STUDY GUIDE

Measurement and Evaluation in Education

Level: MS/M.Phil
Course Code: 3742
Units: 1–9
Credit Hours: 3

Written by
Prof. Dr. Rehana Masrur

DEPARTMENT OF SECONDARY TEACHER EDUCATION
FACULTY OF EDUCATION
ALLAMA IQBAL OPEN UNIVERSITY, ISLAMABAD
### COURSE TEAM

<table>
<thead>
<tr>
<th>Role</th>
<th>Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dean</td>
<td>Prof. Dr. Rehana Masrur</td>
</tr>
<tr>
<td>Chairperson</td>
<td>Prof. Dr. Rehana Masrur</td>
</tr>
<tr>
<td>Course Development Coordinators</td>
<td>Prof. Dr. Rehana Masrur, Dr. Muhammad Arif Zia</td>
</tr>
<tr>
<td>Course Team Members</td>
<td>Dr. Zulkaif Ahmad, Dr. Mussarat Anwar, Dr. Arif Zia, Dr. Rehana Masrur</td>
</tr>
<tr>
<td>Editor</td>
<td>Abdul Wadood</td>
</tr>
<tr>
<td>Course Coordinator &amp; Revision</td>
<td>Prof. Dr. Rehana Masrur</td>
</tr>
</tbody>
</table>
FORWARD

‘Measurement’ and ‘Evaluation’ are very important aspects for the appraisal of any phenomenon. Evaluation helps in ascertaining the level of achievement of objectives. Similarly, measurement helps in rating the worth of any task rendered. As we know that quality of teacher preparation and ongoing professional development indeed is crucial to quality of teacher performance in the classroom and this can be achieved through the awareness and application of the principles relating to measurement and evaluation.

The rapid technological changes and developments have changed the face of the social as well as educational arena. These changes demand more rigorous approach from the academics which consequently have changed their approach towards measurement and evaluation.

The particular fact leads the teacher educators towards more specific approach that the trainee teachers should be well prepared. The particular study guides “Measurement and Evaluation in Education-I” and “Measurement and Evaluation in Education-II” are an attempt of the course development team, who focused upon the modern concepts of the measurement and evaluation.

I appreciate Dean Faculty of Education (coordinator of this course) and course development team, on successful completion of this study guide and selection of web based study materials.

The theoretical and methodological issues in measurement and evaluation addressed in these study guides will help future educators to measure students’ performance on scientific basis.

I congratulate the co-coordinator for her efforts regarding the completion of study guides.

(Prof. Dr. Mehmood-ul-Hassan Butt)
Vice-Chancellor
INTRODUCTION TO THE COURSE

Measurement and Evaluation in Education-I is a study guide designed for the students of MS/M.Phil. Program. The purpose of this study guide is to introduce students to the basic and advanced concepts in educational measurement and assessment. This study guide gives a brief but comprehensive description of concepts related to measurement, assessment and evaluation. Students are advised to study additional and suggested books web material and use given internet sources to get more detailed information. Moreover, practical activities are included in order to encourage students to apply knowledge in practical situations. Allied material in printed form and web material on a CD is also part of this course.

This study guide is written keeping in view of the following general objectives:
1. To develop understanding of the role and importance of measurement and assessment in education that has been discussed in unit 1.
2. To develop ability of stating instructional aims and objectives clearly and effectively that is the theme of unit 2.
3. To develop ability of test construction for measuring both simple and complex learning objectives. These topics are covered in units 3 and 4.
4. To develop ability to evaluate the quality of test items through item analysis. Different techniques of item analysis have been discussed in unit 5.
5. To develop understanding of standardized scores and norms and the ability to use these scores in explaining test results. This is the subject of unit 6.
6. To develop understanding of reliability and validity of test and the ability to use appropriate method for establishing test reliability and validity. Units 7 and Unit 8 deal with these issues.
7. To develop the understanding of aptitude tests and their use in educational settings. This is the topic of unit 9.

In the preparation of this study guide, internet sources have been used. Many websites, related to educational assessment, educational psychology and statistics, are consulted and their references have been given at the end of each unit. Separate folder for each unit has been created. Each folder contains web based material (5-8 on line articles and other internet sources). If studied carefully, these will serve a good source of information for building strong foundations in the field of assessment and evaluation.

Finally, I am thankful to all those who have been helpful in the preparation of this study guide. Maliha Nasir, a student of Ph.D. scholar from International Islamic University, Islamabad, provided me necessary assistance in organizing the material in their relevant folders. Above all it is all due to Allah (SWT) who was and has been the main source of encouragement.

(Prof. Dr. Rehana Masrur)
Chairperson/Dean
Faculty of Education
## CONTENTS

### UNIT 1  MEASUREMENT AND EVALUATION

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contents</td>
<td>3</td>
</tr>
<tr>
<td>1.1 Introduction of the Unit</td>
<td>5</td>
</tr>
<tr>
<td>1.2 Objectives of the Unit</td>
<td>5</td>
</tr>
<tr>
<td>1.3 Definition of Terms Related to Assessment and Evaluation</td>
<td>5</td>
</tr>
<tr>
<td>1.4 Purpose of Measurement, Assessment and Evaluation in Education</td>
<td>7</td>
</tr>
<tr>
<td>1.5 Basic Assumptions that Underlie Educational Evaluation</td>
<td>9</td>
</tr>
<tr>
<td>1.6 Components of Classroom Assessment</td>
<td>10</td>
</tr>
<tr>
<td>1.7 Relation between Instruction and Assessment</td>
<td>11</td>
</tr>
<tr>
<td>1.8 Basic Competencies that all Teachers must Master</td>
<td>12</td>
</tr>
<tr>
<td>1.9 Educational Assessment in the 21st Century</td>
<td>13</td>
</tr>
<tr>
<td>1.10 Self Assessment Questions</td>
<td>15</td>
</tr>
<tr>
<td>1.11 Suggested Books</td>
<td>15</td>
</tr>
<tr>
<td>1.12 Web Based References on CD and Related Websites</td>
<td>15</td>
</tr>
</tbody>
</table>

### UNIT 2  EDUCATIONAL OBJECTIVES

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contents</td>
<td>19</td>
</tr>
<tr>
<td>2.1 Introduction of the Unit</td>
<td>21</td>
</tr>
<tr>
<td>2.2 Objectives of the Unit</td>
<td>21</td>
</tr>
<tr>
<td>2.3 Characteristics of Educational Objectives</td>
<td>21</td>
</tr>
<tr>
<td>2.4 Types of Objectives</td>
<td>22</td>
</tr>
<tr>
<td>2.5 Taxonomy of Educational Objectives</td>
<td>24</td>
</tr>
<tr>
<td>i. Cognitive Domain</td>
<td>24</td>
</tr>
<tr>
<td>ii. Affective Domain</td>
<td>27</td>
</tr>
<tr>
<td>iii. Psychomotor Domain</td>
<td>29</td>
</tr>
<tr>
<td>2.6 What is the Relevance of Affective Domain in Education</td>
<td>29</td>
</tr>
<tr>
<td>2.7 Combining Objectives and Types of Knowledge</td>
<td>31</td>
</tr>
<tr>
<td>2.8 Other Sources of Objectives</td>
<td>33</td>
</tr>
<tr>
<td>2.9 Aligning Objectives and Instruction</td>
<td>33</td>
</tr>
<tr>
<td>2.10 Stating Behavioral Objectives</td>
<td>34</td>
</tr>
<tr>
<td>2.11 Self Assessment Questions</td>
<td>34</td>
</tr>
<tr>
<td>2.12 Suggested Books</td>
<td>34</td>
</tr>
<tr>
<td>2.13 Web Based Material on CD</td>
<td>35</td>
</tr>
</tbody>
</table>

### UNIT 3  MEASUREMENT METHODS AND TECHNIQUES -1

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contents</td>
<td>39</td>
</tr>
<tr>
<td>3.1 Introduction of the Unit</td>
<td>41</td>
</tr>
<tr>
<td>3.2 Objectives of the Unit</td>
<td>41</td>
</tr>
<tr>
<td>3.3 Types of Objective Test Items</td>
<td>42</td>
</tr>
<tr>
<td>3.4 Specifying the Instructional Objectives</td>
<td>43</td>
</tr>
<tr>
<td>3.5 Preparing Test Specifications</td>
<td>43</td>
</tr>
<tr>
<td>3.6 Writing Test Items (General Tips)</td>
<td>44</td>
</tr>
<tr>
<td>3.7 Writing Specific Types of Items</td>
<td>45</td>
</tr>
</tbody>
</table>
3.8 Preparing Test Directions ............................................. 47
3.9 Correction for Guessing .............................................. 47
3.10 Self Assessment Questions ........................................ 48
3.11 Suggested Readings .................................................. 49
3.12 Web Based References and Related Web Sites on CD .... 49

UNIT 4 MEASUREMENT METHODS AND TECHNIQUES-II
Contents.............................................................................. 53
4.1 Introduction of the Unit ............................................... 55
4.2 Objectives of the Unit .................................................. 55
4.3 Purposes of Essay Testing ........................................... 55
4.4 Kinds of Essay Tests .................................................... 56
4.5 Guidelines for Writing Essay Items .............................. 57
4.6 Guidelines for Writing Short-Answer Items .................. 58
4.7 Scoring Essay Tests ..................................................... 59
4.8 Scoring Short-Answer Items ........................................ 61
4.9 Advantages and Limitations of Essay Tests as Achievement Measures... 61
4.10 Aligning Assessment and Instruction ......................... 62
4.11 Self Assessment Questions ......................................... 63
4.12 Recommended Readings ............................................ 63
4.13 Web Based Material on CD ........................................ 63

UNIT 5 APPRAISING THE ACHIEVEMENT TESTS
Contents ............................................................................... 67
5.1 Introduction of the Unit ............................................... 69
5.2 Objectives of the Unit .................................................. 69
5.3 Analyzing the Effectiveness of Test Items ...................... 69
  1. Item Difficulty Index (or Item Difficulty Level) .............. 70
  2. Item Discrimination ................................................... 71
  3. Item Analysis of Speed Tests ........................................ 74
  4. Distracter Analysis ..................................................... 75
5.4 Application of Item Analysis for Teachers ..................... 76
5.5 Using Item Analysis to Improve Items ....................... 78
5.6 Qualitative Item Analysis ............................................ 79
5.7 Using Item Analysis to Improve Classroom Instruction .... 79
5.8 Self Assessment Questions .......................................... 80
5.9 Suggested Readings .................................................. 81
5.10 Web Based References and Related Web Sites ............. 81

