State necessary assumptions
- Identify appropriate conclusions
- Explain cause-effect relationship
- Organize data to support a viewpoint
- Evaluate the quality and worth of an item or action
- Integrate data from several sources

## II. Extended Response Essay Type Items

An essay type item that allows the student to determine the length and complexity of response is called an extended-response essay item. This type of essay is most useful at the synthesis or evaluation levels of cognitive domain. We are interested in determining whether students can organize, integrate, express, and evaluate information, ideas, or pieces of knowledge the extended response items are used.

**Example:**

Identify as many different ways to generate electricity in Pakistan as you can? Give advantages and disadvantages of each. Your response will be graded on its accuracy, comprehension and practical ability. Your response should be 8-10 pages in length and it will be evaluated according to the RUBRIC (scoring criteria) already provided.

**Scoring Essay Type Items**

A rubric or scoring criteria is developed to evaluate/score an essay type item. A rubric is a scoring guide for subjective assessments. It is a set of criteria and standards linked to learning objectives that are used to assess a student's performance on papers, projects, essays, and other assignments. Rubrics allow for standardized evaluation according to specified criteria, making grading simpler and more transparent. A rubric may vary from simple checklists to elaborate combinations of checklist and rating scales. How

elaborative your rubric is, depends on what you are trying to measure. If your essay item is a restricted-response item simply assessing mastery of factual content, a fairly simple listing of essential points would be sufficient. An example of the rubric of restricted response item is given below.

**Test Item:**

*Name and describe five of the most important factors of unemployment in Pakistan. (10 points)*

**Rubric/Scoring Criteria:**

(i) 1 point for each of the factors named, to a maximum of 5 points
(ii) One point for each appropriate description of the factors named, to a maximum of 5 points
(iii) No penalty for spelling, punctuation, or grammatical error
(iv) No extra credit for more than five factors named or described.
(v) Extraneous information will be ignored.

However, when essay items are measuring higher order thinking skills of cognitive domain, more complex rubrics are mandatory. An example of Rubric for writing test in language is given below.

**Table 7.2: Scoring Criteria (Rubrics) for Essay Type Item for 8th grade**

| Sr. No. | Criteria | Unsatisfactory | Proficient | Advance |
|---|---|---|---|---|
| 1 | Length | Length of Text will be according to the Prompt | Length of Text will be according to the Prompt | Length of Text will be according to the Prompt |
| 2 | Layout | Writing is not according to the provided format | Writing is according to the provided format to some extent | Writing is completely according to the provided format |
| 3 | Vocabulary | Expected KEY WORDS* are not used | Expected KEY WORDS* are used to some extent | Expected KEY WORDS* are used mostly |
| 4 | Spelling | Spellings of most words are incorrect | Spellings of some words are incorrect | Spellings of all words are correct |
| 5 | Selection and Organization of Ideas | Few ideas are relevant to the task and the given task organization | Some ideas are relevant to the task and the given task organization | Almost all ideas are relevant to the task and the given task organization |

| | | | | |
|---|---|---|---|---|
| 6 | Punctuation | Very few Punctuation Marks are used | Some Punctuation Marks are used | Almost all Punctuation Marks are used |
| 7 | Grammar | Use of some basic RAMMAR RULES** | Occasional use of basic GRAMMAR RULES** | Use of some basic GRAMMAR RULES** |

* KEY WORDS: Expected Key Words will be provided for each Writing Prompt

**Advantages of Essay Type Items**

The main advantages of essay type tests are as follows:

(i) They can measures complex learning outcomes which cannot be measured by other means.
(ii) They emphasize integration and application of thinking and problem solving skills.
(iii) They can be easily constructed.
(iv) They give examines freedom to respond within broad limits.
(v) The students cannot guess the answer because they have to supply it rather than select it.
(vi) Practically it is more economical to use essay type tests if number of students is small.
(vii) They required less time for typing, duplicating or printing. They can be written on the blackboard also if number of students is not large.
(viii) They can measure divergent thinking.
(ix) They can be used as a device for measuring and improving language and expression skill of examinees.
(x) They are more helpful in evaluating the quality of the teaching process.
(xi) Studies has supported that when students know that the essay type questions will be asked, they focus on learning broad concepts and articulating relationships, contrasting and comparing.
(xii) They set better standards of professional ethics to the teachers because they expect more time in assessing and scoring from the teachers.

**Limitations of Essay Type Items**

The essay type tests have the following serious limitations as a measuring instrument:

(i) A major problem is the lack of consistency in judgments even among competent examiners.

(ii) They have Halo effects. If the examiner is measuring one characteristic, he can be influenced in scoring by another characteristic. For example, a well behaved student may score more marks on account of his good behaviour also.

(iii) They have question to question carry effect. If the examinee has answered satisfactorily in the beginning of the question or questions he is likely to score more than the one who did not do well in the beginning but did well later on.

(iv) They have examinee to examinee carry effect. A particular examinee gets marks not only on the basis of what he has written but also on the basis that whether the previous examinee whose answered book was examined by the examiner was good or bad.

(v) They have limited content validity because of sample of questions can only be asked in essay type test.

(vi) They are difficult to score objectively because the examinee has wide freedom of expression and he writes long answers.

(vii) They are time consuming both for the examiner and the examinee.

(viii) They generally emphasize the lengthy enumeration of memorized facts.

**Suggestions for Writing Essay Type Items**

I. Ask questions or establish tasks that will require the student to demonstrate command of essential knowledge. This means that students should not be asked merely to reproduce material heard in a lecture or read in a textbook. To "demonstrate command" requires that the question be somewhat novel or new. The substance of the question should be essential knowledge rather than trivia that might be a good board game question.

II. Ask questions that are determinate, in the sense that experts (colleagues in the field) could agree that one answer is better than another. Questions that contain phrases such as "What do you think..." or "What is your opinion about..." are indeterminate. They can be used as a medium for assessing skill in written expression, but because they have no clearly right or wrong answer, they are useless for measuring other aspects of achievement.

III. Define the examinee's task as completely and specifically as possible without interfering with the measurement process itself. It is possible to word an essay item so precisely that there is one and only one very brief answer to it. The imposition of such rigid bounds on the response is more limiting than it is helpful. Examinees do need guidance, however, to judge how extensive their response must to be considered complete and accurate.

IV. Generally give preference to specific questions that can be answered briefly. The more questions used, the better the test constructor can sample the domain of knowledge covered by the test. And the more responses available for scoring, the more accurate the total test scores are likely to be. In addition, brief responses can be scored more quickly and more accurately than long, extended responses, even when there are fewer of the latter type.

V. Use enough items to sample the relevant content domain adequately, but not so many that students do not have sufficient time to plan, develop, and review their

responses. Some instructors use essay tests rather than one of the objective types because they want to encourage and provide practice in written expression. However, when time pressures become great, the essay test is one of the most unrealistic and negative writing experiences to which students can be exposed. Often there is no time for editing, for rereading, or for checking spelling. Planning time is short changed so that writing time will not be. There are few, if any, real writing tasks that require such conditions. And there are few writing experiences that discourage the use of good writing habits as much as essay testing does.

VI. Avoid giving examinees a choice among optional questions unless special circumstances make such options necessary. The use of optional items destroys the strict comparability between student scores because not all students actually take the same test. Student A may have answered items 1-3 and Student B may have answered 3-5. In these circumstances the variability of scores is likely to be quite small because students were able to respond to items they knew more about and ignore items with which they were unfamiliar. This reduced variability contributes to reduced test score reliability. That is, we are less able to identify individual differences in achievement when the test scores form a very homogeneous distribution. In sum, optional items restrict score comparability between students and contribute to low score reliability due to reduced test score variability.

VII. Test the question by writing an ideal answer to it. An ideal response is needed eventually to score the responses. It if is prepared early, it permits a check on the wording of the item, the level of completeness required for an ideal response, and the amount of time required to furnish a suitable response. It even allows the item writer to determine if there is any "correct" response to the question.

VIII. Specify the time allotment for each item and/or specify the maximum number of points to be awarded for the "best" answer to the question. Both pieces of information provide guidance to the examinee about the depth of response expected by the item writer. They also represent legitimate pieces of information a student can use to decide which of several items should be omitted when time begins to run out. Often the number of points attached to the item reflects the number of essential parts to the ideal response. Of course if a definite number of essential parts can be determined, that number should be indicated as part of the question.

IX. Divide a question into separate components when there are obvious multiple questions or pieces to the intended responses. The use of parts helps examinees organizationally and, hence, makes the process more efficient. It also makes the grading process easier because it encourages organization in the responses. Finally, if multiple questions are not identified, some examinees may inadvertently omit some parts, especially when time constraints are great.

## 7.4 Administering the Test

### I. Test Assembly

We have discussed various aspects of test planning and construction. If you have written instructional objectives, constructed a test, and written items that match your objectives, then more than likely you will have a good test. All the "raw material" will be there. However, sometimes the raw material, as good as it may be, can be rendered useless because of poorly assembled and administrated test. By now you know it requires a substantial amount of time to write objectives, put together a test blueprint, and write items. It is worth a little more time to properly assemble or packages your test so that your efforts will not be wasted. Assembly of the test comprises the following steps:-

(i) Group together all item of similar format e.g. group all essay type item or MCQ's in one group.
(ii) Arrange test items from easy to hard
(iii) Space the items for easy reading
(iv) Keep items and their options on the same page of the test
(v) Position illustrations, tables, charts, pictures diagrams or maps near descriptions
(vi) Answer keys must be checked carefully
(vii) Determine how students record answers
(viii) Provide adequate and proper space for name and date
(ix) Test directions must be precised and clear
(x) Test must be proofread to make it error free
(xi) Make all the item unbiased (gender, culture, ethnic, racial etc)

**II. Reproduction of the Test**

Most test reproduction in the schools is done by photocopy machines. As you well know, the quality of such copies can vary tremendously. Regardless of how valid and reliable your test might be, poor printing/copies will not have a good impact. Take the following practical steps to ensure that time you spent constructing a valid and reliable test does not end in illegible printing.

- Manage printing of the test if test takers are large in number
- Manage photocopy from a proper/new machine
- Use good quality of the paper and printing
- Retain original test in your own custody
- Be careful while making sets of the test (staple different papers carefully)
- Manage confidentiality of the test

**III. Administration of the Test**

The test is ready. All that remains is to get the students ready and hand out the test. Here are some suggestions to help your students psychologically prepared for the test:-

- Maintain a positive attitude for achievement
- Maximize achievement motivation
- Equalize advantages to all the students

- Provide easy, comfortable and proper seats
- Provide proper system of light, temperature, air and water.
- Clarify all the rules and regulations of the examination center/hall
- Rotate distributions
- Remind the students to check their copies
- Monitor students continuously
- Minimize distractions
- Give time warnings properly
- Collect test uniformly
- Count the answer sheets, seal it in a bag and hand it over to the quarter concerned.

## IV. Test Taking Strategies

To improve test-taking skills, there are three approaches that might prove fruitful. Students need to understand the mechanics of test-taking, such as the need to carefully follow instructions, checking their work, and so forth. Second, they need to use appropriate test-taking strategies, including ways in which test items should be addressed and how to make educated guesses. Finally, they need to practice their test-taking skills to refine their abilities and to become more comfortable in testing situations. By acting upon the following strategies the students may enhance their test taking strategies:-

- Students need to follow directions carefully.
- Students need to understand how to budget their time.
- Students need to check their work.
- For each item, students need to read the entire test item and all the possible answers very carefully.
- Answer the easier questions first and persist to the end of the test.
- Students need to make educated guesses.
- Use test item formats for practice.
- Review the practice items and answer choices with students.
- Practice using answer sheets.

## V. Steps to Prevent Cheating

Cheating is a big issue while administering tests to get reliable and valid data of students learning achievement. Following steps can be followed to prevent cheating:-

i. Take special precautions to keep the test secure during preparation, storage and administration.
ii. Students should be provided sufficient space on their desks to work easily and to prevent use of helping material.
iii. If scratch paper is used have it turned in with the test.

iv. Testing hours must be watched carefully. Walk around the room periodically and observe the students what are they doing.
v. Two forms of the tests can also be used or use some items different in the test to prevent cheating.
vi. Use special seating arrangements while placing the students for the test. Provide sufficient empty spaces between students.
vii. Create and maintain a positive attitude concerning the value of tests for improving learning.

## 7.5 Scoring the Test

### Scoring Objective Test Items

If the student's answers are recorded on the test paper itself, a scoring key can be made by marking the correct answers on a blank copy of the test. Scoring then is simply a matter of comparing the columns of the answers on this master copy with the columns of answers on each student's paper. A strip key which consists merely of strips of paper, on which the columns of answers are recorded, may also be used if more convenient. These can easily be prepared by cutting the columns of answers from the master copy of the test and mounting them on strips of cardboard cut from manila folders.

When separate answer sheets are used, a scoring stencil is more convenient. This is a blank answer sheet with holes punched where the correct answers should appear. The stencil is laid over the answer sheet, and the number of the answer checks appearing through holes is counted. When this type of scoring procedure is used, each test paper should also be scanned to make certain that only one answer was marked for each item. Any item containing more than one answer should be eliminated from the scoring.

As each test paper is scored, mark each item that is scored incorrectly. With multiple choice items, a good practice is to draw a red line through the correct answers of the missed items rather than through the student's wrong answers. This will indicate to the students those items missed and at the same time will indicate the correct answers. Time will be saved and confusion avoided during discussion of the test. Marking the correct answers of the missed items is simple with a scoring stencil. When no answer check appears through a hole in the stencil, a red line is drawn across the hole.

In scoring objective tests, each correct answer is usually counted as one point, because an arbitrary weighting of items make little difference in the students' final scores. If some items are counted two points, some one point, and some half point, the scoring will be more complicated without any accompanying benefits. Scores based on such weightings will be similar to the simpler procedure of counting each item on one point. When a test consists of a combination of objective items and a few, more time-consuming, essay questions, however, more than a single point is needed to distinguish several levels of response and to reflect disproportionate time devoted to each of the essay questions.

When students are told to answer every item on the test, a student's score is simply the number of items answered correctly. There is no need to consider wrong answers or to correct for guessing. When all students answer every item on the test, the rank of the students' scores will be same whether the number is right or a correction for guessing is used.

A simplified form of item analysis is all that is necessary or warranted for classroom tests because most classroom groups consist of 20 to 40 students, an especially useful procedure to compare the responses of the ten lowest-scoring students. As we shall see later, keeping the upper and lower groups and ten students each simplifies the interpretation of the results. It also is a reasonable number for analysis in groups of 20 to 40 students. For example, with a small classroom group, like that of 20 students, it is best to use the upper and lower halves to obtain dependable data, whereas with a larger group, like that of 40 students, use of upper and lower 25 percent is quite satisfactory. For more refined analysis, the upper and lower 27 percent is often recommended, and most statistical guides are based on that percentage.

To illustrate the method of item analysis, suppose we have just finished scoring 32 test papers for a sixth-grade science unit on weather. Our item analysis might then proceed as follows:

1. ... Rank the 32 test papers in order from the highest to the lowest score.
2. ... Select the 10 papers within the highest total scores and the ten papers with the lowest total scores.
3. ... Put aside the middle 12 papers as they will not be used in the analysis.
4. ... For each test item, tabulate the number of students in the upper and lower groups who selected each alternative. This tabulation can be made directly on the test paper or on the test item card.
5. ... Compute the difficulty of each item (percentage of the students who got the item right).
6. ... Compute the discriminating power of each item (difference between the number of students in the upper and lower groups who got the item right).
7. ... Evaluate the effectiveness of distracters in each item (attractiveness of the incorrect alternatives).

Although item analysis by inspection will reveal the general effectiveness of a test item and is satisfactory for most classroom purposes, it is sometimes useful to obtain a more precise estimate of item difficulty and discriminating power. This can be done by applying relatively simple formulas to the item-analysis data.