UNIT 6 TEST SCORES AND NORMS
Contents ............................................................................... 85
6.1 Introduction of the Unit ............................................... 87
6.2 Objectives of the Unit .................................................. 87
6.3 What is a Raw Score? .................................................. 87
6.4 What are Norms? ....................................................... 88
| 6.5 | Norm-Referenced Tests (NRT) and Criterion-Referenced Tests (CRT) | 88 |
| 6.6 | Norm-Referenced Interpretations | 89 |
| 6.7 | The Normal Curve | 90 |
| 6.8 | Derived Scores | 93 |
| 6.9 | Types of Norms | 96 |
| 6.10 | Criterion-Referenced Interpretations | 97 |
| 6.11 | Norm-Referenced VS Criterion-Referenced Tests | 98 |
| 6.12 | Qualitative Description of Scores | 99 |
| 6.13 | Self Assessment Questions | 99 |
| 6.14 | Suggested Readings | 100 |
| 6.15 | Web Based Material | 100 |

**UNIT 7**  
THE RELIABILITY OF MEASUREMENT METHODS  
Contents.................................................................103  
7.1 Introduction of the Unit......................................105  
7.2 Objectives of the Unit .......................................105  
7.3 Reliability.........................................................106  
7.4 Errors of Measurement.........................................107  
7.5 Sources of Measurement Error.............................108  
7.6 Forms of Reliability............................................109  
7.7 Factors Affecting the Reliability Estimates...............111  
7.8 Improving the Test Reliability.............................112  
7.9 The Standard Error of Measurement.......................112  
7.10 Self Assessment Questions..................................113  
7.11 Additional Readings..........................................114  
7.12 Web Based References.......................................114

**UNIT 8**  
THE VALIDITY OF MEASUREMENT METHODS  
Contents.................................................................117  
8.1 Introduction of the Unit......................................119  
8.2 Objectives of the Unit .......................................119  
8.3 Validity.............................................................119  
8.4 Types of Validity................................................120  
8.5 Factors Affecting Validity....................................122  
8.6 Relationship between Reliability and Validity...........123  
8.7 Unitary Concept of Validity..................................124  
8.8 Interpreting Validity Coefficients..........................126  
8.9 Convergent and Discriminant Evidence of Validity ......127  
8.10 Practical Strategies for Teachers..........................128  
8.11 Self Assessment Questions..................................130  
8.12 Suggested Books................................................130  
8.13 Web Based References.......................................130
UNIT 9  APTITUDE TEST
Contents......................................................................................................................133
9.1 Introduction of the Unit..................................................................................135
9.2 Objectives of the Unit ..................................................................................135
9.3 The Nature of Human Abilities ...................................................................135
9.4 Intelligence ....................................................................................................136
9.5 Aptitude Tests ...............................................................................................137
9.6 Intelligence Tests ............................................................................................139
9.7 Diagnostic Tests ............................................................................................140
9.8 The Differential Ability Scales .....................................................................141
9.9 Creativity as Special Ability ..........................................................................142
9.10 Self Assessment Questions ..........................................................................144
9.11 Suggested Readings .....................................................................................145
9.12 Web Based References ................................................................................145
Unit–1

MEASUREMENT AND EVALUATION
CONTENTS

1.1 Introduction of the Unit ................................................................. 5
1.2 Objectives of the Unit ................................................................. 5
1.3 Definition of Terms Related to Assessment and Evaluation ............... 5
1.4 Purpose of Measurement, Assessment and Evaluation in Education........ 7
1.5 Basic Assumptions that Underlie Educational Evaluation...................... 9
1.6 Components of Classroom Assessment ........................................... 10
1.7 Relation between Instruction and Assessment .................................... 11
1.8 Basic Competencies that all Teachers must Master............................... 12
1.9 Educational Assessment in the 21st Century ...................................... 13
1.10 Self Assessment Questions ........................................................... 15
1.11 Suggested Books ........................................................................... 15
1.12 Web Based References on CD and Related Websites .......................... 15
1.1 Introduction of the Unit

Teacher preparation programs require from student-teacher to teach. Teaching should not be conceptualized only as a process of instructions provided to the students to learn. This simple perspective is seen as instruction-learning process. In actual practice, it is more realistic to view assessment as an integral component of the teaching process. We see that teachers spend at least one-third of their professional time to assessment related activities. Since assessment and evaluation provide relevant information that enhances instruction and promote learning. There is indeed a close and reciprocal relationship between instruction, learning and assessment. This broader conceptualization of teaching, learning, assessment and evaluation are integrally related. Assessment and evaluation provide feedback about what the student has learned? What is the quality of learning experiences, how effective the instruction has been? And what information, concepts, and objectives require more attention. If you want to be an effective teacher, you need to know about testing and assessment. Before we proceed to further it will be beneficial for you to review the basic concepts and language of measurement, assessment and evaluation in this unit that you might have encountered during your teaching experiences or as a student of teacher education program. You will also study assessment related topics such as: function and role of assessment and evaluation, relationship between instruction and assessment, basic competencies that all teachers must know and assessment in 21st century.

1.2 Objectives of the Unit

After studying this unit, printed allied material, and web based material on CD you shall be able to:
1. Describe and explain the importance of test, educational measurement, educational assessment and evaluation as indicator of educational effectiveness.
2. Evaluate the purpose of educational assessment and evaluation.
3. Analyze the assumptions underlying educational assessment and measurement.
4. Examine the major application of assessment and evaluation in schools.
5. Describe and explain the role of assessment and evaluation in teaching.
6. Identify the problems faced by the educational administrators in measuring and evaluating students’ progress.
7. Compare some major trends in educational assessment at global level.

1.3 Definition of Terms Related to Assessment and Evaluation

1. Test
   A very common term which explains a procedure, in which a sample of an individual’s behavior is obtained, evaluated and scored using standardized procedures. Some educators may call it a device in which a sample of individual’s behavior is obtained, scored, and evaluated, again using some standardized procedures. This definitely is a general definition, but at this point we will focus on this generic definition. Because a test is only a sample of behavior and it reflects the representative sample of the
behavior. A teacher’s assessment should assess content areas in accordance with relative importance a teacher has assigned to them. The detail of test development have been discussed in unit-3

2. Educational Measurement

The term refers to any device for the general study and practice of testing, scaling and appraising the outcomes of educational process. It includes the administration and application of statistical techniques in the interpretation of obtained measures or test results. This is a broader concept of educational measurement. Popham (2005) while defining the educational measurement states: “A process by which educators use student’s responses to specially created or naturally occurring stimuli in order to make inferences about students’ knowledge, skills or affective status” (p.3).

In its typical meaning we can say that, measurement is the process of quantifying or assigning a number to an attribute of person, object, idea or any measurable thing per Assigning numbers requires setting of rules for assigning numbers of person represent objects, traits, attributes, or behaviors. An educational test is a measuring device and therefore involves rules (e.g. administration guidelines and scoring criteria) for assigning numbers that represent an individual’s performance. In turn these numbers are interpreted as reflecting characteristics of the test taker. For example the number of words spelled correctly on a spelling test might be interpreted as reflecting a student’s spelling skills.

3. Assessment

The process of collecting and interpreting information to aid in decision making is called ‘assessment’. This term embraces diverse kind of tests and measurements. Popham (2005) defines educational assessment in the following words; “Educational assessment is a formal attempt to determine student’s status with respect to educational variables of interest” (p.6).

It is clear from the definition that educators use student’s responses to specially created or naturally occurring stimuli for making inferences about student’s knowledge, skills, or affective status (Popham, 2005). The assessment of student’s knowledge, skills, or affective status requires identification of educational domains. Since educational domains are too large, therefore teachers need to represent the assessment domain the student is expected to learn. Sampling of assessment domain expects from teachers to be knowledgeable about the purpose of assessment.

4. Educational Evaluation

Evaluation is broader in scope and seeks the value of object that has been measured through defined procedures. It is the process of carefully appraising the individual from a variety of information giving devices. Besides testing and other tools of measurement, evaluation seeks additional evidences from various sources of information supplementing each other; like interviews, questionnaires, anecdotal records, cumulative records, case studies, electronic records, or projective techniques, etc., and the selection through careful analysis of data most pertinent to a wise, just, and comprehensive interpretation to make value judgment of individual, or a group. In short words evaluation involves judging of quality about student, instruction, or classroom climate and finds the answer of
‘what value?’ Evaluation further involves the extent to which the learning objectives have been achieved. The following diagram explains the relationship among main components of teaching-learning process:

![Diagram of Teaching-Learning process]

Figure 1.1: Teaching-Learning process

The two-way arrow indicates that evaluation is an ongoing process and that the four concepts are interrelated.

Most of the definitions of evaluation represent one of the following two philosophical view points.

1. Evaluation is the systematic process of collecting and analyzing data in order to determine whether and to what degree objectives have been or are being achieved.
2. Evaluation is the systematic process of collecting and analyzing data in order to make decisions.

Though the following two functions are common in both approaches:

i. Systematic process of data collection, and
ii. Analysis of collected data,

The second definition is most appropriate because it is more inclusive than the first one, since the decision will be based on the achievement status of objective(s).

Read Article 1.1 ‘Measurement, Assessment and Evaluation’ and Article 1.2 ‘Assessment and Evaluation’ in web material folder Unit-1 on CD, for explanations of the terms by Kizlik(2006).

1.4 Purpose of Measurement, Assessment and Evaluation in Education

Educational assessment and evaluation is indispensable and is inherent to effective learning. Its purpose is to improve the curriculum, the educational program, or the teaching-learning process based on the findings and generalizations made through the evaluation of the intellectual, physical, moral, emotional, spiritual, or social growth and development of the pupils.

The major concern of the educational measurement and evaluation is of course the pupils’ present status. However a second important concern is with the contribution of pupils’
past environment, growth & development, and learning to his present and the past can contribute greatly to the effective guidance of the pupil, not only his present classroom and other school experiences but also in his later school activities. Teachers assess for many purposes because they are required to make many different kinds of decisions. The purpose of all assessment is to gather information to facilitate effective decision making. You can get the knowledge of the many purposes teachers have for assessment.

1. **Measurement and Assessment for Maintaining Equilibrium**
   One of the purposes of assessment is to maintain the orderliness in the classroom. As you know that classrooms are complex social settings in which students and teachers interact with one another in a multiple ways. For creating healthy social and learning environment in classrooms order, discipline and cooperation are essential. Orderliness helps facilitate learning and it is needed for successful teaching and learning.

2. **Planning, Structuring, and Conducting Instruction**
   Since instructions make the core of classroom activities, many of the decisions made by teachers focus on planning, structuring and conducting instructional activities. The instruction decisions made by a teacher can be divided into three types: planning decisions, structuring decisions and teaching decisions. Planning and structuring decisions are made for future instructional activities, whereas, actual teaching process in classroom also requires on the spot assessment and decision making. Hence, a large proportion of teacher assessment is concerned with planning and delivering instructions.

3. **Placing Pupils**
   Classroom teachers often have to make decisions about appropriate placement of pupils. Assessment is required for making such decisions as dividing pupils into different groups, arranging cooperative learning groups, making teams for project works, etc.

4. **Providing Feedback and Incentives**
   Classroom assessment is an important source of providing feedback and incentives to pupils. Information about academic performance and social behavior of students is used to provide feedback to pupils and parents. This type of feedback, which is provided during the course of instruction to improve students’ learning, is termed as *formative assessment*. Constant assessment of students’ learning and behavior is required for providing such feedback.

5. **Diagnosing Pupil Problems**
   Teachers keep observing pupils’ learning, emotional, or social problems in the classroom. It helps the teacher to know the difficulties the pupils are facing in the process of learning. Having identified such problems, the teacher can sometimes carry out the remedial activities needed. In our schools teacher helps the pupils with the consultation of parents to overcome the diagnosed problem or may refer for more specialized diagnosis and remediation outside of the classroom. Many private schools in our country provide guidance and counseling service to their pupils. Much of the assessment data that teacher collect is used to identify, understand, and remediate pupils’ problems and learning difficulties.
6. **Grading Academic Learning and Progress**

Teachers spend much of their time on collecting information that have to be used to grade pupils or make final judgment about their academic progress. The collection of information for making final decision and assigning grades at the end of instruction is termed as **summative assessment**. Because of summative evaluation teacher can evaluate the effectiveness of his/her teaching and quality of learning that has taken place in the learner. Read:

*Article 1.2* Black (nd) highlighted three purposes of assessment and evaluation: certification, accountability and improving learning. Read article 1.3 ‘Evaluation and Assessment’ in folder 1 on CD.

Evaluation and measurement plays multifaceted role. It can be summarized as:

1. Testing for Accountability
2. Integrating Instruction and Assessment
3. Classroom Assessment
5. For Making Grading Decisions
6. For Incorporating Recent Trends
7. Integrating Research on Learning
8. 

**NOTE:** These topics have been discussed in detail by McMillan (2007) book please read Chapter 1, ‘Role of Assessment in Teaching’, pp. 1-26 and do the following Action Research:

<table>
<thead>
<tr>
<th>Activity – I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administer a questionnaire to 10 teachers of your school and ask some questions about recent trends in assessment. Consult table 1.4 (p.18) and figure 1.4 (p.22) of McMillan (2007) book for formulation of questions. (Summarize the results)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Activity – II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Construct a questionnaire about the external and internal factors that affect the assessment, grading practices, and decision making (consult figure 1.4 of this chapter 1 of above mentioned book). Collect the views of 20 elementary school teachers and prepare a report.</td>
</tr>
</tbody>
</table>

1.5 **Basic Assumptions that Underlie Educational Evaluation**

After studying the basic concepts and purposes of measurement, assessment, and evaluation now you can draw the underlying assumptions. Cohen and Swerdlik (2002) suggested that adoption of these assumptions has simplified the complex issues related to assessment. After examining these assumptions you will be able to develop better understanding of these complex and interrelated issues.
1. **Measurable Psychological and Educational Construct.**

   In assessment terminology, a construct is simply the trait or characteristic that a test is designed to measure. For example achievement is a construct that shows an individual’s knowledge in areas in which he has received instruction. In schools we are interested in measuring a number of constructs, such as student’s intelligence, achievement in specific content area, or attitude towards learning. This assumption is the realization of the fact that there exists a number of educational and psychological constructs (Reynolds, Livingston, and Willson, 2006: p.6).

2. **Educational and psychological construct can be measured.**

   It is an established fact that whatever exists can be measured. From an idea to a word, from non-physical object to physical object, if it exists in some amount, it can be measured (Cronbach, 1990). Reynolds, Livingston, and Willson (2006, p.11) lamented that “If we accept the assumption that educational and psychological constructs exist, the next question is ‘can these construct be measured?’” Assessment experts believe that educational and psychological constructs can be measured.

3. **Measurement of construct is not perfect.**

   Assessment experts believe that they can measure educational and psychological constructs; they also acknowledge that the measurement process is not perfect. This assumption is based on the fact that measurement in physical sciences can be more reliable and accurate because it can be done in very control condition. Educational and psychological constructs are framed in terms of measurement error and it has effects on the reliability of scores. Some degree of error is inherent in all measurement, and measurement error reduces usefulness of measurement. To overcome this, the experts make considerable efforts to minimize the effects of measurement error (Reynolds, Livingston, and Willson, 2006: p.10)

4. **There are different ways of measuring any given construct.**

   Besides traditional classroom paper-pencil tests there are many other approaches for measuring abilities, skills, different attributes and traits of personalities: like observation, interview, checklist, rating scales etc.

5. **Usefulness of assessment procedures:**

   Assessment devices have their own weaknesses and strength. In other words not all instruments are equally useful. The usefulness of a device depends on the purpose for its use. Moreover there are situations in which the evaluator chooses from different devices the one which is most appropriate and will reveal maximum information.

1.6 **Components of Classroom Assessment**

   In the preceding sections we have discussed the basic concepts of assessment and evaluation; this is an appropriate time to highlight the components of classroom assessment. McMillan (2007) explains four components of classroom assessment. The components are:
1.7 Relation between Instruction and Assessment

In the previous sections common applications of educational assessment have been highlighted in detail. Gronlund has shown a close relationship of assessment and instruction in table 1.1 on page 4 of his book. And he has simplified the model in figure 1.1 (p.5). The focus of his discussion in this model is on the instructional role of placement assessment. How placement assessment takes place at different stages have been elaborated in the preceding pages.

The Gronlund (2006) emphasizes on the relation between instruction and assessment. He argues that the main concern of a teacher during instruction should be on evaluating the progress of students. For monitoring the progress of students formative tests are required. These tests are used to diagnose the learning problems. Formative evaluation detects cognitive strengths and weaknesses and provides feedback on effectiveness of instruction. Gronlund says that “the aim here, as with formative testing, is to monitor learning progress and to provide corrective prescriptions to improve learning” (p. 7). Figure 1.2 at page 8 explains the instructional role of formative assessment. Related with formative evaluation is summative evaluation, which is carried out at the end of instructional program. On the basis of summative evaluation teacher assigns grades.

The main purpose of assessment and evaluation in education is to record the achievements of individual pupils for the purpose of certification. Black (nd) elaborated that assessment provides information/feedback about students’ learning, on the basis of which action may be taken to address learning needs of each pupil. This is what a formative or diagnostic evaluation does. Popham (2006) while conceptualizing the relationship between assessment and instruction presented two approaches. And these two approaches he says are: Instruction-influenced Assessment and Assessment-influenced Instruction approaches. In both approaches the curriculum is guided by the curriculum objectives but in later approach assessment devices are prepared before imparting instruction. Here the assessment plays a major role. See figure 1.2 and 1.3.

1st Approach

**Instruction-Based-Assessment**

![Diagram](Diagram.png)

Figure 1.2: Relationship between Instruction and Assessment
2nd Approach
Assessment-Based-Instruction

Figure 1.3: Relation between Instruction and Assessment

Examining these two approaches and judge which one is more appropriate
McMillan (2007), in chapter 1 of his book, thoroughly elaborates the role of
assessment in teaching from different angles. Concept map on first page of his book
integrates micro level and macro level components of the role of assessment. For example:

1. Integrating instruction and assessment: In this he elaborates component realities of
teaching and instructional decision making. Figure 1.1 (page 7) further highlights
the role of assessment in teaching in eight steps.

2. Classroom assessment focuses on four components: purpose, measurement,
evaluation, and use.

3. Assessment standards: At initial stage assessment standards are decided, and then
developing, interpreting, using, grading, and communicating stages are designed.

4. Internal beliefs, values and some external factors are considered for making the
decisions about grades.

5. Emerging recent trends e.g. assessment for learning, quality standards, students
involvement with assessment, students self evaluation, alternative assessment etc
have been discussed.

6. The role of assessment with respect to cognitive theories instruction.

1.8 Basic Competencies that all Teachers must Master

Assessment plays an important role in schools and teachers devote so much of their
time to assessment-related activities, there are some basic competencies, which all
teachers should master.

1. Teachers should be proficient in developing assessment procedures appropriate for
making instructional designs. In fact, the vast majority of the assessment information
teachers collect and use on a daily basis comes from teacher-made tests. To accomplish
this, teachers must be familiar with the principles and standards for developing a wide
range of assessment techniques including selected-response items, constructed-
response items, performance assessments, and portfolios. Teachers must also be able to
evaluate the technical quality of the instruments they develop.

2. Teachers should be proficient in administering, scoring, and interpreting
professionally developed and teacher-made tests. In addition to being able to
develop good assessment procedures teachers must be able to use them
appropriately. Teachers need to understand the principles of standardization and
prepared to administer tests in standardized manner. They should be able reliably
and accurately to score a wide range of procedures including selected-response items, constructed-response items, and portfolios. Teacher must be able to interpret the scores reported on standardized tests, such as percentile ranks and standard scores. These procedures also require from teachers to be proficient in statistical and psychometric concepts. These concepts will be discussed in later units.

3. Teachers should be competent in using assessment results when making educational decisions. Teachers use assessment results for making a wide range of educational decisions. These decisions include: students’ evaluations, instructional planning, curriculum development, and educational policies. Because teachers play such a pivotal role in using assessment information in the schools they must be able to interpret assessment accurately. They need to comprehend the concept of reliability and validity and be prepared to interpret test results in an appropriate and cautious manner. Teachers should understand and be able to describe implications and limitations of assessment results.

4. Teachers should be competent in developing valid grading procedures. Assigning grades to students is an important aspect of teaching. Teachers must be able to develop and apply fair and valid procedures for assigning grades based on the performance on tests, homework assignments, class projects, and other assessment procedures.

5. Teachers should be able to communicate assessment results. Teachers report assessment results to students and parents, and to stakeholders. Therefore, teachers must be able to use assessment terminology correctly, understand score formats, and explain the meaning and implications of assessment results. Teachers must be able to explain grading practices. They should be able to describe the strength and limitations of different assessment procedures.