**Computing item difficulty:**

The difficulty of a test item is indicated by the percentage of students who get the item right. Hence, we can compute item difficulty (P) by means of following formula, in which R equals the number of students who got the item right, and T equals the total number of students who tried the item.

$$P=(R/T)x\ 100$$

The discriminating power of an achievement test items refers to the degree to which it discriminates between students with high and low achievements. Item discriminating power (D) can be obtained by subtracting the number of students in the lower group who get the item right (RL) from the number of students in the upper group who get the item right (RU) and dividing by one-half the total number of students included in the item analysis (.5T). Summarized in formula form, it is:

$$D= (RU-RL)/.5T$$

An item with maximum positive discriminating power is one in which all students in the upper group get the item right and all the students in the lower group get the item wrong. This results in an index of 1.00, as follows:

$$D= (10-0)/10=1.00$$

An item with no discriminating power is one in which an equal number of students in both the upper and lower groups get the item right. This results in an index of .00, as follows:

$$D= (10-10)/10= .00$$

**Scoring Essay Type Test Items**

According to N.E. Gronlund (1990) the chief weakness of the essay test is the difficulty of scoring. The objectivity of scoring the essay questions may be improved by following a few rules developed by test experts.

a. Prepare a scoring key in advance. The scoring key should include the major points of the acceptable answer, the feature of the answer to be evaluated, and the weights assigned to each. To illustrate, suppose the question is "Describe the main elements of teaching." Suppose also that this question carries 20 marks. We can prepare a scoring key for the question as follows.
    i. Outline of the acceptable answer. There are four elements in teaching these are: the definition of instructional objectives, the identification of the entering behaviour of students, the provision of the learning experiences, and the assessment of the students' performance.
    ii. Main features of the answer and the weights assigned to each.
        - Content: Allow 4 points to each elements of teaching.
        - Comprehensiveness: Allow 2 points.
        - Logical organization: Allow 2 points.
        - Irrelevant material: Deduct upto a maximum of 2 points.
        - Misspelling of technical terms: Deduct 1/2 point for each mistake upto a maximum of 2 points.
        - Major grammatical mistakes: Deduct 1 point for each mistake upto a maximum of 2 points.

- Poor handwriting, misspelling of non-technical terms and minor grammatical errors: ignore.
  Preparing the scoring key in advance is useful since it provides a uniform standard for evaluation.

b. Use an appropriate scoring method. There are two scoring methods commonly used by the classroom teacher. The point method and the rating method.

In the point method, the teacher compares each answer with the acceptable answer and assigns a given number of points in terms of how will each answer approximates the acceptable answer. This method is suitable in a restricted response type of question since in this type each feature of the answer can be identified and given proper point values. For example: Suppose that the question is: "List five hypotheses that might explain why nations go to wars." In the question, we can easily assign a number of point values to each hypothesis and evaluate each answer accordingly.

In the rating method, the teacher reads each answer and places it in one of the several categories according to quality. For example, the teacher may set up five categories: Excellent – 10 points, good – 8 points, average – 6 points, weak – 4 points and poor – 2 points. This method is suitable in an extended response type of question since in this type we make gross judgment concerning the main features of the answer. It's a good practice to grade each feature separately and then add the point values to get the total score.

a. Read a sampling of the papers to get a 'feel' of the quality of the answers. This will give you confidence in scoring and stability in your judgment.
b. Score one question through all of the papers before going on to the next question. This procedure has three main advantages. First, the comparison of answer makes the scoring more exact and just, second, having to keep only one list of points in mind saves time and promotes accuracy and third, it avoids halo effect. A halo effect is defined as the tendency in rating a person to let one of its characteristics influence rating on other characteristics.
c. Adopt a definite policy regarding factors which may not be relevant to learning outcomes being measured. The grading of answer to essay questions is influenced by a large number of factors. These factors include handwriting, spelling, punctuation, sentence structure, style, padding of irrelevant material, and neatness. The teacher should specify which factor would or would not be taken into account and what score values would be assigned to or deducted from each factor.
d. Score the papers anonymously. Have the student record his name on the back or at the end of the paper, rather than at the top of each page. Another way is to let each student have a code number and write it on his paper instead of his name. Keeping the author of the paper unknown will decrease the bias with which the paper is graded.

## 7.6 Activities

- Suppose you are a teacher and you intend to have a quarterly test of grade 5 students in the subject of General Science. Prepare a Table of Specification highlighting hierarchy of knowledge, contents, item types and weightage.
- Locate a question paper of Pakistan Studies for class X of last year board exams and evaluate its MCQs test items with reference to the guidelines you have learnt in this unit and mention short comings.
- Develop an essay type question for class VIII students in the subject of Urdu Language to assess higher order thinking skills and prepare guidelines or scoring criteria (rubrics) for evaluators to minimize the biasness and subjectivity.

## 7.7 Self Assessment Questions

- What strategies will you adopt to plan an annual exam of your class?
- Write down your preferences of selecting Multiple Choice Questions rather than True-False test items.
- "It is difficult to minimize subjectivity or biasness while scoring/ evaluating constructed response test items". Give your point of view to support this statement.
- Write down an instructional objective of Social Study Grade V and develop an Essay type test item with rubric, a Multiple Choice Question and a short question.

## 7.8 References/Suggested Readings


Anastasi, A. (1988). *Psychological Testing* (6th ed.). New York, NY: MacMillan Publishing Company.

Burke, K. (2005). *How to Assess Authentic Learning*. Califonia: A Sage Publication Company

Butler, S. M. & McMunn, N. D. (2006). *Classroom Assessment*. San Francisco: Jossy-Bass

Case, B. J., Jorgensen, M. A. & Zucker, S. (2004). Alignment in Educational Assessment. Harcourt Assessment, Inc. Retrieved from

Cook, H. G. (2005). Aligning English Language Proficiency Tests to English Language Learning Standards: Retrieved on June 06, 2007, fromhttp//: www.ccsso.org/content/pdfs/ELPAlignmentFinalReport.pdf

**Davidson, F. & Lynch, B. K. (2002).** ***Test Craft: a Teacher's Guide to Writing and Using Language Test Specifications.*** Yale University Press

Davidson, F. and Fulcher, G. (2007). "The Common European Framework of Reference (CEFR) and the Design of Language Tests: A Matter of Effect." *Language Teaching* 40, 3, 231 - 241.

Fulcher, G. & Davidson, F. (2007). Language Testing and Assessment: An Advanced Resource Book. Routledge

Fulcher, G. & **Davidson, F.** (2009). Test Architecture, Test Retrofit. *Language Testing.* 26 (1) 123–144

Gronlund, N. E. & Linn, R. L. (2005). *Measurement and Assessment in Teaching*. New Delhi: Baba Barkha Nath Printers.

Haladyna, T. M., (1997). *Writing Test Items to Evaluate Higher Order Thinking*: USA: Allyn & Bacon

Kehoe, Jerard (1995). Writing Multiple-Choice Test Items. *Practical Assessment, Research & Evaluation*, 4(9). Retrieved on December 27, 2008 from http://PAREonline.net/getvn.asp?v=4&n=9

Kline, T. J. B. (2005). *Psychological Testing: A practical Approach to Design and Evaluation.* New Delhi: Sage Publications, Inc.

Kubiszyn, T. & Borich, G. (2003). *Educational Testing and Measurement: Classroom Application and Practice* (7th ed.). New York: John Wiley & Sons.

Linn, R. L. & Gronlund, N. E. (2003). *Measurement and Assessment in Teaching* (8th ed.). Upper Saddle River, NJ: Prentice Hall.

Livingstone, S. A. (2009). *Constructed-Response Test Questions: Why we use them; how we score them. R&D Connections* 11. Princeton, NJ: Educational Testing Service. Retrieved on September 2009 from

ww.ets.org/Media/Research/pdf/RD_Connections11.pdf

Loewenthal, K.M. (2001). *An Introductionto Psychological Tests and Scales* (2nd ed). USA: Taylor & Francis.

McMunn, N. D. & Butler, S. M. (2006). *A Teacher's Guide to Classroom Assessment: Understanding and Using Assessment to Improve Student Learning.* USA. John Wiley & Sons, Inc.

Mehrans, W.A. & Lehmann, I.J. (1984). *Measurement and Evaluation in Education and Psychology*, New York: Holt, Rinehart & Winston.

Miller, M. D., Linn, R.L., & Gronlund, N.E. (2008). *Measurement and Assessment in Teaching* (10th ed.). Upper Saddle River, NJ: Prentice Hall.

Popham, W.J. (2000). *Modern Educational Measurement: Practical Guidelines for Educational Leaders.* Needham, MA: Allyn and Bacon.

Salvia, J. & Ysseldyke, J. E. (1995). *Assessment. (6th ed*). USA. Houghton Mifflin Company.

Solórzano R.W. (2008). High Stakes Testing: Issues, Implications, and Remedies for English Language Learners. *Review of Educational Research. 78 (2)* 260–329

Tim McNamara (2000). *Language testing.* Oxford, UK: Oxford University Press ISBN: 0194372227

Xing, P. and Fulcher, G. (2007). "Reliability Assessment for two Versions of the Vocabulary Levels Test." *System* 35, 2, 182 - 191.

**UNIT–8**

# INTERPRETING TEST SCORES

**Written By:**
**Muhammad Azeem**

**Reviewed By:**
**Dr. Muhammad Tanveer Afzal**

## CONTENT

## INTRODUCTION

Raw scores are considering as points scored in test when the test is scored according to the set procedure or rubric of marking. These points are not meaningful without interpretation or further information. Criterion referenced interpretation of test scores describes students' scores with respect to certain criteria while norm referenced interpretation of test scores describes students' score relative to the test takers. Test results are generally reported to parents as a feedback of their young one's learning achievements. Parents have different academic backgrounds so results should be presented them in understandable and usable way. Among various objectives three of the fundamental purposes for testing are (1) to portray each student's developmental level within a test area, (2) to identify a student's relative strength and weakness in subject areas, and (3) to monitor time-to-time learning of the basic skills. To achieve any one of these purposes, it is important to select the type of score from among those reported that will permit the proper interpretation. Scores such as percentile ranks, grade equivalents, and percentage scores differ from one another in the purposes they can serve, the precision with which they describe achievement, and the kind of information they provide. A closer look at various types of scores will help differentiate the functions they can serve and the interpretations or sense they can convey.

## OBJECTIVES

After completing this unit, the students will be able to:

- understand what are the test score?
- understand what are the measurement scales used for test scores?
- ways of interpreting test score
- clarifying the accuracy of the test scores
- explain the meaning of test scores
- interpret test scores
- usability of test scores
- learn basic and significant concepts of statistics
- understand and usage of central tendency in educational measurements
- understand and usage of measure of variation in educational measurements
- planning and administration of test

## 8.1 Introduction of Measurement Scales and Interpretation of Test Scores

### Interpreting Test Scores

All types of research data, test result data, survey data, etc is called raw data and collected using four basic scales. Nominal, ordinal, interval and ratio are four basic scales for data collection. Ratio is more sophisticated than interval, interval is more sophisticated than ordinal, and ordinal is more sophisticated than nominal. A variable measured on a "nominal" scale is a variable that does not really have any evaluative distinction. One value is really not any greater than another. A good example of a nominal variable is gender. With nominal variables, there is a qualitative difference between values, not a quantitative one. Something measured on an "ordinal" scale does have an evaluative connotation. One value is greater or larger or better than the other. With ordinal scales, we only know that one value is better than other or 10 is better than 9. A variable measured on interval or ration scale has maximum evaluative distinction. After the collection of data, there are three basic ways to compare and interpret results obtained by responses. Students' performance can be compare and interpreted with an absolute standard, with a criterion-referenced standard, or with a norm-referenced standard. Some examples from daily life and educational context may make this clear:

| Sr. No. | Standard | Characteristics | daily life | educational context |
|---|---|---|---|---|
| 1 | Absolute | simply state the observed outcome | He is 6' and 2" tall | He spelled correctly 45 out of 50 English words |
| 2 | criterion-referenced | compare the person's performance with a standard, or criterion. | He is tall enough to catch the branch of this tree. | His score of 40 out of 50 is greater than minimum cutoff point 33. So he must promoted to the next class. |
| 3 | norm-referenced | compare a person's performance with that of other people in the same context. | He is the third fastest ballar in the pakistani squad 15. | His score of 37 out of 50 was not very good; 65% of his class fellows did better. |

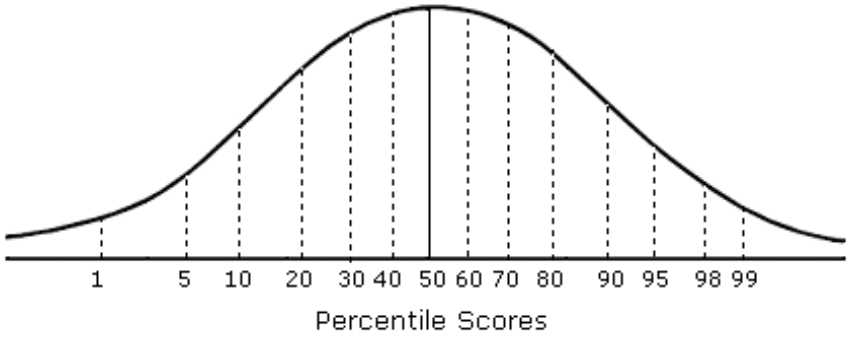
All three types of scores interpretation are useful, depending on the purpose for which comparisons made.

An absolute score merely describes a measure of performance or achievement without comparing it with any set or specified standard. Scores are not particularly useful without any kind of comparison. Criterion-referenced scores compare test performance with a specific standard; such a comparison enables the test interpreter to decide whether the

scores are satisfactory according to established standards. Norm-referenced tests compare test performance with that of others who were measured by the same procedure. Teachers are usually more interested in knowing how children compare with a useful standard than how they compare with other children; but norm-referenced comparisons may also provide useful insights.

## 8.2 Interpreting Test Scores by Percentiles

The students' scores in terms of criterion-referenced scores are most easy to understand and interpret because they are straightforward and usually represented in percentages or raw scores while norm-referenced scores are often converted to derive standard scores or converted in to percentiles. Derived standard scores are usually based on the normal curve having an arbitrary mean to compare respondents who took the same test. The conversion of students' score into student's percentile score on a test indicates what percentage of other students are fell below that student's score who took the same test. Percentiles are most often used for determining the relative standing position of any student in a population. Percentile ranks are an easy way to convey a student's standing at test relative to other same test takers.




For example, a score at the 60th percentile means that the individual's score is the same as or higher than the scores of 60% of those who took the test. The 50th percentile is known as the median and represents the middle score of the distribution.

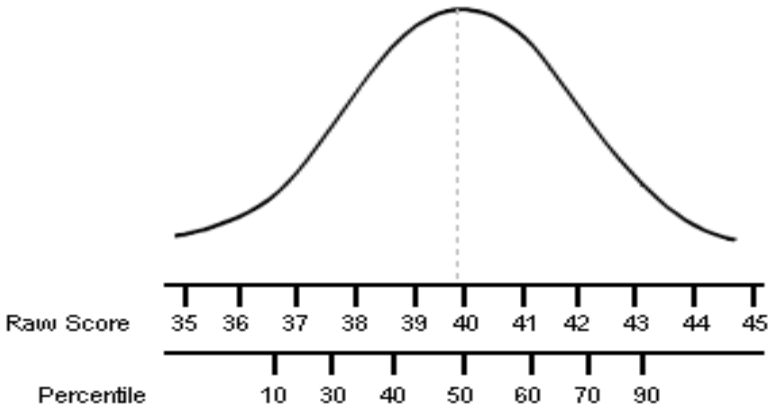
Percentiles have the disadvantage that they are not equal units of measurement. For instance, a difference of 5 percentile points between two individual's scores will have a different meaning depending on its position on the percentile scale, as the scale tends to exaggerate differences near the mean and collapse differences at the extremes.

Percentiles cannot be averaged nor treated in any other way mathematically. However, they do have the advantage of being easily understood and can be very useful when giving feedback to candidates or reporting results to managers.

If you know your percentile score then you know how it compares with others in the norm group. For example, if you scored at the 70th percentile, then this means that you scored the same or better than 70% of the individuals in the norm group.

Percentile score is easily understood when tend to bunch up around the average of the group i.e. when most of the student are the same ability and have score with very small rang.

To illustrate this point, consider a typical subject test consisting of 50 questions. Most of the students, who are a fairly similar group in terms of their ability, will score around 40. Some will score a few less and some a few more. It is very unlikely that any of them will score less than 35 or more than 45.




These results in terms of achievement scores are a very poor way of analyzing them. However, percentile score can interpret results very clearly.

Definition

A **percentile** is a measure that tells us what percent of the total frequency scored at or below that measure. A percentile rank is the percentage of scores that fall at or below a given score. OR

A **percentile** is a measure that tells us what percent of the total frequency scored below that measure. A percentile rank is the percentage of scores that fall below a given score.

Both definitions are seams to same but statistically not same. For Example

Example No.1

If Aslam stand $25^{th}$ out of a class of 150 students, then 125 students were ranked below Aslam.

***Formula:***

To find the percentile rank of a score, $x$, out of a set of $n$ scores, where $x$ is included: $\frac{(B+0.5E)}{n}.100 = \text{percentile rank}$

Where $B$ = number of scores below $x$

$E$ = number of scores equal to $x$

$n$ = number of scores

using this formula Aslam's percentile rank would be:

$$\frac{125+0.5(1)}{150} = \frac{125.5}{150} = .83\overline{6} = 84^{th}\text{ percentile}$$

***Formula:***

To find the percentile rank of a score, $x$, out of a set of $n$ scores, where $x$ is not included:

$$\frac{\text{number of scores below } x}{n}.100 = \text{percentile rank}$$

using this formula Aslam's percentile rank would be:

$$\frac{125}{150}.8\overline{3} = 83^{rd}\text{ percentile}$$

Therefore both definition yields different percentile rank. This difference is significant only for small data. If we have raw data then we can find unique percentile rank using both formulae.

**Example No.2**

The science test scores are: 50, 65, 70, 72, 72, 78, 80, 82, 84, 84, 85, 86, 88, 88, 90, 94, 96, 98, 98, 99 Find the percentile rank for a score of 84 on this test.

**Solution:**

First rank the scores in ascending or descending order

50, 65, 70, 72, 72, 78, 80, 82, 84, |84, 85, 86, 88, 88, 90, 94, 96, 98, 98, 99

Since there are 2 values equal to 84, assign one to the group "above 84" and the other to the group "below 84".

***Solution Using Formula:***

$$\frac{(B+0.5E)}{n}.100 = \text{percentile rank}$$

$$\frac{8+0.5(2)}{20}.100=\frac{9}{20}.100=45^{th}\text{ percentile}$$

***Solution Using Formula:***

$$\frac{\text{number of scores below } x}{n}.100=\text{percentile rank}$$

$$\frac{9}{100}.100=45^{th}\text{ percentile}$$

Therefore score of 84 is at the 45th percentile for this test.

**Example No.3**

The science test scores are: 50, 65, 70, 72, 72, 78, 80, 82, 84, 84, 85, 86, 88, 88, 90, 94, 96, 98, 98, 99. Find the percentile rank for a score of 86 on this test.

**Solution:**

First rank the scores in ascending or descending order

Since there is only one value equal to 86, it will be counted as "half" of a data value for the group "above 86" as well as the group "below 86".

***Solution Using Formula:***

$$\frac{(B+0.5E)}{n}.100=\text{percentile rank}$$

$$\frac{11+0.5(1)}{20}.100=\frac{11.5}{20}.100=58^{th}\text{ percentile}$$

***Solution Using Formula:***

$$\frac{\text{number of scores below } x}{n}.100=\text{percentile rank}$$

$$\frac{11.5}{20}.100=57.5=58^{th}\text{ percentile}$$

The score of 86 is at the 58th percentile for this test.

**Keep in Mind:**

- Percentile rank is a number between 0 and 100 indicating the percent of cases falling at or below that score.
- Percentile ranks are usually written to the nearest whole percent: 64.5% = 65% = $65^{th}$ percentile
- Scores are divided into 100 equally sized groups.
- Scores are arranged in rank order from lowest to highest.
- There is no 0 percentile rank - the lowest score is at the first percentile.
- There is no 100th percentile - the highest score is at the 99th percentile.
- Percentiles have the disadvantage that they are not equal units of measurement.
- Percentiles cannot be averaged nor treated in any other way mathematically.
- You cannot perform the same mathematical operations on percentiles that you can on raw scores. You cannot, for example, compute the mean of percentile scores, as the results may be misleading.
- Quartiles can be thought of as percentile measure. Remember that quartiles break the data set into 4 equal parts. If 100% is broken into four equal parts, we have subdivisions at 25%, 50%, and 75% .creating the:

First quartile (lower quartile) to be at the $25^{th}$ percentile.
Median (or second quartile) to be at the $50^{th}$ percentile.
Third quartile (upper quartile) to be a the $75^{th}$ percentile.

## 8.3 Interpreting Test Scores by Percentages

The number of questions a student gets right on a test is the student's raw score (assuming each question is worth one point). By itself, a raw score has little or no meaning. For example if teacher says that Fatima has scored 8 marks. This information (8 marks) regarding Fatima's result does not convey any meaning. The meaning depends on how many questions are on the test and how hard or easy the questions are. For example, if Umair got 10 right on both a math test and a science test, it would not be reasonable to conclude that his level of achievement in the two areas is the same. This illustrates, why raw scores are usually converted to other types of scores for interpretation purposes. The conversion of raw score into percentage convey students' achievements in understanding and meaningful way. For example if Sadia got 8 questions right out of ten questions then we can say that Sadia is able to solve

$\frac{8}{10}\times 100$=80% questions. If each question carries equal marks then we can say that Sadia has scored 80% marks. If different questions carry different marks then first count marks obtained and total marks the test. Use the following formula to compute % of marks.

$$\frac{\text{Marks Otained}}{\text{Total Marks}} \times 100 = \%\ \text{marks}$$

**Example:**

The marks detail of Hussan's math test is shown. Find the percentage marks of Hussan.

| Question | Q1 | Q2 | Q3 | Q4 | Q5 | Total |
|---|---|---|---|---|---|---|
| Marks | 10 | 10 | 5 | 5 | 20 | 50 |
| Marks obtained | 8 | 5 | 2 | 3 | 10 | 28 |

Solution:

Hussan' s marks = 28

Total marks =50

Hussan got $= \frac{\text{Marks Obtained}}{\text{Total Marks}} \times 100 = \frac{28}{50} \times 100 = 56\ \%$

For example, a number can be used merely to label or categorize a response. This sort of number (nominal scale) has a low level of meaning. A higher level of meaning comes with numbers that order responses (ordinal data). An even higher level of meaning (interval or ratio data) is present when numbers attempt to present exact scores, such as when we state that a person got 17 correct out of 20. Although even the lowest scale is useful, higher level scales give more precise information and are more easily adapted to many statistical procedures.

Scores can be summarized by using either the mode (most frequent score), the median (midpoint of the scores), or the mean (arithmetic average) to indicate typical performance. When reporting data, you should choose the measure of central tendency that gives the most accurate picture of what is typical in a set of scores. In addition, it is possible to report the standard deviation to indicate the spread of the scores around the mean.

Scores from measurement processes can be either absolute, criterion referenced, or norm referenced. An absolute score simply states a measure of performance without comparing it with any standard. However, scores are not particularly useful unless they are compared with something. Criterion-referenced scores compare test performance with a specific standard; such a comparison enables the test interpreter to decide whether the scores are satisfactory according to established standards. Norm-referenced tests compare test performance with that of others who were measured by the same procedure. Teachers

are usually more interested in knowing how children compare with a useful standard than how they compare with other children; but norm referenced comparisons may also provide useful insights.

Criterion-referenced scores are easy to understand because they are usually straightforward raw scores or percentages. Norm-referenced scores are often converted to percentiles or other derived standard scores. A student's percentile score on a test indicates what percentage of other students who took the same test fell below that student's score. Derived scores are often based on the normal curve. They use an arbitrary mean to make comparisons showing how respondents compare with other persons who took the same test.

## 8.4 Interpreting Test Scores by ordering and ranking

Organizing and reporting of students' scores start with placing the scores in ascending or descending order. Teacher can find the smallest, largest, rang, and some other facts like variability of scores associated with scores from ranked scores. Teacher may use ranked scoes to see the relative position of each student within the class but ranked scores does not yield any significant numerical value for result interpretation or reporting.

### 8.4.1 Measurement Scales

Measurement is the assignment of numbers to objects or events in a systematic fashion. Measurement scales are critical because they relate to the types of statistics you can use to analyze your data. An easy way to have a paper rejected is to have used either an incorrect scale/statistic combination or to have used a low powered statistic on a high powered set of data. Following four levels of measurement scales are commonly distinguished so that the proper analysis can be used on the data a number can be used merely to label or categorize a response.

#### 8.4.1.1 Nominal Scale

Nominal scales are the lowest scales of measurement. A nominal scale, as the name implies, is simply some placing of data into categories, without any order or structure. You are only allowed to examine if a nominal scale datum is equal to some particular value or to count the number of occurrences of each value. For example, categorization of blood groups of classmates into A, B. AB, O etc. In The only mathematical operation we can perform with nominal data is to count. Variables assessed on a nominal scale are called **categorical variables;** Categorical data are measured on nominal scales which merely assign labels to distinguish categories. For example, gender is a nominal scale variable. Classifying people according to gender is a common application of a **nominal** scale.

**Nominal Data**

- classification or gatagorization of data, e.g. male or female
- no ordering, e.g. it makes no sense to state that male is greater than female (M > F) etc
- arbitrary labels, e.g., pass=1 and fail=2 etc

#### 8.4.1.2 Ordinal Scale

Something measured on an "ordinal" scale does have an evaluative connotation. You are also allowed to examine if an ordinal scale datum is less than or greater than another value. For example rating of job satisfaction on a scale from 1 to 10, with 10 representing complete satisfaction. With ordinal scales, we only know that 2 is better than 1 or 10 is better than 9; we do not know by how much. It may vary. Hence, you can 'rank' ordinal data, but you cannot 'quantify' differences between two ordinal values. Nominal scale properties are included in ordinal scale.

**Ordinal Data**

- ordered but differences between values are not important. Difference between values may or may not same or equal.
- e.g., political parties on left to right spectrum given labels 0, 1, 2
- e.g., Likert scales, rank on a scale of 1..5 your degree of satisfaction
- e.g., restaurant ratings

#### 8.4.1.3 Interval Scale

An ordinal scale has quantifiable difference between values become interval scale. You are allowed to quantify the difference between two interval scale values but there is no natural zero. A variable measured on an interval scale gives information about more or better as ordinal scales do, but interval variables have an equal distance between each value. The distance between 1 and 2 is equal to the distance between 9 and 10. For example, temperature scales are interval data with 25C warmer than 20C and a 5C difference has some physical meaning. Note that 0C is arbitrary, so that it does not make sense to say that 20C is twice as hot as 10C but there is the exact same difference between 100C and 90C as there is between 42C and 32C. Students' achievement scores are measured on interval scale

**Interval Data**

- ordered, constant scale, but no natural zero
- differences make sense, but ratios do not (e.g., 30°-20°=20°-10°, but 20°/10° is not twice as hot!
- e.g., temperature (C,F), dates

#### 8.4.1.4 Ratio Scale

Something measured on a ratio scale has the same properties that an interval scale has except, with a ratio scaling, there is an absolute zero point. Temperature measured in Kelvin is an example. There is no value possible below 0 degrees Kelvin, it is absolute zero. Physical measurements of height, weight, length are typically ratio variables. Weight is another example, 0 lbs. is a meaningful absence of weight. This ratio hold true regardless of which scale the object is being measured in (e.g. meters or yards). This is because there is a natural zero.

**Ratio Data**

- ordered, constant scale, natural zero
- e.g., height, weight, age, length

One can think of nominal, ordinal, interval, and ratio as being ranked in their relation to one another. Ratio is more sophisticated than interval, interval is more sophisticated than ordinal, and ordinal is more sophisticated than nominal.

## 8.5 Frequency Distribution

**Frequency** is how often something occurs. The frequency (**f**) of a particular observation is the number of times the observation occurs in the data.

**Distribution**

The *distribution* of a variable is the pattern of frequencies of the observation.

**Frequency Distribution**

It is a representation, either in a graphical or tabular format, which displays the number of observations within a given interval. Frequency distributions are usually used within a statistical context.

### 8.5.1 Frequency Distribution Tables

A frequency distribution table is one way you can organize data so that it makes more sense. Frequency distributions are also portrayed as frequency tables, histograms, orpolygons. Frequency distribution tables can be used for both categorical and numeric variables. The intervals of frequency table must be mutually exclusive and exhaustive. Continuous variables should only be used with class intervals. By counting frequencies, we can make a frequency distribution table. Following examples will figure out procedure of construction of frequency distribution table.

Example 1

For example, let's say you have a list of IQ scores for a gifted classroom in a particular elementary school. The IQ scores are: 118, 123, 124, 125, 127, 128, 129, 130, 130, 133, 136, 138, 141, 142, 149, 150, 154. That list doesn't tell you much about anything. You could draw a frequency distribution table, which will give a better picture of your data than a simple list.

**Step 1**:

- Figure out how many classes (categories) you need. There are no hard rules about how many classes to pick, but there are a couple of general guidelines:
- Pick between 5 and 20 classes. For the list of IQs above, we picked 5 classes.
- Make sure you have a few items in each category. For example, if you have 20 items, choose 5 classes (4 items per category), not 20 classes (which would give you only 1 item per category).

**Step 2**:

- Subtract the minimum data value from the maximum data value. For example, our the IQ list above had a minimum value of 118 and a maximum value of 154, so:
  154 – 118 = 36

**Step 3**:

- Divide your answer in Step 2 by the number of classes you chose in Step 1. 36 / 5 = 7.2

**Step 4**:

- Round the number from Step 3 up to a whole number to get the class width. Rounded up, 7.2 becomes 8.

**Step 5**:

- Write down your lowest value for your first minimum data value: The lowest value is 118

**Step 6**:

- Add the class width from Step 4 to Step 5 to get the next lower class limit: 118 + 8 = 126

**Step 7**:

- Repeat Step 6 for the other minimum data values (in other words, keep on adding your class width to your minimum data values) until you have created the number of classes you chose in Step 1. We chose 5 classes, so our 5 minimum data values are:
  118
  126 (118 + 8)
  134 (126 + 8)
  142 (134 + 8)
  150 (142 + 8)

**Step 8**:

- Write down the upper class limits. These are the highest values that can be in the category, so in most cases you can subtract 1 from class width and add that to the minimum data value. For example:
  118 + (8 – 1) = 125
  118 – 125
  126 – 133
  134 – 142
  143 – 149
  150 – 157

**Step 9**:

- Add a second column for the number of items in each class, and label the columns with appropriate headings:

| IQ | Number |
|---|---|
| 118 – 125 | |
| 126 – 133 | |
| 134 – 142 | |
| 143 – 149 | |
| 150 – 157 | |

**Step 10**:

- Count the number of items in each class, and put the total in the second column. The list of IQ scores are: 118, 123, 124, 125, 127, 128, 129, 130, 130, 133, 136, 138, 141, 142, 149, 150, 154.

| **IQ** | **Number** |
|---|---|
| 118 – 125 | 4 |
| 126 – 133 | 6 |

134 – 142 4

143 – 149 1

150 – 157 2

**Example 2**

A survey was taken in Lahore. In each of 20 homes, people were asked how many cars were registered to their households. The results were recorded as follows:

1, 2, 1, 0, 3, 4, 0, 1, 1, 1, 2, 2, 3, 2, 3, 2, 1, 4, 0, 0

Use the following steps to present this data in a frequency distribution table.

1. Divide the results (***x***) into intervals, and then count the number of results in each interval. In this case, the intervals would be the number of households with no car (0), one car (1), two cars (2) and so forth.
2. Make a table with separate columns for the interval numbers (the number of cars per household), the tallied results, and the frequency of results in each interval. Label these columns *Number of cars*, *Tally* and *Frequency*.
3. Read the list of data from left to right and place a tally mark in the appropriate row. For example, the first result is a 1, so place a tally mark in the row beside where 1 appears in the interval column (*Number of cars*). The next result is a 2, so place a tally mark in the row beside the 2, and so on. When you reach your fifth tally mark, draw a tally line through the preceding four marks to make your final frequency calculations easier to read.
4. Add up the number of tally marks in each row and record them in the final column entitled *Frequency*.

Your frequency distribution table for this exercise should look like this:

**Table 1. Frequency table for the number of cars registered in each household**

| Number of cars (x) | Tally | Frequency (f) |
|---|---|---|
| 0 | 𝍸 | 4 |
| 1 | 𝍸 I | 6 |
| 2 | 𝍸 | 5 |
| 3 | III | 3 |
| 4 | II | 2 |

By looking at this frequency distribution table quickly, we can see that out of 20 households surveyed, 4 households had no cars, 6 households had 1 car, etc.

**Relative frequency and percentage frequency**

An analyst studying these data might want to know not only how long batteries last, but also what proportion of the batteries falls into each class interval of battery life.

This *relative frequency* of a particular observation or class interval is found by dividing the frequency (**f**) by the number of observations (**n**): that is, (**f ÷ n**). Thus:

**Relative frequency = frequency ÷ number of observations**

The *percentage frequency* is found by multiplying each relative frequency value by 100. Thus:

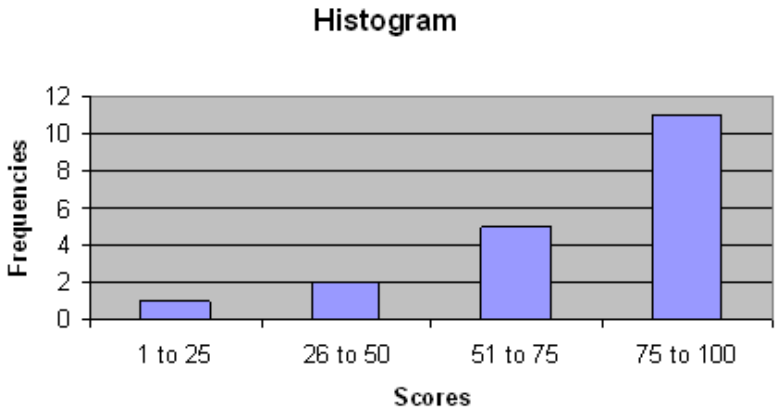
**Percentage frequency = relative frequency X 100 =** f ÷ n X 100

**8.6 Interpreting Test Scores by Graphic Displays of Distributions**

The data from a frequency table can be displayed graphically. A graph can provide a visual display of the distributions, which gives us another view of the summarized data. For example, the graphic representation of the relationship between two different test scores through the use of scatter plots. We learned that we could describe in general terms the direction and strength of the relationship between scores by visually examining the scores as they were arranged in a graph. Some other examples of these types of graphs include histograms and frequency polygons.