6. Teachers should be able in recognizing unethical, illegal, and inappropriate uses of assessment procedures. It is essential that teachers could be familiar with the ethical codes that apply to the educational assessment procedures.

7. Teachers should be able in selecting appropriate assessment tools for making instructional decisions. This competency requires from teachers to select professionally developed tests that are appropriate for the situation, technically appropriate, fair, and that provide useful information. This requires that teachers should be familiar with wide range of educational and psychological tests that are applicable in school setting. Teachers need to be able to locate, interpret, and use technical information and critical reviews of professional tests. McMillan (2007) discussed eleven basic principles to guide the assessment training and professional development of teachers and administrators.

Read Article 1.4 ‘Fundamental Assessment Principles’ in web based material folder Unit- 1 on CD.

1.9 Educational Assessment in the 21st Century

The field of educational assessment is dynamic and is on going activity. Though classical test theory is still very much useful, however, many aspects of educational assessment and evaluation are evolving because of number of internal and external factors. Theoretical and technical advances at global level have led the teachers to bring
changes in the patterns of assessment. It is important for better assessment for the professionals to stay informed regarding new developments that are taking place at national and international levels.

Similarly research on learning, motivation, instruction, and curriculum has important implications for assessment. Research on learning suggests that domain referenced interpretation will emerge and abilities will be interpreted in reference with cognitive processes and structure that are required to solve the assessment problems and tasks. Instead of focusing on quantitative aspects of performance, future assessment will focus on qualitative aspects of test performance. And dynamic testing will become an important force in ability testing. It is predicted that self efficacy and self confidence will emerge significant determinants of motivation. Feedback about how well students are doing serves as motivational force.

As long as the curriculum is concerned, a standard based movement is emerging at global level. Higher standards of learning will enhance the quality of learning. Relevancy of content with the world outside of school needs to be applied.

Although researchers expect changes to occur rapidly over the next few decades, but the basic psychometric principles will still be important and will remain fundamental aspects of the profession.

Many advanced countries have brought significant changes in curriculum theory and assessment. They are setting high standards of learning for all students. Such approaches have been discussed in detail in allied material (print form) and web material provided you in CD.

**Activity – III**

Students are directed to study first chapter of McMillan book (2007), page no. 18-20 and explain the following terms:

High-Stakes, NCLB, and AYP

**NOTE:**

1. Please check your understanding of the content of this unit by solving the self-instructional review exercises given at the end of allied material related to this unit.
2. After reading allied material and assessing your understanding on evaluation measurement and assessment, you are advised to read web based journal articles. These articles will expose you to the practical applications of assessment beyond classroom.

*For improving the assessment Matter (1999) has discussed some strategies in Article 1.5 in web Material folder Unit- 1 on CD*

After reading the unit, allied material and related web material, try the following self-assessment questions.
1.10 Self Assessment Questions

1. Give at least four examples while differentiating among test, measurement, assessment, and evaluation.
2. What are the main components of classroom assessment? Keeping in view educational setup of our public schools how could we make the best uses of assessment?
3. Discuss your point of view about the assumption that ‘measurement of psychological construct cannot be perfect’.
6. Explain the theoretical and practical advances in educational assessment at national and global level.
7. Prepare a lesson /lecture on the importance of assessment and evaluation in education that you are supposed to teach to elementary teacher trainees.
8. Discuss and explain your point of view about “Assessment-Influenced Assessment and Instruction-Influenced Assessment”.

Suggested Books


1.12 Web Based Material on CD and Related Websites

EDUCATIONAL OBJECTIVES
## CONTENTS

2.1 Introduction of the Unit .................................................................21
2.2 Objectives of the Unit .................................................................21
2.3 Characteristics of Educational Objectives .......................................21
2.4 Types of Objectives .................................................................22
2.5 Taxonomy of Educational Objectives ............................................24
   i. Cognitive Domain .................................................................24
   ii. Affective Domain ...............................................................27
   iii. Psychomotor Domain ........................................................29
2.6 What is the Relevance of Affective Domain in Education ..................29
2.7 Combining Objectives and Types of Knowledge ..............................31
2.8 Other Sources of Objectives ........................................................33
2.9 Aligning Objectives and Instruction ............................................33
2.10 Stating Behavioral Objectives ......................................................34
2.11 Self Assessment Questions .......................................................34
2.12 Suggested Books ........................................................................34
2.13 Web Based Material on CD ........................................................35
2.1 Introduction of the Unit

Assessment procedure should closely be tied to the curriculum and its objectives. The initial steps in developing a test are to specify the educational objectives, develop a table of specification, and select the type of items you will include in your assessment. These activities provide the foundation for all classroom tests. The identification and statement of educational objectives is an important first step in developing tests. Educational objectives are simply goals, that is, what we hope the student will learn or accomplish. Educational objectives are also referred as instructional or learning objectives. These objectives are sometimes clearly stated and sometimes implicit. A good classroom test can be written from clearly stated objectives much more easily than can one from vague or poorly developed objectives. Clearly stated objectives help to make sure that the test measures what has been taught in class and greatly facilitate the test development process. Establishing explicit, clearly stated objectives also has the added benefit of enhancing the quality of teaching. When a teacher knows what the educational objectives then he/she will be able to teach effectively. This unit summarizes the types and characteristics of objectives, sources of objective, stating cognitive, affective and psychomotor objective and how to align objectives and instruction.

2.2 Objectives of the Unit

After reading and studying this unit, printed allied material and web based material on CD you should be able to:
1. Describe the importance of educational objectives in terms of both instruction and assessment.
2. Describe characteristics of educational objectives.
3. Describe and give examples of how educational objectives can differ in terms of scope.
4. Describe Bloom’s taxonomy of cognitive objectives. Explain and give examples of each category.
5. Describe and give examples of the three domains covered by educational objectives.
6. Describe and give examples of behavioral and non-behavioral objectives.
7. Illustrate a thorough understanding of principles for writing affective educational objectives
8. Write objectives for a specified content area.

2.3 Characteristics of Educational Objectives

Classroom tests should reflect what was taught in class, and test should emphasize what was emphasized in class. It is probably best to begin by describing some of the characteristics of educational objectives. The most prominent characteristics of educational objectives involve their scope and domain. The description of these characteristics as follows:
Scope
Scope refers to how broad or narrow an objective is. An example of a broad objective is:

- The student will be able to analyze and discuss the effects of 1971 Indo-Pak war on Pakistani politics.

An example of a narrow or specific objective is:

- The student will be able to list the causes of 1971 Indo-Pak war.

Clearly different kinds of student responses would be expected for tests questions developed from such objectives. Objectives with a broad scope are often broken down into objectives with a more narrow scope. It is probably best to strike a balance between broad objectives and narrow objectives. If you use only broad educational objectives, you may not have specific information needed to help you develop tests with good measurement characteristics. The balance between broad and narrow objectives can best be accomplished using two approaches. The first approach requires writing objective at intermediate level of specificity. Here the goal is to write objective that provide the specificity necessary to guide test development but are not so narrow as to limit assessment to low level abilities. The second approach is to use a combination of broad and specific objectives. That is, write broad objectives that are broken down into more specific objectives. Either of these approaches can help you develop well-organized tests with good measurement characteristics.

Domain
Objectives are written at several levels, including programs, grades, subjects, classroom, unit plan, and at several degrees of specificity, from broad to precise. The most popular domains to formulating objectives are based on the work of Bloom, Krathwohl, and Dave. The cognitive domain developed by a committee of 36 researchers from different universities headed by Benjamin Bloom. This domain includes objectives that are related recall or recognition of knowledge, and the development of higher intellectual skills and abilities. Affective domain by David Krathwohl and associate is concerned with aims and objectives related to interests, attitudes, and feelings. The description of psychomotor domain, deals with manipulative and motor skills.

2.4. Types of Objectives:

Objectives can range from very general to very specific. The following two examples will clarify this range:

i. The pupil will become mathematically literate.

ii. The pupil can add three two-digit numbers.

In this example objective (i) is more general and objective (ii) is more specific. Since objectives vary widely in specificity, therefore they can be divided in three levels of abstraction that represents the degree of objective specificity. Examine the following three levels:
1. **Goals**

These global objectives often mentioned as ‘Goals’ are broad, complex learning outcomes that require longer time and instruction to accomplish. They are very general, encompassing a large number of more specific objectives. The objective in the above example in global term is a ‘Goal’ that is broadly inclusive. Because they are broadly inclusive therefore are rarely used in classroom assessment unless they are broken down into more specific objectives. Narrow and specific objective are needed to meet classroom needs.

2. **Educational Objectives**

Educational objectives represent a middle level of abstraction. Examples of educational objectives include: “The pupil can read English poetry loud”. They are sufficiently narrow and more specific to help teachers plan and focus teaching. These middle level objectives suggest a range of possible students outcome associated with the objectives.

3. **Instructional Objectives**

These types of objectives indicate the specific behaviors that students must demonstrate to indicate that learning has occurred and are termed as behavioral objectives. Instructional objectives are the least abstract and most specific type of objective. For example: “The pupil can correctly punctuate sentences”, “The pupil can add two four-digit numbers”. These examples show that instructional objectives focus teaching on relatively narrow topic of learning in content area. These concrete objectives are used in planning daily lessons.

For further detail read the following material:


**Gronlund, N.E. (2006).** *Assessment of Student Achievement.* (Chapter 4, pp. 41-53)

(Open folder 2 and read the following articles for the clarification of learning objectives:

*Article 2.1 by Ritz (2005) on ‘Aims, Goals and Objectives’*

*Article 2.2 by Arreola (1998) on ‘Writing Learning Objectives’*

*Article 2.3 on ‘General Learning Objectives’*

McMillan spreads out the concept of learning targets and standards into following components:

1. Standards
2. Sources of Learning Targets and Standards
3. Criteria for Selecting Learning Targets and Standards
4. Types of Learning Targets
5. Goals, Objectives, and Targets
Read the chapter of his book (mentioned in the box above) carefully and try to solve self-instructional review exercises given on page 53. McMillan has given five suggestions for carrying out Action Research. Do it as directed by the author.

2.5. **Taxonomy of Educational Objectives**

Taxonomy” simply means “classification”, so the well-known taxonomy of learning objectives is an attempt (within the behavioral paradigm) to classify forms and levels of learning. It identifies three “domains” of learning (see below), each of which is organized as a series of levels or pre-requisites. It is suggested that one cannot effectively — or ought not try to — address higher levels until those below them have been covered (it is thus effectively serial in structure). As well as providing a basic sequential model for dealing with topics in the curriculum, it also suggests a way of categorizing levels of learning, in terms of the expected ceiling for a given programme. Thus in the Cognitive domain, training for technicians may cover knowledge, comprehension and application, but not concern itself with analysis and above, whereas full professional training may be expected to include this and synthesis and evaluation as well.