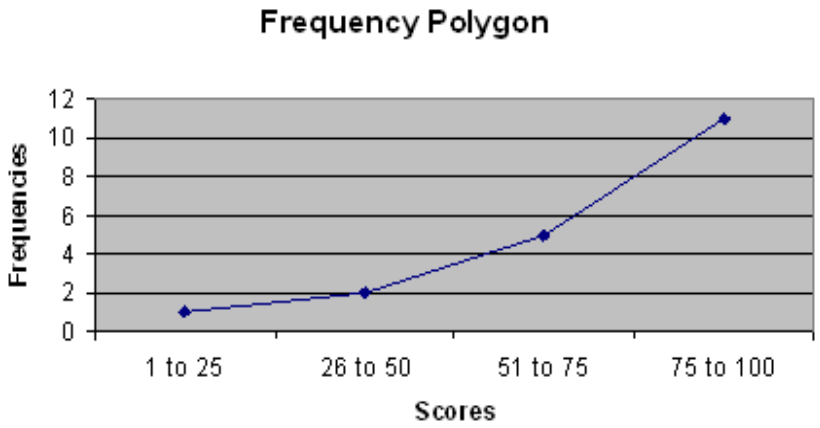
A **histogram** is a bar graph of scores from a frequency table. The horizontal x-axis represents the scores on the test, and the vertical y-axis represents the frequencies. The frequencies are plotted as bars.

**Histogram of Mid-Term Language Arts Exam**



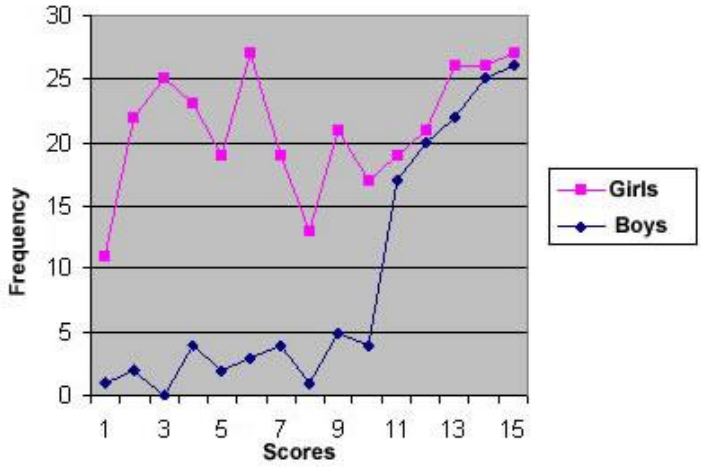
A **frequency polygon** is a line graph representation of a set of scores from a frequency table. The horizontal x-axis is represented by the scores on the scale and the vertical y-axis is represented by the frequencies.

**Frequency Polygon of Mid-Term Language Arts Exam**




A frequency polygon could also be used to compare two or more sets of data by representing each set of scores as a line graph with a different color or pattern. For example, you might be interested in looking at your students' scores by gender, or comparing students' performance on two tests (see Figure 9.4).

**Frequency Polygon of Midterm by Gender**




Frequency polygons are a graphical device for understanding the shapes of distributions. They serve the same purpose as histograms, but are especially helpful in comparing sets of data. Frequency polygons are also a good choice for displaying cumulative frequency distributions.

To create a frequency polygon, start just as for histograms, by choosing a class interval. Then draw an X-axis representing the values of the scores in your data. Mark the middle of each class interval with a tick mark, and label it with the middle value represented by the class. Draw the Y-axis to indicate the frequency of each class. Place a point in the middle of each class interval at the height corresponding to its frequency. Finally, connect the points. You should include one class interval below the lowest value in your data and one above the highest value. The graph will then touch the X-axis on both sides.

A frequency polygon for 642 psychology test scores is shown in Figure 1. The first label on the X-axis is 35. This represents an interval extending from 29.5 to 39.5. Since the lowest test score is 46, this interval has a frequency of 0. The point labeled 45 represents the interval from 39.5 to 49.5. There are three scores in this interval. There are 150 scores in the interval that surrounds 85.

You can easily discern the shape of the distribution from Figure 1. Most of the scores are between 65 and 115. It is clear that the distribution is not symmetric inasmuch as good scores (to the right) trail off more gradually than poor scores (to the left). In the terminology of Chapter 3 (where we will study shapes of distributions more systematically), the distribution is **skewed**.

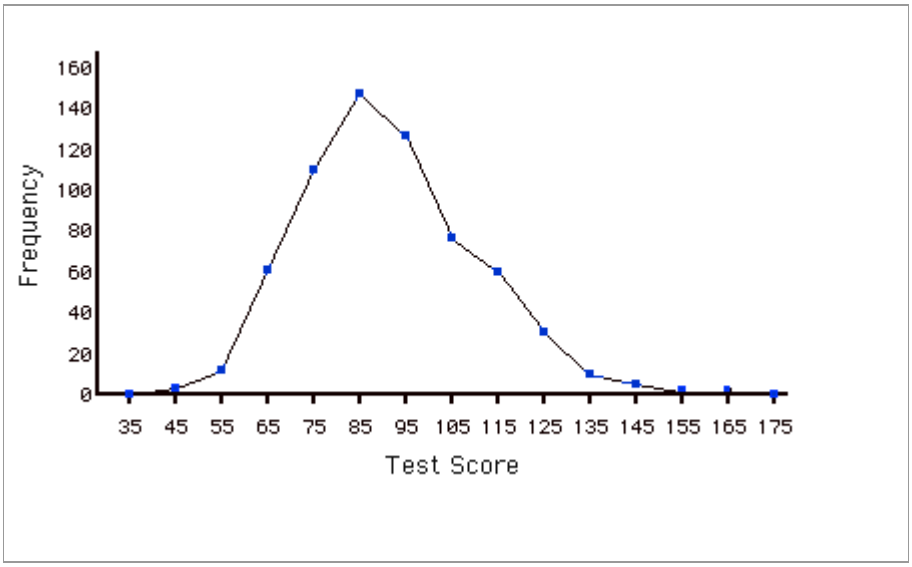



**Figure 1:** Frequency polygon for the psychology test scores.

A cumulative frequency polygon for the same test scores is shown in Figure 2. The graph is the same as before except that the Y value for each point is the number of students in the corresponding class interval plus all numbers in lower intervals. For example, there are no scores in the interval labeled "35," three in the interval "45,"and 10 in the interval "55."Therefore the Y value corresponding to "55" is 13. Since 642 students took the test, the cumulative frequency for the last interval is 642.

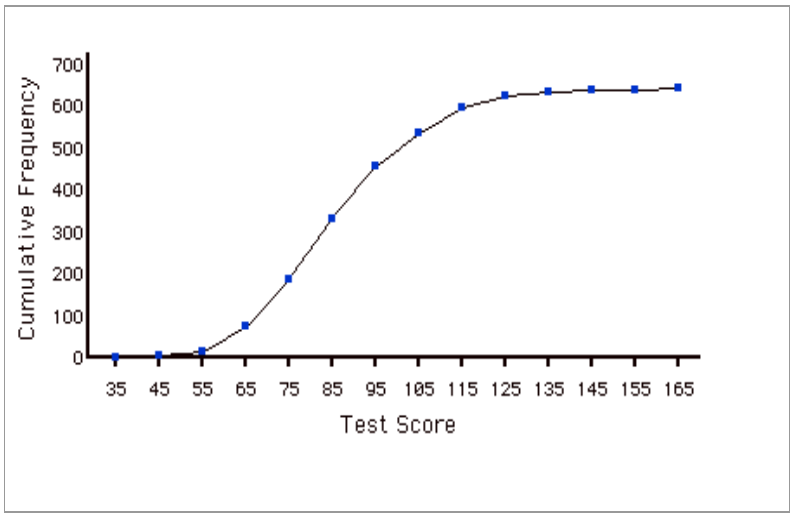



***Figure 2: Cumulative frequency polygon for the psychology test scores.***

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets. Figure 3 provides an

example. The data come from a task in which the goal is to move a computer mouse to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial. The two distributions (one for each target) are plotted together in Figure 3. The figure shows that although there is some overlap in times, it generally took longer to move the mouse to the small target than to the large one.

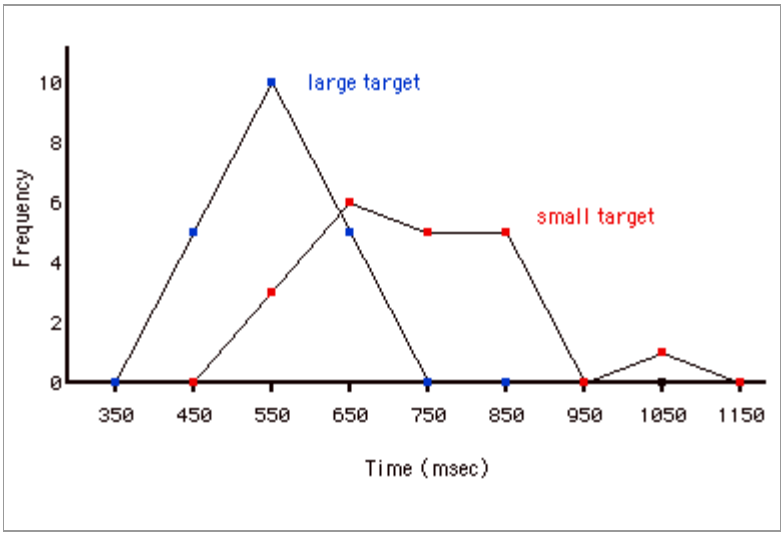



***Figure 3: Overlaid frequency polygons.***

It is also possible to plot two cumulative frequency distributions in the same graph. This is illustrated in Figure 4 using the same data from the mouse task. The difference in distributions for the two targets is again evident.

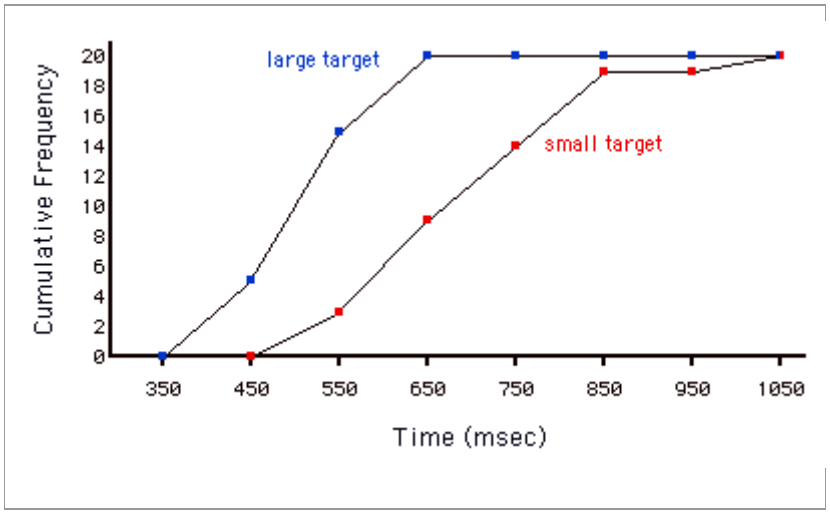


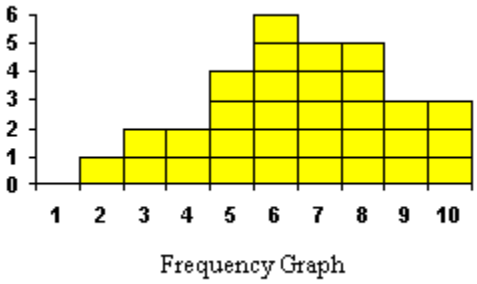
***Figure 4: Overlaid cumulative frequency polygons.***

The raw scores for the 10 pt. quiz are:

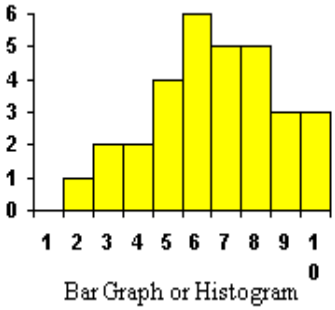
10 9 8 8 7 7 6 6 5 4 2 10 9 8 8 7 6 6 5 5 3 10 9 8 7 7 6 6 5 4 3

Draw frequency graph, bar graph, frequenvy  polygone, and frequency curve

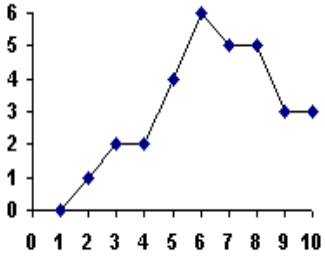
Solution



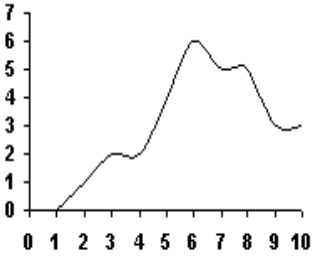
Frequency Graph



Bar Graph or Histogram



Frequency Polygon



Frequency Curve

## 8.7 Measures of Central Tendency

Suppose that a teacher gave the same test to two different classes and following results are obtained:

Class 1: 80%, 80%, 80%, 80%, 80%

Class 2: 60%, 70%, 80%, 90%, 100%

If you calculate the mean for both sets of scores, you get the same answer: 80%. But the data of two classes from which this mean was obtained was very different in the two cases. It is also possible that two different data sets may have same mean, median, and mode. For example:

Class A: 72 73 76 76 78

Class B: 67 76 76 78 80

Therefore class A and class B has same mean, mode, and median.

The way that statisticians distinguish such cases as this is known as measuring the variability of the sample. As with measures of central tendency, there are a number of ways of measuring the variability of a sample.

Probably the simplest method is to find the range of the sample, that is, the difference between the largest and smallest observation. The range of measurements in Class 1 is 0, and the range in class 2 is 40%. Simply knowing that fact gives a much better understanding of the data obtained from the two classes. In class 1, the mean was 80%, and the range was 0, but in class 2, the mean was 80%, and the range was 40%.

Statisticians use summary measures to describe patterns of data. **Measures of central tendency** refer to the summary measures used to describe the most "typical" value in a set of values.

Here, we are interested in the typical, most representative score. There are three most common measures of central tendency are mean, mode, and median. A teacher should be familiar with these common measures of central tendencies.

### 8.7.1 Mean

The mean is simply the arithmetic average. It is sum of the scores divided by the number of scores. it is computed by adding all of the scores and dividing by the number of scores. When statisticians talk about the mean of a population, they use the Greek letter μ to refer to the mean score. When they talk about the mean of a sample, statisticians use the symbol $\overline{X}$ to refer to the mean score.

It is symbolized as: $$\overline{X} = \frac{\sum X}{N}$$

$\overline{X}$ (read as "X-Bar") when computed on a sample

Computation - Example: find the mean of 2,3,5, and 10.

$$\overline{X} = \frac{\sum X}{N} = \frac{2+3+5+10}{4} = \frac{20}{4} = 5$$

Since means are typically reported with one more digit of accuracy that is present in the data, I reported the mean as 5.0 rather than just 5.

**Example 1**

The marks of seven students in a mathematics test with a maximum possible mark of 20 are given below:

15 13 18 16 14 17 12

Find the mean of this set of data values.

**Solution***:*

$$\begin{aligned}\text{Mean} &= \frac{\text{Sum of all data values}}{\text{Number of data values}} \\ &= \frac{15+13+18+16+14+17+12}{7} \\ &= \frac{105}{7} \\ &= 15\end{aligned}$$

So, the mean mark is 15.

Symbolically, we can set out the solution as follows:

$$\begin{aligned}\overline{x} &= \frac{\sum x}{n} \\ &= \frac{15+13+18+16+14+17+12}{7} \\ &= \frac{105}{7} \\ &= 15\end{aligned}$$

So, the mean mark is 15.

When working with grouped frequency distributions, we can use an approximation:

$$\overline{X} = \frac{\sum (\text{Mdpt} * f)}{N}$$

Where Mdpt. is midpoint of the group

**For example:**

| Interval | Midpoint | f | Mid*f |
|---|---|---|---|
| 95-99 | 97 | 1 | 97 |
| 90-94 | 92 | 3 | 276 |
| 85-89 | 87 | 5 | 435 |
| 80-84 | 82 | 6 | 492 |
| 75-79 | 77 | 4 | 308 |
| 70-74 | 72 | 3 | 216 |
| 65-69 | 67 | 1 | 67 |
| 60-64 | 62 | 2 | 124 |
| | | **f=25=N** | **Mid*f=2015** |

$$\bar{X}=\frac{\sum(Mdpt*f)}{N}$$

$$\bar{X}=\frac{2015}{25}=80.6$$

When computed on the raw data, we get:

$$\bar{X}=\frac{\sum X}{N}=\frac{2014}{25}=80.56$$

Thus the formula for computing the mean with grouped data gives us a good approximation of the actual mean. In fact, when we report the mean with one decimal more accuracy than what is in the data, the two techniques give the same result.

### 8.7.2 **Median** or $\mathbf{M_d}$

The score that cuts the distribution into two equal halves (or the middle score in the distribution).

The **median** of a set of data values is the middle value of the data set when it has been arranged in ascending order. That is, from the smallest value to the highest value.

**Example**

The marks of nine students in a geography test that had a maximum possible mark of 50 are given below:

47 35 37 32 38 39 36 34 35

Find the median of this set of data values.

**Solution:**

Arrange the data values in order from the lowest value to the highest value:

32 34 35 35 36 37 38 39 47

The fifth data value, 36, is the middle value in this arrangement.

$\therefore$ Median = 36

***In general:***

Median = $\frac{1}{2}(n+1)$th value, where n is the number of data values in the sample.

If the number of values in the data set is even, then the **median** is the average of the two middle values.

**Fortunately, there is a <u>formula</u>** to take care of the more complicated situations, including computing the median for grouped frequency distributions.

$$M_d = L + \left( \frac{N/2 - n_b}{n_w} \right) i$$

**Where:**

| | |
|---|---|
| **L** | = **L**ower exact limit of the interval containing $M_d$. |
| $\mathbf{n_b}$ | = **n**umber of scores **b**elow L. |
| $\mathbf{n_w}$ | = **n**umber of scores **w**ithin the interval containing $M_d$. |
| **i** | = the width of the **i**nterval (for ungrouped data i=1). |
| **N** | = the **N**umber of scores. |

**Using our last example:**

$$M_d = L + \left( \frac{\frac{N}{2} - n_b}{n_w} \right) i$$

$$= 4.5 + \left( \frac{\frac{6}{2} - 1}{3} \right) 1 = 4.5 + \left( \frac{3-1}{3} \right)$$

$$= 4.5 + \left( \frac{2}{3} \right) = 4.5 + .67 = 5.2$$

### 8.7.3 Mode

Mode is the most frequently occurring score. Note:

- There can be more than one. Can have bi- or tri-modal distributions and then speak of major and minor modes.
- It is symbolized as $M_o$.

Example: Find the mode of 2,2,6,0,9 6,8 5,4,5,4,6,4,7,4

Solution: 4 is most frequent occurring score therefore mode is 4.

## 8.8 Measures of Variability

*Variability* refers to the extent to which the scores in a distribution differ from each other. An equivalent definition (that is easier to work with mathematically) says that variability refers to the extent to which the scores in a distribution differ from their mean. If a distribution is lacking in variability, we may say that it is *homogenous* (note the opposite would be *heterogenous*).

We will discuss four measures of variability for now: the *range*, *mean* or *average deviation*, *variance* and *standard deviation*.

### 8.8.1 Range

Probably range is the simplest method to find variability of the sample, that is, the difference between the largest/maximum/highest and smallest/minimum/lowest observation.

Range = Highest value - Lowest value

**$R = X_H - X_L$**

**Example:**

The range of the saleem's four tests scores (3, 5, 5, 7) is:

$$X_H = 7 \text{ and } X_L = 3$$

Therefore $R = X_H - X_L = 7 - 3 = 4$

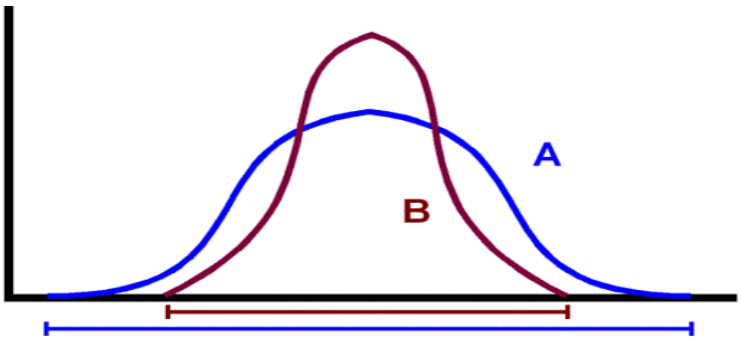
**Example**

Consider the previous example in which results of the two different classes are:

Class 1: 80%, 80%, 80%, 80%, 80%

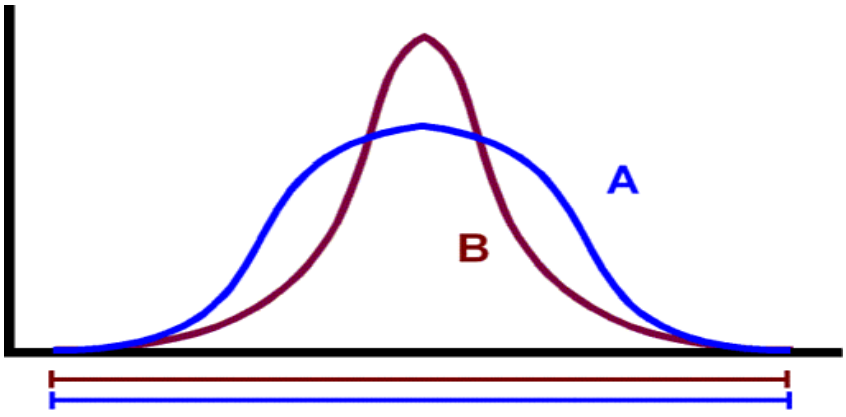
Class 2: 60%, 70%, 80%, 90%, 100%

The range of measurements in Class 1 is 0, and the range in class 2 is 40%. Simply knowing that fact gives a much better understanding of the data obtained from the two classes. In class 1, the mean was 80%, and the range was 0, but in class 2, the mean was 80%, and the range was 40%. The relationship between rang and variability can be graphically show as:




Distribution A has a larger range (and more variability) than Distribution B.

Because only the two extreme scores are used in computing the range, however, it is a crude measure. For example:




The range of Distribution A and B is the same, although Distribution A has more variability.

**Co-efficient of Range**

It is relative measure of dispersion and is based on the value of range. It is also called range co-efficient of dispersion. It is defined as:

Co-efficient of Range = (XH – XL) / (XH + XL)

Let us take two sets of observations. Set A contains marks of five students in Mathematics out of 25 marks and group B contains marks of the same student in English out of 100 marks.

Set A: 10, 15, 18, 20, 20

Set B: 30, 35, 40, 45, 50

The values of range and co-efficient of range are calculated as:

| | **Range** | **Coefficient of Range** |
|---|---|---|
| **Set A: (Mathematics)** | 20–10=10 | $\frac{20-10}{20+10}=0.33$ |
| **Set B: (English)** | 50–30=20 | $\frac{50-30}{50+30}=0.25$ |

In set A the range is 10 and in set B the range is 20. Apparently it seems as if there is greater dispersion in set B. But this is not true. The range of 20 in set B is for large observations and the range of 10 in set A is for small observations. Thus 20 and 10 cannot be compared directly. Their base is not the same. Marks in Mathematics are out of 25 and marks of English are out of 100. Thus, it makes no sense to compare 10 with 20. When we convert these two values into coefficient of range, we see that coefficient of range for set A is greater than that of set B. Thus there is greater dispersion or variation in set A. The marks of students in English are more stable than their marks in Mathematics.

### 8.8.2 Mean Deviation

If a deviation (**MD**) is the difference of a score from its mean and variability is the extent to which the scores differ from their mean, then summing all the deviations and dividing by the number of them should give us a measure of variability. The problem though is that the deviations sum to zero. However, computing the absolute value of the deviations before summing them eliminates this problem. Thus, the formula for the MD is given by:

$$M.D = \frac{\sum|x|}{N} = \frac{\sum\left|X - \overline{X}\right|}{N}$$

Thus for sample data in which the suitable average is the $\overline{X}$, the mean deviation ($M.D$) is given by the relation:

$$M.D = \frac{\sum \left| X - \overline{X} \right|}{n}$$

For frequency distribution, the mean deviation is given by

$$M.D = \frac{\sum f \left| X - \overline{X} \right|}{\sum f}$$

**Example:**

Calculate the mean deviation from arithmetic mean in respect of the marks obtained by nine students gives below and show that the mean deviation from median is minimum.

Marks (out of 25): 7, 4, 10, 9, 15, 12, 7, 9, 7

**Solution:**

After arranging the observations in ascending order, we get

Marks: 4, 7, 7, 7, 9, 9, 10, 12, 15

$$Mean = \frac{\sum X}{n} = \frac{80}{9} = 8.89$$

| Marks X | $\left\|X-\overline{X}\right\|$ |
|---|---|
| 4 | 1.89 |
| 7 | 1.89 |
| 7 | 1.89 |
| 7 | 1.89 |
| 9 | 0.11 |
| 9 | 0.11 |
| 10 | 1.11 |
| 12 | 3.11 |
| 15 | 6.11 |
| **Total** | 21.11 |

$$M.D\,\text{from mean} = \frac{\sum\left|X-\overline{X}\right|}{n} = \frac{21.11}{9} = 2.35$$

### 8.8.3 Variance

Variance is another absolute measure of dispersion. It is defined as **the average of the squared difference between each of the observations in a set of data and the mean**. For a sample data the variance is denoted is denoted by $S^2$ and the population variance is denoted by $\sigma^2$ (sigma square).

That is:

$$\text{V=MS=}\frac{\text{SS}}{\text{N}}=\frac{\sum x^2}{\text{N}}=\frac{\sum(\text{X-}\bar{\text{X}})^2}{\text{N}}$$

Thus another name for the Variance is the ***Mean of the Squared Deviations About the Mean*** (or more simply, the ***Mean of Squares (MS)***). The problem with the MS is that its units are squared and thus represent space, rather than a distance on the X axis like the other measures of variability.

**Example:**

Calculate the variance for the following sample data: 2, 4, 8, 6, 10, and 12.

**Solution:**

| **X** | $\left\|X-\overline{X}\right\|^2$ |
|---|---|
| **2** | $(2-7)^2 = 25$ |
| **4** | $(4-7)^2 = 9$ |
| **8** | $(8-7)^2 = 1$ |
| **6** | $(6-7)^2 = 1$ |
| **10** | $(10-7)^2 = 9$ |
| **12** | $(12-7)^2 = 25$ |
| $\Sigma X=42$ | $\Sigma\left(X-\overline{X}\right)^2 =70$ |

$$\overline{X} = \frac{\Sigma X}{n} = \frac{42}{6} = 7$$

$$S^2 = \frac{\Sigma\left(X-\overline{X}\right)^2}{n}$$

$$S^2 = \frac{\Sigma\left(X-\overline{X}\right)^2}{n}$$

$$S^2 = \frac{70}{6} = \frac{35}{3} = 11.67$$

**Variance = $S^2$ = 11.67**

Variance is another absolute measure of dispersion. It is defined as **the average of the squared difference between each of the observations in a set of data and the mean.**

### 8.8.4 Standard Deviation

The standard deviation is defined as **the positive square root of the mean of the square deviations taken from arithmetic mean of the data.**

A simple solution to the problem of the MS representing a space is to compute its square root. That is:

$$SD=\sqrt{V}=\sqrt{MS}=\sqrt{\frac{SS}{N}}=\sqrt{\frac{\sum \chi^2}{N}}=\sqrt{\frac{\sum (X-\bar{X})^2}{N}}$$

Since the standard deviation can be very small, it is usually reported with 2-3 more decimals of accuracy than what is available in the original data.

The standard deviation is in the same units as the units of the original observations. If the original observations are in grams, the value of the standard deviation will also be in grams. The standard deviation plays a dominating role for the study of variation in the data. It is a very widely used measure of dispersion. It stands like a tower among measure of dispersion. As far as the important statistical tools are concerned, the first important tool is the mean $\bar{x}$ and the second important tool is the standard deviation $S$ . It is based on all the observations and is subject to mathematical treatment. It is of great importance for the analysis of data and for the various statistical inferences.

**Properties of the Variance & Standard Deviation**:

1. Are always positive (or zero).
2. Equal zero when all scores are identical (i.e., there is no variability).
3. Like the mean, they are sensitive to all scores.

**Example: in previous example**

**Variance = $S^2$ = 11.67**

**Therefore SD= S = $\sqrt{S^2}$ = $\sqrt{11.67}$ = 3.41**

### 8.8.9 Estimation

Estimation is the goal of inferential statistics. We use sample values to estimate population values. The symbols are as follows:

| Measure | Sample | Population |
|---|---|---|
| Mean | $\overline{X}$ | μ |
| Variance | $s^2$ | $x^2$ |
| Standard Deviation | s | x |

It is important that the sample values (estimators) be unbiased. An ***unbiased estimator*** of a parameter is one whose average over all possible random samples of a given size equals the value of the parameter.

While $\overline{X}$ is an unbiased estimator of **x**, $\mathbf{s^2}$ is not an unbiased estimator of $\mathbf{x^2}$.

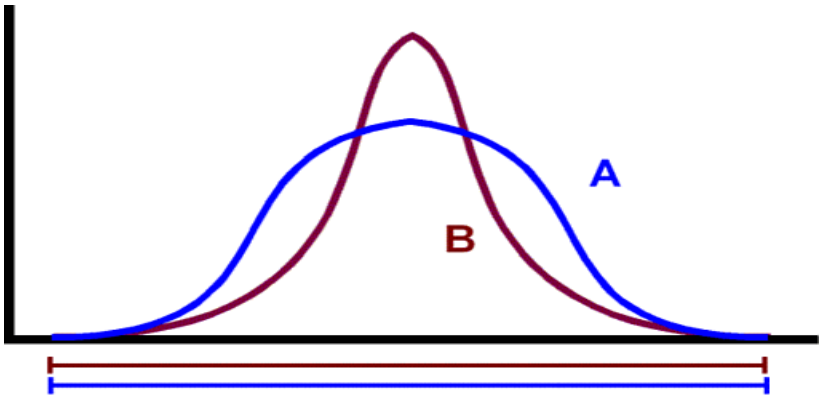
In order to make it an unbiased estimator, we use N-1 in the denominator of the formula rather than just N. Thus:

$$s^2=V=MS=\frac{\sum\left(X-\bar{X}\right)^2}{N-1}=\frac{\sum\chi^2}{N-1}$$

Note that this is a *defining* formula and, as we will see below, is not the best choice when actually doing the calculations.

**Overall Example**

Let's reconsider an example from above of two distributions (A & B):



Consider a possibility for the scores that go with these distributions:

| **Distribution** | **A** | **B** |
|---|---|---|
| **Data** | 150 | 150 |
| | 145 | 110 |
| | 100 | 100 |
| | 100 | 100 |
| | 55 | 90 |
| | 50 | 50 |
| | **600** | **600** |
| **N** | **6** | **6** |
| $\overline{X}$ | **100** | **100** |
| **Range** | **150-50+1=101** | **150-50+1=101** |

Notice that the central tendency and range of the two distributions are the same. That is, the mean, median, and mode all equal 100 for both distributions and the range is 101 for both distributions. However, while Distributions A and B have the same measures of central tendency and the same range, they differ in their variability. Distribution A has more of it. Let us prove this by computing the standard deviation in each case. First, for Distribution A:

| | A | $\overline{X}$ | X | $X^2$ |
|---|---|---|---|---|
| | 150 | 100 | 50 | 2500 |
| | 145 | 100 | 45 | 2025 |
| | 100 | 100 | 0 | 0 |
| | 100 | 100 | 0 | 0 |
| | 55 | 100 | -45 | 2025 |
| | 50 | 100 | -50 | 2500 |
| | **600** | | **0** | **9050** |
| **N** | **6** | | | |

Plugging the appropriate values into the *defining formula* gives:

| Measure | A |
|---|---|
| $s^2 = \frac{\sum \chi^2}{N-1}$ | $= \frac{9050}{6-1} = \frac{9050}{5} = 1810$ |
| $s = \sqrt{s^2}$ | $= \sqrt{1810} = 42.50$ |

Note that calculating the variance and standard deviation in this manner requires computing the mean and subtracting it from each score. Since this is not very efficient and can be less accurate as a result of rounding error, a *computational formula* is typically used. It is given as follows:

$$s = \sqrt{\frac{N\sum X^2 - (\sum X)^2}{N(N-1)}}$$

Redoing the computations for Distribution A in this manner gives:

| | A | $X^2$ |
|---|---|---|
| | 150 | 22500 |
| | 145 | 21025 |
| | 100 | 10000 |
| | 100 | 10000 |
| | 55 | 3025 |
| | 50 | 2500 |
| | **600** | **69050** |
| **N** | **6** | |

Then, plugging in the appropriate values into the computational formula gives:

$$s=\sqrt{\frac{6\times 69{,}050-(600)^2}{6(6-1)}}$$
$$=\sqrt{\frac{414{,}300-360{,}000}{6(5)}}$$
$$=\sqrt{\frac{54{,}300}{30}}=\sqrt{1810}=42.5$$

Note that the defining and computational formulas give the same result, but the computational formula is easier to work with (and potentially more accurate due to less rounding error).