In addition to the scope of educational objectives, they also differ in the domain or the type of ability characteristics being measured. The domains typically addressed by educational objectives involve cognitive, affective, and psychomotor abilities or characteristics. These three domains are usually presented as hierarchies that involve different levels that reflect varying degrees of complexity.

1. **Cognitive Domain**

The most-used of the domains, refers to knowledge structures (although sheer “knowing the facts” is its bottom level). It can be viewed as a sequence of progressive contextualization of the material (Based on Bloom, 1956)

Bloom’s taxonomy provides a useful way of describing the complexity of an objective by classifying it into one of six hierarchical categories ranging from the most simple to the most complex. Table 2.1 and figure 2.1 provide a summary of Bloom’s Taxonomy. The categories include the following:

i) **Knowledge**

The simplest level of taxonomy is knowledge. Objectives at the knowledge level involve learning or memorizing specific facts, terms, names, dates, and so forth. Examples of educational objectives in the knowledge category include;

- The student will be able to name each provincial capital.
- They will be able to list the names of prime ministers of Pakistan in the order they served.

ii) **Comprehension**

Objectives at the comprehension level require understanding, not simply rote memorizing. Objectives at this level often use verb such as interpret, translate, explain, and summarize. Examples of educational objectives at the comprehension level include:
The student will be able to describe the use of each symbol on Pakistan’s geographical map.

The student will be able to explain how interest rates affect unemployment.

Figure 2.1. Hierarchical levels of Bloom’s taxonomy (web material: Atherton, 2005)

Table 2.1 Bloom’s Taxonomy of Educational Objectives

<table>
<thead>
<tr>
<th>Sr.No.</th>
<th>Level</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>Comprehension</td>
<td>Summarize, interpret or explain material.</td>
<td>Summarize the use of every symbol on the map.</td>
</tr>
<tr>
<td>3.</td>
<td>Application</td>
<td>Use general rules and principles to solve new problems.</td>
<td>Write directions for traveling by numbered roads from any city to any other city.</td>
</tr>
<tr>
<td>4.</td>
<td>Analysis</td>
<td>Reduction of concepts into parts and show relationship of parts to whole.</td>
<td>Describe maps in terms of function and form.</td>
</tr>
<tr>
<td>5.</td>
<td>Synthesis</td>
<td>Creation of new ideas or results from existing concepts.</td>
<td>Construct a map of a hypothetical country with given characteristics.</td>
</tr>
<tr>
<td>6.</td>
<td>Evaluation</td>
<td>Judgment of value or worth.</td>
<td>The students will evaluate the usefulness of a map to enable him/her to travel from one place to another.</td>
</tr>
</tbody>
</table>

Source: Based on Bloom’s et. al (1956)
iii) Application
Objectives at application level involve the use of general rules, principles, or abstract concepts to solve a problem not previously encountered. Examples of objectives at the application level include:
- The student will be able to apply multiplication and division of double digits in applied math problems.
- The student will be able to write directions for traveling by numbered roads from any city on a map to any other city.

iv) Analysis
Objectives at analysis level require the students to reduce or breakdown a complex concept into its basic parts or elements in a manner that illustrates the relationship of parts to whole. Examples of educational objectives at this level include:
- The student will describe map in terms of functions and forms.

v) Synthesis
Objectives at synthesis level require the student to blend existing elements in such a way that they form new structure or patterns. Example of objective at synthesis level will include:
- The student will construct a map of a hypothetical country.

vi) Evaluation
Objectives at evaluation level will require from a student to make evaluative judgments regarding the quality, value, or worth of something for a stated purpose. Examples of objectives at the evaluation level include:
- The student will evaluate the usefulness of a map to enable him or her to travel from one place to another.

The model above is included because it is still being used, but Anderson and Krathwohl, 2001 have made some apparently minor but actually significant modifications, to come up with a new model (web based material). In Figure 2.2 old and new version of Bloom’s taxonomy have been taken from a web reference.

Figure 2.2: Revised taxonomy of the cognitive domain following Anderson and Krathwohl (2001). (web material: Atherton, 2005)

Note the new top category, which is about being able to create new knowledge within the domain, and the move from nouns to verbs.
2. Affective Domain

Affective Learning Outcomes
Affective learning outcomes involve attitudes, motivation, and values. The expression of these often involves statements of opinions, beliefs, or an assessment of worth (Smith & Ragan, 1999). The Affective domain has received less attention, and is less intuitive than the Cognitive. It is concerned with values, or more precisely perhaps with perception of value issues, and ranges from mere awareness (Receiving) to being able to distinguish implicit values through analysis.

The affective domain involves characteristics such as values, attitudes, interests and behavioral actions. As a result, affective objectives involve the attitudes and actions of students in relation to a school subject. For example:

- The student will demonstrate interest in earth science by conducting a science fair project in some area of earth science.

As a general rule, affective objectives are emphasized more at elementary school curricula than secondary curricula (Reynolds, Livingston, & Willson, 2006). Taxonomy of affective objectives developed by Krathwohl, Bloom, and Masia (1964) is presented in table 2.2 and in figure 2.3:

Table 2.2 Krathwohl’s Taxonomy of Affective Objectives

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
<th>Sublevels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Receiving (attending)</td>
<td>Being aware of and willing to attend something e.g. instructions.</td>
<td>Awareness, willingness to attend and selective attention.</td>
</tr>
<tr>
<td>Responding</td>
<td>Actively participating in an activity or process.</td>
<td>Acquiescence, willingness and satisfaction.</td>
</tr>
<tr>
<td>Valuing</td>
<td>Assigning value or worth to an activity or idea.</td>
<td>Acceptance, preference and commitment.</td>
</tr>
<tr>
<td>Organization</td>
<td>Ideas and values become internalized and organized into one’s personal system of values and beliefs.</td>
<td>Conceptualization and organization.</td>
</tr>
<tr>
<td>Characterization by a value or value complex</td>
<td>Individual values are exemplified in a characteristic set of behavior and actions.</td>
<td>Generalized set and characterization.</td>
</tr>
</tbody>
</table>

Source: Based on Krawthwohl et al. (1964)
This taxonomy involves levels of increasing sophistication with each level building on preceding levels. It depicts a process whereby new ideas, values, and beliefs are gradually accepted and internalized as one’s own. In schools affective objectives are almost always adjuncts to cognitive objectives. For example, we want our students to learn science and as a result to appreciate or enjoy it. Classroom tests predominantly focus on cognitive objectives, but affective objectives are found in school curricula, either explicitly or implicitly. Because affective objectives appear in school curriculum, their specification enhances the chance of being achieved.

Affective objectives vary from simple attention to selected phenomena to complex but internally consistent qualities of character and conscience. We found a large number of such objectives in the literature expressed as interests, attitudes, appreciations, values, and emotional sets or biases. (McMillan, 2007).

Here are descriptions of each step in the taxonomy, starting at the most basic level.

i) **Receiving** is being aware of or sensitive to the existence of certain ideas, material, or phenomena and being willing to tolerate them. Examples include: to differentiate, to accept, to listen (for), to respond to, etc.

ii) **Responding** is committed in some small measure to the ideas, materials, or phenomena involved by actively responding to them. Examples are: to comply with, to follow, to commend, to volunteer, to spend leisure time in, to acclaim, etc.

iii) **Valuing** is willing to be perceived by others as valuing certain ideas, materials, or phenomena. Examples include: to increase measured proficiency in, to relinquish, to subsidize, to support, to debate, etc.

iv) **Organization** is to relate the value to those already held and bring it into a harmonious and internally consistent philosophy. Examples are: to discuss, to theorize, to formulate, to balance, to examine, etc.
v) **Characterization** by value or value set is to act consistently in accordance with the values he or she has internalized. Examples include: to revise, to require, to be rated high in the value, to avoid, to resist, to manage, to resolve, etc.

### 2.6 What is the relevance of the affective domain in education?

Affective topics in educational literature include attitudes, motivation, communication styles, classroom management styles, learning styles, use of technology in the classroom and nonverbal communication. It is also important not to turn students off by subtle actions or communications that go straight to the affective domain and prevent students from becoming engaged.

In the educational literature, nearly every author introduces their paper by stating that the affective domain is essential for learning, but it is the least studied, most often overlooked, the most nebulous and the hardest to evaluate of Bloom's three domains. In formal classroom teaching, the majority of the teacher's efforts typically go into the cognitive aspects of the teaching and learning and most of the classroom time is designed for cognitive outcomes. Similarly, evaluating cognitive learning is straightforward but assessing affective outcomes is difficult. Thus, there is significant value in realizing the potential to increase student learning by tapping into the affective domain. Similarly, students may experience affective roadblocks to learning that can neither be recognized nor solved when using a purely cognitive approach.

*Read Article 2.5 in folder Unit 2 by Atherton (2007) on ‘Bloom Taxonomy’*

### 3. Psychomotor Domain

Psychomotor objectives focus on physical and kinesthetic skills (including keyboarding, using technical instruments and other skills). This domain is characterized by progressive levels of behaviors from observation to mastery of a physical skill. The importance and value of physical skills to the individual and to society in general appears to be maintained and indeed increased. The need to master new physical skills has become a part of lifelong learning for a greater proportion of the population. Dave (1975) developed five levels for measuring psychomotor objectives. Figure 2.4 shows the psychomotor levels, their description and examples given by Dave.
Romiszowski (1999) reported that value of instruction in this area lies in the value of the skills and competencies achieved by the learner, as perceived by the learners themselves, their employers, and society in general. The hierarchical levels of psychomotor domains are given in Table 2.3.