Doing the same calculations for Distribution B yields:

| | **B** | **$X^2$** |
|---|---|---|
| | 150 | 22500 |
| | 110 | 12100 |
| | 100 | 10000 |
| | 100 | 10000 |
| | 90 | 8100 |
| | 50 | 2500 |
| | **600** | **65200** |
| **N** | **6** | |

Then, plugging in the appropriate values into the computational formula gives:

$$s=\sqrt{\frac{6\times 65{,}200-(600)^2}{6(6-1)}}$$
$$=\sqrt{\frac{391{,}200-360{,}000}{6(5)}}$$
$$=\sqrt{\frac{31{,}200}{30}}=\sqrt{1040}=32.25$$

**8.10 Planning the Test**

One essential step in planning a test is to decide why you are giving the test. (The word "test" is used although we are using it in a broad sense that includes performance assessments as well as traditional paper and pencil tests.)

Are you trying to sort the students (so you can compare them, giving higher scores to better students and lower scores to poor students)? If so, you will want to include some difficult questions that you expect only a few of the better students will be able to answer correctly. Or do you want to know how many of the students have mastered the content? If your purpose is the latter, you have no need to distribute the scores, so very difficult questions are unnecessary. You will, however, have to decide how many correct answers are needed to demonstrate mastery. Another way to address the "why" question is to identify if this is to be a formative assessment to help you diagnose students' problems and guide future instruction, or a summative measure to determine grades that will be reported to parents.

Airasian (1994) lists six decisions usually made by the classroom teacher in the test development process: 1. what to test, 2. how much emphasis to give to various objectives, 3. what type of assessment (or type of questions) to use, 4. how much time to allocate for the assessment, 5. how to prepare the students, and 6. whether to use the test from the textbook publisher or to create your own. Other decisions, such as whether to use a separate answer sheet, arise later.

You, as the teacher, decide what to assess. The term "assess" is used here because the term "assess" is frequently associated only with traditional paper and pencil assessments, to the exclusion of alternative assessments such as performance tasks and portfolios. Classroom assessments are generally focused on content that has been covered in the class, either in the immediate past or (as is the case with unit, semester, and end-of-course tests) over a longer period of time. For example, if we were constructing a test for preservice teachers on writing test questions, we might have the following objectives:

The student will:

1. Know the advantages and disadvantages of the major selection-types of questions.
2. Be able to differentiate between well and poorly written selection-type questions.
3. Be able to construct appropriate selection-type questions using the guidelines and rules that were presented in class.

We could have listed only the topics we have covered (e.g., true-false questions, short-

answer questions, multiple-choice questions, and test format) instead of the objectives.

Now that we have made the what decision, we can move to the next step: deciding how much emphasis to place on each objective. We can look at the amount of time in class we have devoted to each objective. We can also review the number and types of assignments the students have been given. For this example, let's assume that 20% of the assessment will be based on knowing the advantages and disadvantages, 40% will be on differentiating between well written and poorly written questions, and the other 40% will be on writing good questions. Now our planning can be illustrated with the use of a table of specifications (also called a test plan or a test blueprint) as shown in table below.

**Table of Specifications:**

| **Objectives/Content area/Topics** | **Knowledge** | **Comprehension** | **Application** | **# items/ % of test** |
|---|---|---|---|---|
| 1. Know the advantages & disadvantages of the major selection-types of questions. | | | | 20% |
| 2. Be able to differentiate between well and poorly written selection-type questions | | | | 40% |
| 3. Be able to construct appropriate selection-type questions using the guidelines and rules that were presented in class. | | | | 40% |

A table of specifications is a two-way table that matches the objectives or content you

have taught with the level at which you expect students to perform. It contains an estimate of the percentage of the test to be allocated to each topic at each level at which it is to be measured. In effect we have established how much emphasis to give to each objective or topic.

In estimating the time needed for this test, students would probably need from 5 to 10 minutes for the 20 True-False questions (15-30 seconds each), 5-7 1/2 minutes for the five comprehension questions (60-90 seconds each), and 20-30 minutes (rough estimate) to read the material and write the four questions measuring application. The total time needed would be from 30 to 48 minutes. If you are a middle or high school teacher, estimated response time is an important consideration. You will need to allow enough time for the slowest students to complete your test, and it will need to fit within a single class period.

Another consideration in planning a classroom test may be alignment with standardized tests used in your state to measure similar areas of student learning. How are those tests constructed? What objectives are measured on those tests? How are they measured; i.e., what kinds of items are used and what levels of learning (knowledge, comprehension, application, etc.) are emphasized? On your classroom test you need to measure what you have taught in the ways you have taught it, but in both the teaching and the testing, consider that your work is part of a broader educational system.

The final step in planning the test will be to write the test questions. If more information is needed on item writing, please consult the other modules that correspond to the types of questions of interest to you.

**Accommodations**

Accommodations may be needed for some of your students. It is helpful to keep those students in mind as you plan your assessments. Some examples of accommodations include:

Providing written instructions for students with hearing problems

Using large print, reading or recording the questions on audiotape (The student could record the answers on tape.)

Having an aide or assistant write/mark the answers for the student who has coordination problems, or having the student record the answers on audiotape or type the answers Using written assessments for students with speech problems

Administering the test in sections if the entire test is too long for the attention of a student Asking the students to repeat the directions to make sure they understand what they are to do

Starting each sentence on a new line helps students identify it as a new sentence

Including an example with each type of question, showing how to mark answers

## 8.11 Constructing and Assembling The Test

- Before beginning to construct your own test, you may want to compare your table of specifications with test items provided by the publisher or other sources to see what, if anything, from those sources can be incorporated into your assessment.
- Begin with simpler item types, then proceed to more complex, from easy to difficult, from concrete to abstract. Usually this means going from selection to supply-type items. Selection-type items would usually begin with the most limited selection type (true-false) and progress to multiple choice or matching in which options can be used more than once. The objective is to determine what the student knows. If more difficult items appear early in the test, the student may spend too much time on them and not get to the simpler ones that he/she can answer. For the test, we were planning in example 1d of this module, we would begin with true-false, followed in order by short answer, multiple choice, and the performance tasks
- Group items of the same type (true-false, multiple choice, etc.) together so that you only write directions for that item type once. Once you have a good set of directions for a particular type of item, save them so you can use them again the next time you use that same type of item.
- Check to see that directions for marking/scoring (point values, etc.) are included with each type of item.
- Provide directions for recording responses, and have students circle or underline correct responses when possible rather than writing them to avoid problems arising from poor handwriting.
- If a group of items of the same type (multiple choice, etc.) carry over from one page to another, repeat the directions at the top of the second page.
- All parts of an item should be on the same page.
- If graphs, tables, charts, or illustrations are used, put them near the questions based on them (on the same page, if at all possible).
- Check to see that items are independent (one item does not supply the answer or a clue to the answer of another question).
- Make sure the reading level is appropriate for your students. (This may be a problem with tests supplied by textbook publishers).
- Space the items for easy reading.
- Leave appropriate space for writing answers if completion/short answer, listing, or essay questions are used. (Younger children need larger spaces than older students because their print/handwriting is larger.)
- When possible, have answers recorded in a column down either the left or right side of the paper to facilitate scoring.

- Decide if students are to mark answers on the test, use a separate answer sheet, or use a blank sheet of paper. Usually separate answer sheets are not recommended for students in primary or early elementary grades.
- Include on the answer sheet (or on the test if students put answers on the test itself) a place for the student's name and the date.
- Make an answer key. (This is easy to do as you write the questions.)
- Check the answer key for a response pattern. If necessary, rearrange the order of questions within a question type so the correct answers appear to be in a random order.
- Set the test aside for awhile.
- Re-read the questions; proofread the test one last time before duplication. If possible, have someone else read the test as well.
- Prepare a copy of the test for each student (plus 2 or 3 extra copies). Questions written on the board may cause difficulties for students with visual problems. Reading the test questions to the students (except in the case of spelling tests) can be problematic for students with deficiencies in attention, hearing, comprehension, or short-term memory.
- Plan accommodations for individual students when appropriate.

## 8.12 Test Administration

A teacher's test administration procedures can have great impact on student test performance. As you will see in the guidelines below, test administration involves more than simply handling out and collecting the test.

**Before the test:**

- Avoid instilling anxiety
- Give as many of the necessary oral directions as possible before distributing the tests, but keep them to a minimum.
- Tell students purpose of the test.
- Give test-taking hints about guessing, skipping and coming back, etc.
- Tell students the amount of time allowed for the test. You may want to put the length of time remaining for the test on the board. This can be changed periodically to help students monitor their progress. If a clock is prominently available, an alternative would be to write the time at which they must be finished.
- Tell the students how to signal you if they have a question.
- Tell the students what to do with their papers when they are finished (how papers are to be collected).
- Tell the students what they are to do when they are finished, particularly if they are to go on to another activity (also write these directions on the chalkboard so

they can refer back to them).

- Rotate the method of distributing papers so you don't always start from the left or the front row.
- Make sure the room is well lighted and has a comfortable temperature.
- If a student is absent, write his/her name on a blank copy of the test as a reminder that it needs to be made up.

**After Distributing Test Papers**

- Remind students to put their names on their papers (and where to do so).
- If the test has more than one page, have each student check to see that all pages are there.

**During the Test**

- Minimize interruptions and distractions.
- Avoid giving hints.
- Monitor to check student progress and discourage cheating.
- Give time warnings if students are not pacing their work appropriately.
- Make a note of any questions students ask during the test so that items can be revised for future use.

**After the Test**

- Grade the papers (and add comments if you can); do test analysis (see the module on test analysis) after scoring and before returning papers to students if at all possible. If it is impossible to do your test analysis before returning the papers, be sure to do it at another time. It is important to both evaluation of your students and improvement of your tests.
- If you are recording grades, record them in pencil in your grade book before returning papers. If there are errors/adjustments in grading, they (grades) are easier to change when recorded in pencil.
- Return papers in a timely manner.
- Discuss test items with the students. If students have questions, agree to look over their papers again, as well as the papers of others who have the same question. It is usually better not to agree to make changes in grades on the spur of the moment while discussing the tests with the students but to give yourself time to consider what action you want to take. The test analysis may have already alerted you to a problem with a particular question that is common to several students, and you may already have made a decision regarding that question (to disregard the question and reduce the highest possible score accordingly, to give all students credit for that question, etc.).

## 8.13 Self Assessment Questions

1. The control group scored 47.26 on the pretest. Does this score represent nominal, ordinal, or interval scale data?

2. The control group's score of 47.26 on the pretest put it at the 26th percentile. Does this percentile score represent nominal, ordinal, or interval scale data?

3. The control group had a standard deviation of 7.78 on the pretest. Does this standard deviation represent nominal, ordinal, or interval scale data?

4. Construct a frequency distribution with suitable class interval size of marks obtained by 50 students of a class are given below:
   23, 50, 38, 42, 63, 75, 12, 33, 26, 39, 35, 47, 43, 52, 56, 59, 64, 77, 15, 21, 51, 54, 72, 68, 36, 65, 52, 60, 27, 34, 47, 48, 55, 58, 59, 62, 51, 48, 50, 41, 57, 65, 54, 43, 56, 44, 30, 46, 67, 53

5. The Lakers scored the following numbers of goals in their last twenty matches:
   3, 0, 1, 5, 4, 3, 2, 6, 4, 2, 3, 3, 0, 7, 1, 1, 2, 3, 4, 3

6. Which number had the highest frequency?
7. Which letter occurs the most frequently in the following sentence?

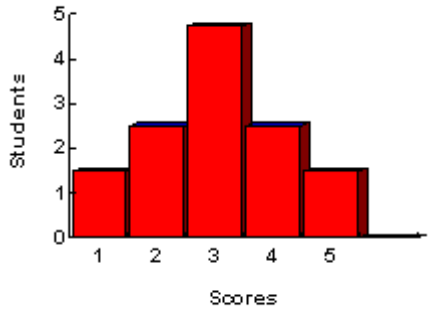
THE SUN ALWAYS SETS IN THE WEST.

8. Pi is a special number that is used to find the area of a circle. The following number gives the first 100 digits of the number pi:

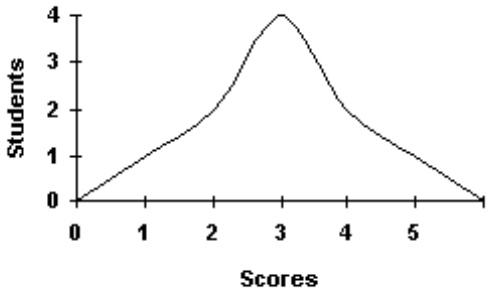
   141 592 653 589 793 238 462 643 383 279 502 884 197 169 399 375 105 820 974 944 592 307 816 406 286 208 998 628 034 825 342 117 067
   Which of the digits 0 to 9 occurs most frequently in this number?

9. Identify by correctly labeling the following graphic illustrations of results of a five point quiz taken by ten students.


Students
5
4
3
2
1
0
1 2 3 4 5
Scores


Students
5
4
3
2
1
0
1 2 3 4 5
Scores


Students
4
3
2
1
0
0 1 2 3 4 5
Scores


Students
4
3
2
1
0
0 1 2 3 4 5
Scores

1. In each data set given, find the mean of the group

a) Times were recorded when learners played a game

| Time in seconds | 36 - 45 | 46 - 55 | 56 - 65 | 66 - 75 | 76 - 85 | 86 - 95 | 96 - 105 |
|---|---|---|---|---|---|---|---|
| Frequency | 5 | 11 | 15 | 26 | 19 | 13 | 6 |

b) The following data were collected from a group of learners

| Time in seconds | 41 - 45 | 46 - 50 | 51 - 55 | 56 - 60 | 61 - 65 | 66 - 70 | 71 - 75 | 76 - 80 |
|---|---|---|---|---|---|---|---|---|
| Frequency | 3 | 5 | 8 | 12 | 14 | 9 | 7 | 2 |

11. Following are the wages of 8 workers of a factory. Find the range and the coefficient of range. Wages in (Rs) 14000, 14500, 15200, 13800, 14850, 14950, 15750, 14400.

12. The following distribution gives the numbers of houses and the number of persons per house.

| **Number of Persons** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Number of Houses** | 26 | 113 | 120 | 95 | 60 | 42 | 21 | 14 | 5 | 4 |

Calculate the range and coefficient of range.

## 8.14 References Suggested Readings


Huff, D. (1954). *How* to lie with statistics. New York: Norton.

Bertrand, A., & Cehula, J. P. (1980). *Tests, measurement, and evaluation: A developmental approach.* Reading, MA: Addison-Wesley. Chapter 7 provides an innovative presentation of most of the topics covered in the present chapter.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall. Chapters 7 through 12 are especially useful for helping teachers develop classroom achievement tests. Chapter 14 discusses observation and informal data collection techniques. Chapters 16 through 18 provide useful information on using standardized tests.

Hills, J. R. (1986). *All of Hills' handy hints*. Columbus, OH: Merrill. This is a collection of articles originally published in*Educational measurement: Issues and practice.* The articles offer practical and interesting insights into fallacies in the interpretation of test scores. (Incidentally, the original journal provides theoretically sound guidelines that are easy to understand.)

Kubiszyn, T., & Borich, G. (1987). *Educational tests and measurement: Classroom application and practice*. Glenview, IL: Scott, Foresman and Company. The chapter on data presentation provides useful and practical guidelines for communicating data effectively through graphs and diagrams.

Lyman, H. B. (1986). *Test scores and what they mean* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall. This book provides a detailed discussion of the interpretation of test scores.

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphic Press. This book offers interesting examples of how to display information and discusses strategies for presenting data graphically.

Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher, 21*, 14-23. This article presents strategies for employing and interpreting sophisticated yet understandable graphs to display quantitative data.

Worthen, B. R., Borg, W. R., & White, K. R. (1993). *Measurement and evaluation in the schools.* New York: Longman. Chapter 5 presents a practical discussion of the meaning of test scores.

Gellman, E. (1995). School *testing: What parents and educators need to know*. Westport, CT:

Praeger. Hamill, D. (1987) *Assessing the abilities and instructional needs of students.* Austin, TX: Pro-Ed. Salvia, J. & Ysseldyke, J. (1992)*Assessment in special and remedial education,* 5th edition. Boston:

Houghton-Mifflin.