Table 2.3 Psychomotor Domain (Romiszowski, 1999)

<table>
<thead>
<tr>
<th>Sr.#</th>
<th>Level</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Observing</td>
<td>Active mental attention to a physical event.</td>
<td>The learner observes a more experienced person in his/her performance of a skill. Asked to observe sequences and relationships and to pay particular attention to the finished product. Direct observation may be supplemented by reading or watching a video. Thus, the learner may read about the topic and then watch a performance.</td>
</tr>
<tr>
<td>2</td>
<td>Imitating</td>
<td>Attempted copying of a physical behavior.</td>
<td>The learner begins to acquire the rudiments of the skill. The learner follows directions and sequences under close supervision. The total act is not important, nor is timing or coordination emphasized. The learner is conscious of deliberate effort to imitate the model.</td>
</tr>
<tr>
<td>3</td>
<td>Practicing</td>
<td>Trying a specific physical act</td>
<td>The entire sequence is performed repeatedly. All aspects of the act are performed in a sequence. Conscious effort fades as the performance becomes more or less habitual. Timing and coordination are emphasized. Here the person has acquired the skill but is not an expert.</td>
</tr>
<tr>
<td>4</td>
<td>Adapting</td>
<td>Fine tuning. Making minor adjustments in the physical activity in order to perfect it.</td>
<td>Perfection of the skill. Minor adjustments are made that influence the total performance. Coaching is often very valuable here. This is how a good player becomes a better player.</td>
</tr>
</tbody>
</table>
Simpson (1972) developed hierarchical system of psychomotor domain. This classification is given in figure 2.5.

![Hierarchical Model of Psychomotor Domain](image)

Figure 2.5: Hierarchical Model of Psychomotor Domain (Simpson, 1972; Reference McMillan, 2007).

McMillan spreads out the hierarchies presented by Simpson. From the most basic level to higher level of origination the author has described each level in terms of activities that have to be assessed.

Note: You will read this topic in detail in Chapter 8 of McMillan book, and Chapter 9 of Gronlund book)

Please see again articles 2.4 & 2.5 in Folder Unit -2 on CD.

2.7. Combining Objectives and Types of Knowledge

Bloom’s cognitive taxonomy presents six levels of abstraction for stating objectives. The range of these six different cognitive processes (verbs) encompasses three different types of content knowledge (nouns). These three types predominate in teacher’s objectives, instruction, and assessments. They are;

1. Factual Knowledge
2. Conceptual Knowledge
3. Procedural Knowledge

Table 2.4 summarizes the characteristics of three types of content knowledge, related nouns and their general description.
Table: 2.4 Types of Knowledge

<table>
<thead>
<tr>
<th>Sr.No.</th>
<th>Type of knowledge</th>
<th>Related Nouns</th>
<th>General Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Factual</td>
<td>Terminology, names, Facts, definitions, dates</td>
<td>Knowledge of specific facts</td>
</tr>
<tr>
<td>2.</td>
<td>Conceptual</td>
<td>Classifications, categories, Generalizations, theories, principles</td>
<td>Knowledge of concepts and categories</td>
</tr>
<tr>
<td>3.</td>
<td>Procedural</td>
<td>Methods, criteria, recipes, Algorithms, strategies, Steps in a process</td>
<td>Knowledge of how to apply methods and procedures</td>
</tr>
</tbody>
</table>

Examining the table you will notice that there are three ways to show main ideas. For example an English teacher expects that:

‘Pupil can state the main idea of a short story.’

The process (verb) in this objective is ‘state’ and the content (noun) is ‘main idea’. Now there are multiple ways to learn a main idea, e.g.

1. To learn main idea by memorizing the teacher’s statement of it during instruction.
2. Pupils have to infer it from on the basis of interpretation of the key concepts of the story.
3. Pupil can learn a main idea by following the steps the teacher has taught them during instruction.

Because there are different ways to learn main idea, there should be a different objective for each way. The statements 1, 2 and 3 require the objectives to be stated:

1. The pupils can remember the main idea of a short story.
2. The pupil can understand the concept of a main idea in a short story.
3. The pupils can apply a procedure to find the main idea of a short story.

It is important for teachers to recognize that which one of three objectives has to be assessed. Teacher should think about and include information about the knowledge dimensions of objectives for the following reasons:

a) Pupils are expected to master different types of knowledge and these types should be made explicit when teachers formulate objectives.

b) Different types of knowledge require different types of instruction.

c) Different types of knowledge require different types of assessment.

For aligning lesson plans and instruction clarity of the objectives is essential. Please read this topic in detail in the following supporting material:


Read Article 2.6 Cown (2006) on ‘Use of objectives to improve learning’ in web material Folder Unit- 2 on CD
2.8 Other Sources of Objectives

1. National Objectives
The aims and goals set by the legislatures while formulating Educational Policies of a country provide a frame work to administrators, teachers and all personnel who are attached with education department. National level objectives of education are developed on the bases of national philosophy, religious and sociological aspect. Global perspective in this regard gives direction for maintaining the equilibrium among all aspects.

2. Textbook Objectives
The modern textbooks are accompanied with a list of learning objectives, information to help teachers to plan, deliver and assess their instructions. The objectives and other resources can be very useful for classroom teacher. Teacher should assess textbook objectives for their usefulness on the following bases:
1. Objectives are stated in observable and measurable terms.
2. Objectives and text material are clearly stated. Especially the clarity of process and knowledge.
3. The objectives are aligned with the text material. The objectives include most learning outcomes to the topic presented. It requires the comprehensiveness of the objectives.
4. Both higher-and lower-level thinking behavior is included.
5. Objectives are suitable for pupil of this class and age.
6. Pupils have the prerequisite skills needed to master objectives.
7. Appropriate time required to master the objectives is provided.

2.9 Aligning Objectives and Instruction

Teachers use different instructional approaches when planning instruction. Each approach involves a set of steps or a variety of strategies. Before teaching in any instructional situation, it is important to know what facts, concepts, and procedures are needed for pupils to learn something new. It is critical to determine what pupils know and do not know about what they are to learn so that the teacher can start instruction at level appropriate to the pupil’s readiness. If pupils do not remember needed factual knowledge, it may be fruitless to begin instruction with conceptual knowledge. Similarly, trying to teach procedural knowledge will be difficult if the conceptual knowledge that underlines a procedure is not available to the students. In most cases teachers develop their own instructional objectives based on the textbook objectives and aims.

Teachers may select instructional objectives from different sources; these sources should be in line with school goals and national goals. The teacher after preparing the list of content must evaluate the final list of objectives. For evaluating the final list of objectives Gronlund (2006) has provided a list of questions to help the teachers while reviewing the selected objectives (chapter 1).
(i) Also read McMillan’s book, Chapter 2
(ii) La Marca(2001) discussed the concept of alignment and the role it plays in assessment and accountability systems in article 2.7 titled: ‘Alignment of standards and Assessment as an accountability criterion’ in Folder Unit-2 on CD.

2.10 Stating Behavioral Objectives

Three common steps are used by teachers for stating the instructional objectives:
1. Identify the process (verb) that best represents the activity pupils needed to perform.
2. Identify the type of knowledge (noun) that describes the information that pupils will work with.
3. Check for the compatibility of the process and knowledge.

A well stated objective provides a description of intended outcome or the type of needed product. Gronlund (2006) states:

In stating instructional objectives it is helpful to keep in mind that we are not describing the teaching procedures, the instructional activities, or the learning process we are simply describing the student performance to be demonstrated at the end of the learning experience as evidence of learning (p. 46).

Following material in Folder Unit-2 on CD can be helpful in stating effective behavioral objectives:

Article 2.8 by Kizlik (2008) on ‘Examples of Behavioral Verbs’
Article 2.9 ‘Applying Bloom’s Taxonomy’

2.11 Self-Assessment Questions

1. Is it necessary to write down one’s objectives? Even if you are an experienced teacher.
2. Explain with examples how educational objectives can differ in terms of scope?
3. What domains are covered by educational objectives? Give examples.
4. Describe Bloom’s Taxonomy of cognitive objectives? What developments have been made in the revised taxonomy?
5. After carefully studying this unit, allied material and web material, write behavioral objectives for a lesson of English textbook for class 10.
6. Keeping in mind the types of knowledge discuss the importance of aligning objectives and instruction.

2.12 Suggested Readings

2. 13 Web Based Material on CD


Unit–3

MEASUREMENT: METHODS AND TECHNIQUES–I
CONTENTS

3.1 Introduction of the Unit ................................................................. 41
3.2 Objectives of the Unit ................................................................. 41
3.3 Types of Objective Test Items ..................................................... 42
3.4 Specifying the Instructional Objectives ....................................... 43
3.5 Preparing Test Specifications ..................................................... 43
3.6 Writing Test Items (General Tips) ............................................... 44
3.7 Writing Specific Types of Items .................................................. 45
3.8 Preparing Test Directions ........................................................... 47
3.9 Correction for Guessing .............................................................. 47
3.10 Self Assessment Questions ......................................................... 48
3.11 Suggested Readings ................................................................. 49
3.12 Web Based References and Related Web Sites on CD .................. 49
3.1 Introduction of the Unit

Teacher made tests are the major basis for evaluating students’ progress in school. It is an established fact that tests play an important role in the life of teacher and student. There are different types of tests, which are used according to their purposes and objectives. There are two types of tests accordingly, one is subjective type test and the other is objective type test. Objective type tests are responded by the examinee to the point answer. Keeping in view the behavioral objectives, we construct objective type items. Objective type items are the best achiever of the instructional objective in the classroom. For construction of objective type items we have to keep in mind the instructional objectives and course content. There are also certain guidelines to write the objective type test items. Table of specifications help a teacher to construct items containing cognitive, affective and psychomotor domains. To achieve these domains we develop objective type items. These are fruitful to know about students’ knowledge, behavior and skills. Good tests require appropriate planning so that instructional objectives, curriculum, content, and instructional material are related in meaningful way. Therefore, teachers are expected to write their own classroom tests. Most of these tests are subject related and focus on specific domain of learning and assess whether a subject has been understood. This unit and related objective reading material on CD and in McMillan book describe some basic concepts about types and nature of an objective test items, specifying instructional objectives, preparing table of specification, writing general and specific test items (multiple-choice, binary choice, matching, etc), and preparing test direction.

3.2 Objectives

After reading this unit you should be able to:
1. Explain different types of objective tests.
2. Describe the major types of selected-response items and their characteristics.
3. Illustrate a thorough understanding of the principles for developing a table of specifications by developing one for a specified content area.
4. Describe the components and types of multiple-choice items and give examples.
5. Describe the principles involved with developing effective multiple-choice items.
6. Discuss the strengths and weaknesses of multiple-choice items.
7. Describe the principles involved in writing binary-choice items.
8. Develop effective binary-choice items for a given content area.
9. Discuss the strengths and weaknesses of binary-choice items.
10. Describe the principles involved with developing effective matching items.
11. Develop matching items for a given content area.
12. Discuss the strength and weaknesses of matching items.
13. Be able to write guide lines/instructions for an objective type test.
14. Be able to apply and interpret a correction for guessing.
3.3 Types of Objective Test Items

Objective tests demand hard work and thinking for their proper construction throughout the course. In regard to the response (answer) these are categorized as;

i. Free response tests

ii. Limited response tests

a. Free response tests
   These tests need only one word or short statement to complete. No option is provided in these tests

b. Limited response tests
   These tests are categorized as:
   - Fill in the blanks
   - Alternate/Binary choice response items
   - Multiple choice items
   - Matching items

Fill in the blanks
   In this type of test a statement is written but an important word or words are left blank. The students are told to write appropriate word or words in the provided space.