**UNIT 9**

# REPORTING TEST SCORES

***Written By:***
***Dr. Muhammad Saeed***

***Reviewed By:***
***Dr. Naveed Sultana***

## CONTENTS

***Sr. No*** ***Topic*** ***Page No***

# INTRODUCTION

The unit "Reporting Test Scores" is about measuring the performance of students by providing a profile of their progress and reporting the scores of tests in different ways in context to the different purposes. There is a long tradition that students' skills are measured by some of testing procedures. Invariably, the product of testing is a score, a 'yardstick' by which an individual student is compared with others and/or by which progress is documented. Teachers and other educators use tests, and subsequently test scores in a variety of ways.

The first major topic of the unit deals with the functions of test scores and progress reports of students after taking any test. As there are different functions of grading and reporting systems with respect to its uses like instructional uses, providing feedback to students for administrative use and guidance and informing parents about their children's performance.

The second key topic in the unit discussed is the "Types of Test Scores and Progress Reports". Here two types of reporting test scores are discussed. First is Norm-referenced tests which include raw scores, grade norms, percentiles, stanines, and standard scores. Second is Criterion-referenced test which include system of pass-fail and the other types of the practices that are used to report the progress of students.

The third major theme is "Calculating CGPA and Assigning Letter Grades" It includes the method of calculating CGPA and different steps which are concerned with assigning letter grades in reporting test scores such as combining the data, selecting the proper frame of reference for grading and determining the distribution of grades etc.

The last major theme of the unit is "Conducting Parent-Teacher Conferences". This section includes the information and important preparations for conducting the parent teacher conferences, mentioning the "Do's" and "Don'ts" of the parent teacher conferences.

# OBJECTIVES

After studying the Unit, the students will be able to:

1. understand the purpose of reporting test scores
2. explain the functions of test scores
3. describe the essential features of progress report
4. enlist the different types of grading and reporting systems
5. calculate CGPA
6. conduct parent teacher conferences

## 9.1 Functions of Test Scores and Progress Reports

The task of grading and reporting students' progress cannot be separated from the procedures adopted in assessing students' learning. If instructional objectives are well defined in terms of behavioural or performance terms and relevant tests and other assessment procedures are properly used, grading and reporting become a matter of summarizing the results and presenting them in understandable form. Reporting students' progress is difficult especially when data is represented in single letter-grade system or numerical value (Linn & Gronlund, 2000).

Assigning grades and making referrals are decisions that require information about individual students. In contrast, curricular and instructional decisions require information about groups of students, quite often about entire classrooms or schools (Linn & Gronlund, 2000).

There are three primary purposes of grading students. First, grades are the primary currency for exchange of many of the opportunities and rewards our society has to offer. Grades can be exchanged for such diverse entities as adult approval, public recognition, college and university admission etc. To deprive students of grades means to deprive them of rewards and opportunities. Second, teachers become habitual of assessing their students' learning in grades, and if teachers don't award grades, the students might not well know about their learning progress. Third, grading students motivate them. Grades can serve as incentives, and for many students incentives serve a motivating function.

The different functions of grading and reporting systems are given as under:

### *1. Instructional uses*

The focus of grading and reporting should be the student improvement in learning. This is most likely occur when the report: a) clarifies the instructional objectives; b) indicates the student's strengths and weaknesses in learning; c) provides information concerning the student's personal and social development; and d) contributes to student's motivation.

The improvement of student learning is probably best achieved by the day-to-day assessments of learning and the feedback from tests and other assessment procedures. A portfolio of work developed during the academic year can be displayed to indicate student's strengths and weaknesses periodically.

Periodic progress reports can contribute to student motivation by providing short-term goals and knowledge of results. Both are essential features of essential learning. Well-designed progress reports can also help in evaluating instructional procedures by identifying areas need revision. When the reports of majority of students indicate poor progress, it may infer that there is a need to modify the instructional objectives.

### *2. Feedback to students*

Grading and reporting test results to the students have been an on-going practice in all the educational institutions of the world. The mechanism or strategy may differ from country to country or institution to institution but each institution observes this practice in any way. Reporting test scores to students has a number of advantages for them. As the students move up through the grades, the usefulness of the test scores for personal academic planning and self-assessment increases. For most students, the scores provide feedback about how much they know and how effective their efforts to learn have been. They can know their strengths and areas need for special attention. Such feedback is essential if students are expected to be partners in managing their own instructional time and effort. These results help them to make good decisions for their future professional development.

Teachers use a variety of strategies to help students become independent learners who are able to take an increasing responsibility for their own school progress. Self-assessment is a significant aspect of self-guided learning, and the reporting of test results can be an integral part of the procedures teachers use to promote self-assessment. Test results help students to identify areas need for improvement, areas in which progress has been strong, and areas in which continued strong effort will help maintain high levels of achievement. Test results can be used with information from teacher's assessments to help students set their own instructional goals, decide how they will allocate their time, and determine priorities for improving skills such as reading, writing, speaking, and problem solving. When students are given their own test results, they can learn about self-assessment while doing actual self-assessment. (Iowa Testing Programs, 2011).

Grading and reporting results also provide students an opportunity for developing an awareness of how they are growing in various skill areas. Self-assessment begins with self-monitoring, a skill most children have begun developing well before coming to kindergarten.

### *3. Administrative and guidance uses*

Grades and progress reports serve a number of administrative functions. For example, they are used for determining promotion and graduation, awarding honours, determining sports eligibility of students, and reporting to other institutions and employers. For most administrative purposes, a single letter-grade is typically required, but of course, technically single letter-grade does not truly interpret student's assessment.

Guidance and Counseling officers use grades and reports on student's achievement, along with other information, to help students make realistic educational and vocational plans. Reports that include ratings on personal and social characteristics are also useful in helping students with adjustment problems.

### *4. Informing parents about their children's performance*

Parents are often overwhelmed by the grades and test reports they receive from school

personnel. In order to establish a true partnership between parents and teachers, it is essential that information about student progress be communicated clearly, respectfully and accurately. Test results should be provided to parents using; a) simple, clear language free from educational and test jargon, and b) explanation of the purpose of the tests used (Canter, 1998).

Most of the time parents are either ignored or least involved to let them aware of the progress of their children. To strengthen connection between home and school parents need to receive comprehensive information about their children achievement. If parents do not understand the tests given to their children, the scores, and how the results are used to make decisions about their children, they are prohibited from helping their children learn and making decisions.

According to Kearney (1983), the lack of information provided to consumers about test data has sweeping and negative consequences. He states;

Individual student needs are not met, parents are not kept fully informed of student progress, curricular needs are not discovered and corrected, and the results are not reported to various audiences that need to receive this information and need to know what is being done with the information.

In some countries, there are prescribed policies for grading and reporting test results to the parents. For example, Michigan Educational Assessment Policy (MEAP) is revised periodically in view of parents' suggestions and feedback. MEAP consists of criterion-referenced tests, primarily in mathematics and reading, that are administered each year to all fourth, seventh and tenth graders. MEAP recommends that policy makers at state and local levels must develop strong linkages to create, implement and monitor effective reporting practices. (Barber, Paris, Evans, & Gadsden, 1992).

Without any doubt, it is more effective to talk parents to face about their children's scores than to send a score report home for them to interpret on their own. For a variety of reasons, a parent-teacher or parent-student-teacher conference offers an excellent occasion for teachers to provide and interpret those results to the parents.

1. Teachers tend to be more knowledgeable than parents about tests and the types of scores being interpreted.
2. Teachers can make numerous observations of their student's work and consequently substantiate the results. In-consistencies between test scores and classroom performance can be noted and discussed.
3. Teachers possess work samples that can be used to illustrate the type of classroom work the student has done. Portfolios can be used to illustrate strengths and to explain where improvements are needed.
4. Teachers may be aware of special circumstances that may have influenced the scores, either positively or negatively, to misrepresent the students' achievement level.

5. Parents have a chance to ask questions about points of misunderstanding or about how they can work. The student and the teacher in addressing apparent weaknesses and in capitalizing on strengths wherever possible, test scores should be given to the parents at the school. (Iowa Testing Program, 2011).

Under the Act of 1998, schools are required to regularly evaluate students and periodically report to parents on the results of the evaluation, but in specific terms, the NCCA guidelines make a recommendation that schools should report twice annually to parents – one towards the end of 1st term or beginning of 2nd term, and the other towards the end of school year.

Under existing data protection legislation, parents have a statutory right to obtain scores which their children have obtained in standardized tests. NCCA have developed a set of reports card templates to be used by schools in communicating with parents and taken in conjunction with the Circular 0138 which was issued by the Department of Education in 2006.

In a case study conducted in the US context (www.uscharterschools.org) it was found that 'the school should be a source for parents, it should not dictate to parents what their role should be'. In other words, the school should respect all parents and appreciate the experiences and individual strengths they offer their children.

## 9.2 Types of Test Reporting and Marking

Usually two types of tests are used in schools, criterion-referenced and norm-referenced. Criterion-referenced tests are used to measure student mastery of instructional objectives or curriculum rather than to compare one student's performance with another or to rank students. They are often used as benchmarks to identify areas of strengths and/or weaknesses in a given curriculum. Norm-referenced tests compare an individual's performance to that of his/her classmates, thus emphasizing relative rather an absolute performance. Scores on norm-referenced tests indicate the students' ranking relative position to that group. Typical scores used with norm-referenced tests include raw scores, grade norms, percentiles, stanines, and standard scores.

### *1. Raw scores*

The raw score is simply the number of points received on a test when the test has been scored according to the directions. For example, if a student responds to 65 items correctly on an objective test in which each correct item counts one point, the raw score will be 65.

Although a raw score is a numerical summary of student's test performance, it is not very meaningful without further information. For example, in the above example, what does a raw score of 35 mean? How many items were in the test? What kinds of the problems were asked? How the items were difficult?

### *2. Grade norms*

Grade norms are widely used with standardized achievement tests, especially at elementary level. The grade equivalent that corresponds to a particular raw score identifies the grade level at which the typical student obtains that raw score. Grade equivalents are based on the performance of students in the norm group in each of two or more grades.

### *3. Percentile ranking*

A percentile is a score that indicates the rank of the score compared to others (same grade/age) using a hypothetical group of 100 students. In other words, a percentile rank (or percentile score) indicates a student's relative position in the group in terms of percentage of students.

Percentile rank is interpreted as the percentage of individuals receiving scores equal or lower than a given score. A percentile of 25 indicates that the student's test performance is equal or exceeds 25 out of 100 students on the same measure.

### *4. Standard scores*

A standard score is also derived from the raw scores using the normal information gathered when the test was developed. Instead of indicating a student's rank compared to others, standard scores indicate how far above or below the average (Mean) an individual score falls, using a common scale, such as one with an average of 100. Basically standard scores express test performance in terms of standard deviation (SD) from the Mean. Standard scores can be used to compare individuals of different grades or age groups because all are converted into the same numerical scale. There are various forms of standard scores such as z-score, T-score, and stanines.

Z-score expresses test performance simply and directly as the number of SD units a raw score is above or below the Mean. A z-score is always negative when the raw score is smaller than Mean. Symbolic representation can be shown as: z-score = X-M/SD.

T-score refers to any set of normally distributed standard cores that has a Mean of 50 and SD of 10. Symbolically it can be represented as: T-score = 50+10(z).

Stanines are the simplest form of normalized standard scores that illustrate the process of normalization. Stanines are single digit scores ranging from 1 to 9. These are groups of percentile ranks with the entire group of scores divided into nine parts, with the largest number of individuals falling in the middle stanines, and fewer students falling at the extremes (Linn & Gronlund, 2000).

### *5. Norm reference test and traditional letter-grade system*

It is the most easiest and popular way of grading and reporting system. The traditional system is generally based on grades A to F. This rating is generally reflected as: Grade A

(Excellent), B (Very Good), C (Good), D (Satisfactory/Average), E (Unsatisfactory/ Below Average), and F (Fail).

This system does truly assess a student's progress in different learning domains. First shortcoming is that using this system it is difficult to interpret the results. Second, a student's performance is linked with achievement, effort, work habits, and good behaviour; traditional letter-grade system is unable to assess all these domains of a student. Third, the proportion of students assigned each letter grade generally varies from teacher to teacher. Fourth, it does not indicate patterns of strengths and weaknesses in the students (Linn & Gronlund, 2000). Inspite of these shortcomings, this system is popular in schools, colleges and universities.

### *6. Criterion reference test and the system of pass-fail*

It is a popular way of reporting students' progress, particularly at elementary level. In the context of Pakistan, as majority of the parents are illiterate or hardly literate, therefore they have concern with 'pass or fail' about their children's performance in schools. This system is mostly used for courses taught under a pure mastery learning approach i.e. criterion-referenced testing.

This system has also many shortcomings. First, as students are declared just pass or fail (successful or unsuccessful) so many students do not work hard and hence their actual learning remains unsatisfactory or below desired level. Second, this two-category system provides less information to the teacher, student and parents than the traditional letter-grade (A, B, C, D) system. Third, it provides no indication of the level of learning.

### 7. *Checklist of Objectives*

To provide more informative progress reports, some schools have replaced or supplemented the traditional grading system with a list of objectives to be checked or rated. This system is more popular at elementary school level. The major advantage of this system is that it provides a detailed analysis of the students' strengths and weaknesses. For example, the objectives for assessing reading comprehension can have the following objectives.

- Reads with understanding
- Works out meaning and use of new words
- Reads well to others
- Reads independently for pleasure (Linn & Gronlund, 2000).

### *8. Rating scales*

In many schools students' progress is prepared on some rating scale, usually 1 to 10, instead letter grades; 1 indicates the poorest performance while 10 indicates as the

excellent or extra-ordinary performance. But in the true sense, each rating level corresponds to a specific level of learning achievement. Such rating scales are also used by the evaluation of students for admissions into different programmes at university level. Some other rating scales can also be seen across the world.

In rating scales, we generally assess students' abilities in the context of 'how much', 'how often', 'how good' etc. (Anderson, 2003). The continuum may be qualitative such as 'how good a student behaves' or it may quantitative such as 'how much marks a student got in a test'. Developing rating scales has become a common practice now-a-days, but still many teachers don't possess the skill of developing an appropriate rating scale in context to their particular learning situations.

### *9. Letters to parents/guardians*

Some schools keep parents inform about the progress of their children by writing letters. Writing letters to parents is usually done by a fewer teachers who have more concern with their students as it is a time consuming activity. But at the same time some good teachers avoid to write formal letters as they think that many aspects are not clearly interpreted. And some of the parents also don't feel comfortable to accept such letters.

Linn and Gronlund (2000) state that although letters to parents might provide a good supplement to other types of reports, their usefulness as the sole method of reporting progress is limited by several of the following factors.

- Comprehensive and thoughtful written reports require excessive amount of time and energy.
- Descriptions of students learning may be misinterpreted by the parents.
- Fail to provide a systematic and organized information

### *10. Portfolio*

The teachers of some good schools prepare complete portfolio of their students. Portfolio is actually cumulative record of a student which reflects his/her strengths and weaknesses in different subjects over the period of the time. It indicates what strategies were used by the teacher to overcome the learning difficulties of the students. It also shows students' progress periodically which indicates his/her trend of improvement. Developing portfolio is really a hard task for the teacher, as he/she has to keep all record of students such as teacher's lesson plans, tests, students' best pieces of works, and their assessments records in an academic year.

An effective portfolio is more than simply a file into which student work products are placed. It is a purposefully selected collection of work that often contains commentary on the entries by both students and teachers.

No doubt, portfolio is a good tool for student's assessment, but it has three limitations. First, it is a time consuming process. Second, teacher must possess the skill of developing

portfolio which is most of the time lacking. Third, it is ideal for small class size and in Pakistani context, particularly at elementary level, class size is usually large and hence the teacher cannot maintain portfolio of a large class.

### *11. Report Cards*

There is a practice of report cards in many good educational institutions in many countries including Pakistan. Many parents desire to see the report cards or progress reports in written form issued by the schools. Although a good report card explains the achievement of students in terms of scores or marks, conduct and behaviour, participation in class activities etc. Well written comments can offer parents and students' suggestions as to how to make improvements in specific academic or behavioural areas. These provide teachers opportunities to be reflective about the academic and behavioural progress of their students. Such reflections may result in teachers gaining a deeper understanding of each student's strengths and needs for improvement. Bruadli (1998) has divided words and phrases into three categories about what to include and exclude from written comments on report cards.

A. Words and phrases that promote positive view of the student

1. Gets along well with people
2. Has a good grasp of …
3. Has improved tremendously
4. Is a real joy to have in class
5. Is well respected by his classmates
6. Works very hard

B. Words and phrases to convey the students need help

1. Could benefit from …
2. Finds it difficult at time to …
3. Has trouble with …
4. Requires help with …
5. Needs reinforcement in …

C. Words and phrases to avoid or use with extreme caution

1. Always
2. Never
3. Can't )or unable to)

4. Won't

Report card usually carries two shortcomings: a) regardless of how grades are assigned, students and parents tend to use them normatively; and b) many students and parents (and some teachers) believe that grades are far more precise than they are. In most grading schemes, an 'F' denotes to fail or unsatisfactory. Hall (1990) and Wiggins (1994) state that not only grades imprecise, they are vague in their meaning. They do not provide parents or students with a thorough understanding of what has been learned or accomplished.

### *12. Parent-teacher conferences*

Parent-teacher conferences are mostly used in elementary schools. In such conferences portfolio are discussed. This is a two-way flow of information and provides much information to the parents. But one of the limitations is that many parents don't come to attend the conferences. It is also a time consuming activity and also needs sufficient funds to hold conferences.

Literature also highlights 'parent-student-teacher conference' instead 'parent-teacher conference', as student is also one of the key components of this process since he/she is directly benefitted. In many developed countries, it has become the most important way of informing parents about their children's work in school. Parent-teacher conferences are productive when these are carefully planned and the teachers are skilled and committed.

The parent-teacher conference is an extremely useful tool, but it shares three important limitations with informal letter. First, it requires a substantial amount of time and skills. Second, it does not provide a systematic record of student's progress. Third, some parents are unwilling to attend conferences, and they can't be enforced.

Parent-student-teacher conferences are frequently convened in many states of the USA and some other advanced countries. In the US, this has become a striking feature of Charter Schools. Some schools rely more on parent conferences than written reports for conveying the richness of how students are doing or performing. In such cases, a school sometimes provides a narrative account of student's accomplishments and status to augment the parent conferences. (www.uscharterschools.org).

### *13. Other ways of reporting students results to parents*

There are also many other ways to enhance communication between teacher and parent, e.g. phone calls. The teachers should contact telephonically to the parents of the children to let them inform about child's curriculum, learning progress, any special achievement, sharing anecdote, and invite parents in open meetings, conferences, and school functions.

## 9.3 Calculating CGPA and Assigning Letter Grades

CGPA stands for Cumulative Grade Point Average. It reflects the grade point average of all subjects/courses regarding a student's performance in composite way. To calculate CGPA, we should have following information.

- Marks in each subject/course
- Grade point average in each subject/course
- Total credit hours (by adding credit hours of each subject/course)

Calculating CGPA is very simple that total grade point average is divided by total credit hours. For example if a student MA Education programme has studied 12 courses, each of 3 credits. The total credit hours will be 36. The average of GPA, in all the twelve course will be the CGPA. In the following table the GPA calculated for astudent of MA Education program is given as example.

| **Sr. #** | **Course Title** | **Credits** | **Marks** | **Grade** | **GPA** | **CGPA** |
|---|---|---|---|---|---|---|
| 1. | Philosophy of Education | 3 | 85 | A | 4.0 | |
| 2. | Curriculum and Instruction | 3 | 78 | B+ | 3.3 | |
| 3. | Edul. Admin.& Supervision | 3 | 72 | B | 3.0 | |
| 4. | Computer in Education | 3 | 77 | B+ | 3.3 | |
| 5. | Educational Technology | 3 | 77 | B+ | 3.3 | |
| 6. | Instructional Technology | 3 | 71 | B | 3.0 | |
| 7. | Teacher Edu. in Islamic Pers. | 3 | 79 | B+ | 3.3 | |
| 8. | History of TE in Pakistan | 3 | 76 | B+ | 3.3 | |
| 9. | Master Research Project | 3 | 81 | A- | 3.7 | |
| 10. | Islamic System of Education | 3 | 85 | A | 4.0 | |
| 11. | Research Methods in Edu. | 3 | 86 | A | 4.0 | |
| 12. | Edul. Assessment & Evalu. | 3 | 75 | B+ | 3.3 | |
| 13. | Comparative Education | 3 | 82 | A- | 3.7 | |
| 14. | Methods of Teaching Islamiat | 3 | 85 | A | 4.0 | |
| 15. | Teaching of Urdu | 3 | 80 | A- | 3.7 | |
| 16. | Islamic Ideology & Ideology | 3 | 81 | A- | 3.7 | |
| 17. | Student Teaching & Obs. I | 3 | 80 | A- | 3.7 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 18. | Student Teaching & Obs. II | 3 | 88 | A | 4.0 | |
| 19. | Education in Pakistan | 3 | 88 | A | 4.0 | |
| 20. | Teaching of Social Studies | 3 | 81 | A- | 3.7 | |
| **21.** | **Total** | **60** | | | | |

The average of GPA, will represent (GPA) $\text{CGPA} = \frac{\text{sum of GPA}}{\text{total course}}$

**Assigning letter grades**

Letter grade system is most popular in the world including Pakistan. Most teachers face problems while assigning grades. There are four core problems or issues in this regard; 1) what should be included in a letter grade, 2) how should achievement data be combined in assigning letter grades?, 3) what frame of reference should be used in grading, and 4) how should the distribution of letter grades be determined?

### *1. Determining what to include in a grade*

Letter grades are likely to be most meaningful and useful when they represent achievement only. If they are communicated with other factors or aspects such as effort of work completed, personal conduct, and so on, their interpretation will become hopelessly confused. For example, a letter grade C may represent average achievement with extraordinary effort and excellent conduct and behaviour or vice versa.

If letter grades are to be valid indicators of achievement, they must be based on valid measures of achievement. This involves defining objectives as intended learning outcomes and developing or selecting tests and assessments which can measure these learning outcomes.

### *2. Combining data in assigning grades*

One of the key concerns while assigning grades is to be clear what aspects of a student are to be assessed or what will be the tentative weightage to each learning outcome. For example, if we decide that 35 percent weightage is to be given to mid-term assessment, 40 percent final term test or assessment, and 25% to assignments, presentations, classroom participation and conduct and behaviour; we have to combine all elements by assigning appropriate weights to each element, and then use these composite scores as a basis for grading.

### *3. Selecting the proper frame of reference for grading*

Letter grades are typically assigned on the basis of one of the following frames of reference.

a) Performance in relation to other group members (relative grading)

b) Performance in relation to specified standards (absolute grading)

c) Performance in relation to learning ability (amount of improvement)

Assigning grades on relative basis involves comparing a student's performance with that of a reference group, mostly class fellows. In this system, the grade is determined by the student's relative position or ranking in the total group. Although relative grading has a disadvantage of a shifting frame of reference (i.e. grades depend upon the group's ability), it is still widely used in schools, as most of the time our system of testing is 'norm-referenced'.

Assigning grades on an absolute basis involves comparing a student's performance to specified standards set by the teacher. This is what we call as 'criterion-referenced' testing. If all students show a low level of mastery consistent with the established performance standard, all will receive low grades.

The student performance in relation to the learning ability is inconsistent with a standard-based system of evaluating and reporting student performance. The improvement over the short time span is difficult. Thus lack of reliability in judging achievement in relation to ability and in judging degree of improvement will result in grades of low dependability. Therefore such grades are used as supplementary to other grading systems.

### *4. Determining the distribution of grades*

The assigning of relative grades is essentially a matter of ranking the student in order of overall achievement and assigning letter grades on the basis of each student's rank in the group. This ranking might be limited to a single classroom group or might be based on the combined distribution of several classroom groups taking the same course.

If grading on the curve is to be done, the most sensible approach in determining the distribution of letter grades in a school is to have the school staff set general guidelines for introductory and advanced courses. All staff members must understand the basis for assigning grades, and this basis must be clearly communicated to users of the grades. If the objectives of a course are clearly mentioned and the standards for mastery appropriately set, the letter grades in an absolute system may be defined as the degree to which the objectives have been attained, as followed.

A = Outstanding (90 to 100%)

B = very Good (80-89%)

C = Satisfactory (70-79%)

D = Very Weak (60-69%)

F = Unsatisfactory (Less than 60%)

## 9.4 Conducting Parent-Teacher Conferences

The first conference is usually arranged in the beginning of the school year to allow parents and teachers to get acquaintance and preparing plan for the coming months. Teachers usually receive some training to plan and conduct such conferences. Following steps may be observed for holding effective parent-teacher conferences.

***1. Prepare for the conference***

- Review the goals and objectives
- Organize the information to present
- If portfolios are to discuss, these are well-arranged
- Start and keep positive focus
- Announce the final date and time as per convenience of the parents and children
- Consider socio-cultural barriers of students / parents
- Check with other staff who works your advisee
- Develop a packet of conference including student's goals, samples of work, and reports or notes from other staff.

***2. Rehearse the conference with students by role-playing***

- Students present their goals, learning activities, samples of work
- Students ask for comments and suggestions from parents

***3. Conduct conference with student, parent, and advisor. Advisee takes the lead to the greatest possible extent***

- Have a comfortable setting of chairs, tables etc.
- Notify a viable timetable for the conferences
- Review goals set earlier
- Review progress towards goals
- Review progress with samples of work from learning activities
- Present students strong points first
- Review attendance and handling of responsibilities at school and home
- Modify goals for balance of the year as necessary
- Determine other learning activities to accomplish goals
- Describe upcoming events and activities

- Discuss how the home can contribute to learning
- Parents should be encouraged to share their thoughts on students' progress
- Ask parents and students for questions, new ideas

***4. Do's of parent-teacher conferences***

- Be friendly
- Be honest
- Be positive in approach
- Be willing to listen and explain
- Be willing to accept parents' feelings
- Be careful about giving advice
- Be professional and maintain a positive attitude
- Begin with student's strengths
- Review student's cumulative record prior to conference
- Assemble samples of student's work
- List questions to ask parents and anticipate parents' questions
- Conclude the conference with an overall summary
- Keep a written record of the conference, listing problems and suggestions, with a copy for the parents

***5. Don'ts of the parent teacher conference***

- Don't argue
- Don't get angry
- Don't ask embarrassing questions
- Don't talk about other students, parents and teachers
- Don't bluff if you don't know
- Don't reject parents' suggestions
- Don't blame parents
- Don't talk too much; be a good listener (www.udel.edu.)

## 9.5 Activities

**Activity 1:**

Enlist three pros and cons of test scores.

**Activity 2:**

Give a self-explanatory example of each of the types of test scores.

**Activity 3:**

Write down the different purposes and functions of test scores in order of importance as per your experience. Add more purposes as many as you can.

**Activity 4:**

Compare the modes of reporting test scores to parents by MEAP and NCCA. Also conclude which is relatively more appropriate in the context of Pakistan as per your point of view.

**Activity 5:**

In view of the strengths and shortcomings in above different grading and reporting systems, how would you briefly comment on the following characteristics of a multiple grading and reporting system for effective assessment of students' learning?

a) Grading and reporting system should be guided by the functions to be served.

b) It should be developed cooperatively by parents, students, teachers, and other school personnel.

c) It should be based on clear and specific instructional objectives.

d) It should be consistent with school standards.

e) It should be based on adequate assessment.

f) It should provide detailed information of student's progress, particularly diagnostic and practical aspects.

g) It should have the space of conducting parent-teacher conferences.

**Activity 6:**

Explain the differences between relative grading and absolute grading by giving an example of each.

**Activity 7:**

Faiza Shaheen, a student of MA Education (Secondary) has earned the following marks, grades and GPA in the 22 courses at the Institute of Education & Research, University of the Punjab. Calculate her CGPA. Note down that that maximum value of GPA in each course is 4.

**Activity 8:**

Write Do's and Don'ts in order of priority as per your perception. You may add more points or exclude what have been mentioned above.

## 9.6 Self-Assessment Questions

**Part-I: MCQs**:

Encircle the best/correct response against each of the following statements.

1. Comparing a students' performance in a test in relation to his/her classmates is referred to as:
   a) Learning outcomes
   b) Evaluation
   c) Measurement
   d) Norm-referenced assessment
   e) Criterion-referenced assessment

2. The first test data on a test is called as:
   a) Frequency
   b) Numeric
   c) Raw score
   d) True score
   e) Cleaned data

3. A student's relative position in the group in terms of percentage is referred to as:
   a) Mean
   b) Median
   c) Mode
   d) Standard deviation
   e) Percentile

4. A z-score is always negative when:
   a) Raw score is smaller than mean
   b) Raw score is greater than mean
   c) Raw score is equal to mean
   d) T-score = 50
   e) None of the above

5. The simplest form of normalized standard score is:
   a) Standard score
   b) Z-score
   c) True score
   d) Stanines
   e) T-score

6. Grading and reporting works better when:
   a) Assessment procedures rarely used
   b) Assessment procedures mostly used
   c) Assessment procedures properly used
   d) Students perform better
   e) Awards are given to students

7. Periodic assessment is almost synonymous to:
   a) Evaluation
   b) Measurement
   c) Summative assessment
   d) Formative assessment
   e) Monthly assessment

8. A student's best work is generally compiled by a teacher in the form of:
   a) Cumulative record
   b) Portfolio
   c) Assessment report
   d) Comments by the teacher
   e) Written comments

9. Self-assessment begins with:
   a) Excellent work

b) Any academic contribution

c) Self-monitoring

d) Skill development

e) Knowledge updating

10. Who said that 'lack of information provided to consumers about test data has negative and sweeping consequences'

a) Hopkins & Stanley

b) Anderson

c) Linn & Gronlund

d) Barber et al.

e) Kearney

11. Michigan Educational Policy (MEAP) is revised by parents' suggestions:

a) Quarterly

b) Biannually

c) Annually

d) Every three years

e) Periodically

12. The system used in our BISEs is based on:

a) Letter grade

b) Pass-fail

c) Checklist of objectives

d) Rating scales

e) Portfolio

13. The contribution to report cards is of:

a) Hopkins & Stanley

b) Hall

c) Wiggins

d) Anderson

e) Bruadli

14. The first stage in parent-teacher conferences is:

a) Start and keep positive focus

b) Planning

c) Implementing/conducting

d) Rehearsal

e) Role play

**Part-II: Short Answer Questions**

1. How do Z-scores and T-scores differ?
2. Write down two strengths and two shortcomings of test scores.
3. Explain briefly the function of 'instructional uses' of grading and reporting.
4. What type of grading system is employed in public sector elementary schools of Pakistan?
5. How does 'pass-fail' system not truly assess students' performance?
6. What do you mean by 'checklist of objectives' in context to a type of grading and reporting?
7. Enlist the activities that a teacher can consider in developing a portfolio of a student.
8. Report cards are a good means of reporting results to parents. Comment.
9. What is the importance of assigning letter grades to assess students' assessment?

**Part-III: Essay-type Questions**

1. Describe the various types of reporting test scores by giving examples from our country context.
2. In what way parent-teacher conferences play significant role in regard to provide feedback to parents about their children academic growth and development?
3. What should be essentials of a good progress report? Discuss in detail with respect to public school system in Pakistan.

**Key to MCQs**

| Q. No. | Correct response | Q. No. | Correct Response |
|---|---|---|---|
| 1. | D | 2. | C |
| 3. | E | 4. | A |
| 5. | D | 6. | C |
| 7. | D | 8. | B |
| 9. | C | 10. | E |
| 11. | D | 12. | B |
| 13. | E | 14. | B |

## 9.7 References/Suggested Readings


Anderson, L.W. (2003). *Classroom Assessment – Enhancing the Quality of Teacher Decision Making*. London: Lawerence Erlbaum Associates, Publishers.

Barber, B.L., Paris, S.G., Evans, M., & Gadsden, V.L. (1992). *Policies for Reporting test Results to Parents*. USA: Pennsylvania State University.

Brualdi, A. (1998). Teacher comments on report cards. *Practical Assessment, Research & Evaluation, 6*(5).

Canter, A. (1998). *Understanding test scores*. Accessible at: http://www.wyanclotle.org/SpecialEd/Understanding_test_scores.htm

Hall, K. (1990). *Determining the Success of Narrative Report Cards*. Unpublished Manuscript. (ERIC Documents No. 334 013).

Hopkins, K.D. & Stanley, J.C. (1981). *Educational and Psychological Measurement and Evaluation* (6th ed.). New Dehli: Pearson Education.

Iowa Testing Programs (2011). *Reporting results – Interpreting test scores – ITBS: Iowa Tests of Basic Skills.* Iowa: The University of Iowa College of Education. Accessible at: http://www.education.uiowa.edu/itp/itbs_interp_rpts.aspx

Kearney, C.P. (1983). Uses and Abuses of Assessment and Evaluation data by Policy Makers. *Educational Measurement: Issues and Practices, 2*, 9-17.

Linn, R.L. & Gronlund, N.E. (2000). *MEASUREMENT and ASSESSMENT in TEACHING* (8th ed.). New Dehli: Pearson Education.

Wiggins, G. (1994). Towards better report cards. *Educational Leadership, 52*(2), 28-37.

www.uscharterschools.org

www.udel.edu.