Alternate/Binary choice response items
   It carries only two options to be responding, therefore, this test is also called two options multiple choice test. Two options are provided like:
   - True or False
   - Right or Wrong
   - Correct or Incorrect
   - Yes or No
   - Same or Opposite

Multiple choice items
   In this type of test every item carries two or more than two options like four or five. Only one option is true or more appropriate than the others. Students are instructed to tick or encircle the best or more appropriate option than others.

Matching items
   Usually this kind of test needs three columns. First two columns are filled with equal number of options but options of both columns have relation with any aspect. Some times one column is given more options than the other to reduce the chance or guess factor. Students are told to write the correct option in the third column from the second column according to the first column.
3.4 Specifying the Instructional Objectives

Teacher made test may be more useful in the classroom in accordance to the instructional objectives. One can administer every item with specification to the instructional objectives. So, in order to measure the achievement of instructional objectives it is more appropriate to construct objective type items. In this way intended learning outcomes can be easily related to the items. Objective type items help the teacher to relate them with the specifications of the instructional objectives. In this way a large amount of content can be covered. With the help of table of specifications each item can be selected from the course content and also keeping in view the intended learning outcomes. Students performance can be evaluated like:

Knowledge Outcomes
What is the name of .................................?
What is the location of .................................?
What are the characteristics of ..........................?
What is the function of .................................?

Understanding Outcomes
What is the reason for .................................?
What is the relationship between ...........................
Which of these is an example of ..........................
Which of these best summarizes ..........................

Application Outcomes
What method would be best for ..........................
What steps should be followed to construct ...............
Which of these indicates correct application of ............
Which of these solutions is correct form .................

This makes the user easy to construct an item where the essay type test is not feasible because it specifies according to the instructional objectives.

3.5 Preparing Test Specifications

Test specifications relate to instructional objectives and to the course content and specify that relative emphasis should be given to each type of learning outcomes. Table of specifications possesses:

- Obtaining a list of instructional objective
- Outlining the course content
- Preparation of two-way chart

1. Obtaining a List of Instructional Objective

Every unit contains its instructional objectives which are supposed to be achieved after teaching of that unit. These are formulated in such a way that they are achieved in
the paper pencil test. So, while planning an objective test these specific learning outcomes and instructional objectives should be kept in mind.

2. **Outlining the Course Contents**
   
   Course content is the area of performance of the students to be displayed. Actually this is derived from the instructional objectives. Course content may be a series of topics and subtopics of that specific unit in a sequence. The amount of content in each topic depends upon the test also.

3. **Preparation of Two-Way Chart (Table of Specification)**
   
   This step includes the formulation of two-way chart in regard to table of specifications. We relate instructional objectives to the course content and this specifies the nature of test sample.

   Two-Way chart is of this type:

<table>
<thead>
<tr>
<th>Content Area</th>
<th>Knows Common Terms</th>
<th>Knows Specific Terms</th>
<th>Understands principles &amp; Generalizations</th>
<th>Applies Principles &amp; Generalizations</th>
<th>Interprets Charts &amp; Graphics</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>Clothing</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>Transportation</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Communication</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Shelter</td>
<td>-</td>
<td>-</td>
<td>5</td>
<td>5</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>City life</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>-</td>
<td>20</td>
</tr>
<tr>
<td>Farm life</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>-</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>25</td>
<td>10</td>
<td>100</td>
</tr>
</tbody>
</table>

_Read Article 3.1 by Chase(1999) on ‘Table of Specification’ in Folder Unit-3 on CD._

**Activity 3.1**

Take three chapter from 10th class General Science or Pak Studies book and prepare a two way table of specification.

3.6 **Writing Test Items (General Tips)**

i. Express items as precisely, clearly and simply as possible.
ii. Include all qualifications necessary to provide a reasonable basis for responding.
iii. Emphasize general tasks rather than small details.
iv. Avoid jargon and textbook language.
v. Locate and delete irrelevant clues.
vi. Eliminate irrelevant sources of difficulty.
vii. Place all items of a given type together in the test.
viii. Prepare keys of model answers in advance of test administration.
ix. Arrange for competent review of the items.
3.7 Writing Specific Types of Items

1. Multiple Choice Items
   These are the most popular objective type of test items. The basic form of multiple choice items is a stem or lead, which defines the problem to be completed by one of a number of alternatives or choice. There should be only one correct response and the other alternatives/options should be plausible but incorrect. For this reason the incorrect alternatives are sometimes referred to as ‘Distracters’. In most cases four or five alternatives are given. The use of plausible distracters help the teacher to control the difficulty of the test. They should not be tricky or trivial. The idea in writing the question is to have the knowledgeable students choose the correct answer and not be distracted by the other alternatives serve to distract the less knowledgeable. The effect of guessing is reduced by increasing the number of alternatives. The major limitation of multiple choice format is that the distracters are often to construct, particularly as the number of choices increases.

   Below are suggestions for writing multiple choice items:

   i. State the central idea of the problem in the stem. It should be a singular statement, topic, or problem.
   ii. Include one correct or most defensible answer.
   iii. Select diagnostic foils or distracters.
   iv. Option should be presented in a logical, systematic order.
   v. Option should be grammatically parallel and consistent with the stem.
   vi. Option will be mutually exclusive.
   vii. Ensure that correct responses are not consistently shorter or longer than the foils.
   viii. Eliminate grammatical or verbal clues.
   ix. Present the problem in novel terms.
   x. Use negatively stated items infrequently.
   xi. Beware of ‘none of these’, ‘none of the above’, ‘all of these’, and ‘all of the above’ phrases.
   xii. Alter item difficulty by making options more alike or less alike in meaning.
   xiii. Correct response should be in random order. Do not use particular letter more often than others or create a pattern for the placement of correct response.
   xiv. Avoid using items directly from the text or workbook, since this practice encourages memorization.

   Read following articles in folder Unit-3 on CD:
   Article 3.2 by Kehoe (1995) on ‘Writing Multiple-Choice items’
   Article 3.3 by McKenna & Bull (1999) on ‘Designing effective objective test questions’

2. Alternate/Binary Choice Items
   The main advantage of true-false items is their ease of construction and ease of scoring. A teacher can cover a large content area and a large number of items can be presented in
a prescribed time period. This allows the teacher to obtain a good estimate of the student’s knowledge. If the items are carefully constructed, they can also be used to test understanding of principles. Guessing is the biggest disadvantage to true-false test. When students guess, they have a 50-50 chance of being right. The purpose of such test is to measure what students know not how lucky or clever they are. The followings are the general principles for constructing binary choice items:

i. Include only one idea in each item. Each true-false item should test an important concept or piece of information. The knowledge being tested should be significant.

ii. Place the idea being tested at the end of the statement. Most students focus more attention at the last portion of the statement.

iii. Avoid unfamiliar language and wordiness since it confuses the students.

iv. Avoid verbatim textbook statements, use of such statements encourages memorization.

v. Avoid partly true-partly false items.

vi. Avoid specific determiners.

vii. Ensure that true false items are approximately equal in length.

viii. Balance the number of true items and false items.

ix. Eliminate vague terms of degree or amount.

x. Use caution in writing negative item statements and double negative statements.

xi. Avoid qualifying statements and words that involve judgment and interpretation. For example, few, always, never, usually, etc. avoid using them in statement considered to be false.

3. Matching Items

In a matching test there are usually two columns of items. For each item in one column, the student is required to select a correct (or matching) in the other. The items may be names terms, places, phrases, questions, statements, or events. The basis for choosing must be carefully explained in the directions. Matching items have the advantages of covering a large amount and variety of content and easy to score. These are considered a modification of multiple choice items. One problem with multiple choice items is, finding homogenous test and response items that are significant in terms of objectives and learning outcomes. One of limitation is this, that matching items require recall rather than comprehension and sophisticated level of thinking. Higher level of cognition may be called far in matching questions that involve analogies, cause and effect, complex relationships, and theories.

i. The direction should be brief and clear.

ii. There should be two columns: Column A and Column B for matching items.

iii. An entire matching question should appear on a single page.

iv. Wording of items in column A should shorter than those in column B.

v. Column A should contain no more than 10 test items: 5 or 6 items are more appropriate.

vi. There should be more alternative in Column B than there are items in Column A to prevent answering the last one or two items by simple elimination. Column should contain 6 or 7 items if column A contains 5. A list of 10 items in column A should be accompanied by about 12 items in column B.
vii. Include homogeneous material in each exercise in terms of content, form, grammar, and length. Otherwise the test-wise students will guess the correct matching.

viii. Eliminate irrelevant clues.

ix. Reduce the influence of clues and thereby increase matching item difficulty.

x. Compose the response list of single words or very short phrases.

xi. Arrange the responses in systematic order: Alphabetical, chronological, etc.

*For detail please read following:*

| McMillan: Chapter 6 | pp. 167-180 |

*Read the following Articles in folder Unit-3 on CD:*

i. *Article 3.2 ‘Writing Multiple-Choice Items’*

ii. *Article 3.3, ‘Writing Test Items’*

iii. *Article 3.4 ‘Designing Effective Objective Test Items’*

After reading above mentioned articles perform the following activity:

<table>
<thead>
<tr>
<th>Activity 3.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collect question papers for the last two years of your examination board, prepare a specification table and place the objective types of items according to the classification in cognitive domain.</td>
</tr>
</tbody>
</table>

### 3.8 Preparing Test Directions

Following points should be taken into consideration while preparing test directions:

1. The purpose of the test should be made clear to the students. Students need to know why they are taking the test and how the results will be used.

2. Students need to know exactly how much time they will have to complete the test. It is helpful to indicate to the students that how they should distribute their time among various parts of the test.

3. Students should be made clear about what they are to do to answer the question, i.e. how to respond.

4. Students should be made clear about how to show their answers, e.g. circle the right answer, write the word in the blank, etc.

5. Students should be made clear about the basis on which their tests, particularly constructed response items will be graded.

### 3.9 Correction for Guessing

To ascertain whether the score in the form of raw score is accurate indicator of true standing of the examinee on the test or whether it has been just examinee’s guessing, the users of test scores has developed different approaches. It is a common practice with the objective types of items that students often guess and it improves their chance of
improvement in scores. Correcting for the effect of guessing on certain standardized tests (GRE and SAT) the following most popular correction-for-guessing formula can be used:

\[
S = R - \frac{W}{k-1} \quad (3.1)
\]

In this formula:
- \(S\) is corrected score
- \(R\) is number of items that the examinee gets right
- \(W\) is number of items that the examinee gets wrong
- \(K\) is number of options

For true-false items it is simply calculated as:

\[
S = R - W
\]

On most classroom assessment a correction for guessing is not necessary. Nitco (2001) recommended that the relative ranking of students using corrected and uncorrected scores remains the same. It has been noticed that examinees after knowing that their score will be corrected that skip items. If they are sure that their score will be not corrected they attempt every item.

### 3.10 Self Assessment Questions

Q1. Encircle the best one option given below every statement

- Alternate response items contain number of options;  
  a) One  
  b) Two  
  c) Three  
  d) Four

- Matching tests has three columns and the answer is written in the---column  
  a) First  
  b) Second  
  c) Third  
  d) Fourth

- Instructional objectives are best achieved  
  a) In the classroom test  
  b) In GRE test  
  c) In the annual exams.  
  d) In the practical

- Table of specifications needs  
  a) Four way chart  
  b) Three way chart  
  c) Two way chart  
  d) One way chart

Q2. Explain different types of objective tests. Also give your Opinion which type is better than the others.

Q3. What are the instructional objectives? Give at least five examples according to their level of complexity.

Q4. What are test specifications? Develop a table of specifications keeping in view the social studies course of class 7th.

Q5. With the help of the above table of specification of Q4, construct 50 MCQs.

Q6. Write the rules of writing objective type tests; also give common rules which are used in all types of objective tests?

Q7. Compare the strength and weaknesses of multiple-choice and binary-choice test items. Give examples.
Q.8 Compare the strength and weaknesses of binary-choice and matching items. Give examples.

Q.9 Compare the strength and weaknesses of multiple-choice and matching items.

Q.10 Illustrate general principle for developing effective objective test items. Support your discussion with the help of example.

Q.11 What is guessing? How would you deal with guessing component for an objective test of 100 items.

Q.12 Below are five objectives. For each objective write two test items of the type specified in parentheses to assess the objective.

1. The student can match the symbols of chemical elements to their names. (matching)
2. The student can identify the nouns in a sentence that contains more than one noun. (multiple choice)
3. The student can indicate whether a statement about the Pakistan constitution is true or false. (True-false)
4. The student can identify the definition of adverb. (multiple choice)
5. The student can write the name of the capital of a country. (completion item)

Q.13 What are examples of clues to be avoided

3.11 Suggested Readings


Web Based Material on CD

Unit–4

MEASUREMENT: METHODS AND TECHNIQUES–II
CONTENTS

4.1 Introduction of the Unit .................................................................55
4.2 Objectives of the Unit .................................................................55
4.3 Purposes of Essay Testing ............................................................55
4.4 Kinds of Essay Tests .................................................................56
4.5 Guidelines for Writing Essay Items .............................................57
4.6 Guidelines for Writing Short-Answer Items ................................58
4.7 Scoring Essay Tests .................................................................59
4.8 Scoring Short-Answer Items .....................................................61
4.9 Advantages and Limitations of Essay Tests as Achievement Measures ..........61
4.10 Aligning Assessment and Instruction ..........................................62
4.11 Self Assessment Questions .......................................................63
4.12 Recommended Readings ............................................................63
4.13 Web Based Material on CD .......................................................63
4.1 Introduction of the Unit

An essay item is a test item that poses a question or problem for the students to respond in a written format accordingly. Being a constructed-response item, the student must respond by constructing a response, not by selecting among alternatives. Essay items can be scored in terms of content, style, and grammar. The essay items provide considerable freedom to the student in composing a response. Good essay items challenge the student to organize, analyze, integrate, and synthesize information. At their best, essay items elicit novel and creative cognitive processes from students. We will begin by discussing how essay item can be developed and can vary according to their educational purpose. The other topics discussed will be types of essay tests, guidelines for writing essay test, and guidelines for scoring essay items.

4.2 Objectives of the Unit

After reading this unit, you will be able to:
1. Explain how essay item can differ in terms of purpose and level of complexity.
2. Compare and contrast restricted-response and extended-response items.
3. Describe the principles involved with developing effective essay items.
4. Develop effective essay items for a given content area.
5. Discuss the strengths and weaknesses of essay items.
6. Describe the principles involved with grading essays.
7. Demonstrate the ability to grade essays in a reliable and valid manner.
8. Evaluate the principles involved with developing effective short-answer items.
9. Develop effective short-answer items for a given content area.
10. Analyze the strength and weaknesses of short-answer items.
11. Learn aligning assessment and instruction.

4.3 Purposes of Essay Testing

The major purposes of essay testing are the assessment of:

i. Content
ii. Style
iii. Grammar

The content element represents testing solely for cognitive achievement. When scoring for this purpose the teacher attends only on the content of the response and ignores student’s achievement in writing style and in grammar. The purpose is to determine what the student knows and can produce. Through essay items all the levels of cognitive taxonomy can be measured.

The style element has to do with writing composition. The student is asked to pick a topic and write in a specified manner. The measurement of response is based on objectives related to organization, structure, phrasing, transition and other components of the writing process.

The grammar element is one in which the purpose is to examine the students ability to apply grammatical rules. This involves all aspects of writing mechanics.
4.4 Kinds of Essay Tests

i. Restricted Response Items
ii. Extended Response Items
iii. Short-Answer Items

1. Restricted Response Items
   Restricted response items are good for assessing objectives at the knowledge, comprehension, and at the application levels. They can be answered in a timely fashion by the students, which allows the test developer to include more items and they are easier to score in a reliable manner. Restricted response items are highly structured and clearly specify the form and scope of a student’s response. Such items typically require students to list, define, describe, or give reasons. These items may specify time or length limits for the response. We can say that:
   i. The restricted response questions limit both the content and the response.
   ii. The answer of these questions requires few lines or a brief paragraph.

   EXAMPLES
   • Why is the barometer one of the most useful instruments for forecasting weather?
     Answer in a brief paragraph.
   • List the types of muscle tissue and state the function of each.

2. Extended Response Items
   Extended-response items provide more latitude and flexibility in how the students can respond to the question. There is little or no limit on the form or scope of the response. When limitations are provided they are usually held to a minimum (e.g. page and time limits). Extended- response items often require students to compose, summarize, formulate, compare/interpret, and so forth. We can say that extended response questions give freedom to the students to demonstrate their ability, to select, organize, integrate and evaluate ideas.

   EXAMPLES
   • Summarize and write a critical evaluation on the research on global warming. Include a detailed analysis of the strength and weaknesses of the empirical research.
   • Describe the influence of Mendel’s laws of heredity on the development of Biology as a science.
   • Write a scientific evaluation of Copernican theory of solar system. Include scientific observation that supports your statements.

   The above examples show that extended response items assess higher-level of cognitive objectives. However, they are difficult to score in a reliable manner and as they take considerable time for students to complete, the teacher typically have to limit the test to relatively few items. It is important to write and score extended response items as carefully as possible and take into consideration the limitations.

3. Short Answer Items
   Short-answer items are the items that require the student to supply a word, phrase, number or symbol. Such type of items can also be written in an incomplete sentence format instead of a direct question format. This type of format if sometimes referred to as
a completion item. The short-answer format presents the problem with direct question. We can say that:

i. Relative to essay items, short-answer item place stricter limits on the nature and length of the response.

ii. They tend to assess mainly factual knowledge or comprehension.

iii. Typically the student is asked to reply with a word, phrase, name or sentence, rather than with a more extended response.

EXAMPLES

i. Next to each country write the name of its capital city.
   Malaysia ________________
   Japan ________________
   Iran ________________

ii. Scientists who specialize in the study of plants are called.

   ______________________

iii. What is the membrane surrounding the nucleus called?

   ______________________

Items (ii) and (iii) are the examples of direct-question format.

4.5 Guidelines for Writing Essay Items

Working along the following guidelines can improve the quality of essay items:

1. Write in a clear straightforward manner
   This is one of the most important criteria for a good essay item which should specify the assessment task. If you want the student to list reasons, specify that you want a list. If you want a restricted response, specify that. If you want an extended response, make that clear. When appropriate indicate the point value of the item and how much time students should devote to it.

2. Consider carefully the amount of time students will need to respond to the essay items
   Allow sufficient time for students to answer the question. This is a practical recommendation that a test developer pays attention to the amount of time the student will need to complete each essay item. As a teacher we may estimate only the time necessary to write the response whereas students actually need time to collect and organize their thoughts before even starting writing process. As a rule of thumb, construct a test you think is appropriate to the available time and reduce it in length by about 25%.

3. Do not give choice to select the items to which they will respond
   Some teachers provide a number of items and ask the students to select a specified number of items to respond to. It is common practice that students are asked to choose any five questions out of eight questions to respond to. This rule promotes selected study and students refrain from learning of important concepts. As a general rule this practice is to be avoided.
4. Use more restricted-response items in place of a number of extended response items

Restricted-response items have measurement characteristics that may make them preferable over extended-response items. They are easier to score in a reliable manner and students can respond to a larger number of items in a given amount of time. This can provide a superior sampling of content domain. Several shorter questions are generally preferable to fewer longer questions for general achievement testing.

5. Limit the use of essay items to educational objectives that cannot be measured using selected-response items

While essay items are extremely popular among teachers and have their strengths, they do have limitations that will be discussed in next sections. It is recommended that you restrict the use of essay items to the measurement of objectives that cannot be measured adequately using selected-response items, For example a student’s ability to organize and present material in a written format.

6. Have a critique of the test from your colleague

An opinion should be sought about the emphases and breadth of coverage, the appropriateness or difficulty of the questions, and the adequacy of the model answers. The small amount of effort required will make test more reliable.

7. Assign weightage

Determine in advance how much weight will be given to each question or part of a question. Give this information on the test, and score accordingly.

8. Ask thoughtful questions

Ask questions that require considerable thought. Use essay questions to focus on organizing data, analysis, interpretation, formulating theories, rather than on reporting facts.

See Article 4.1 on folder Unit-4 on CD, an uiowa Bulletin on ‘Preparing and Evaluating Essay Test Questions’

4.6 Guidelines for Writing Short-Answer Items

While writing short answer items, following guidelines should be taken into consideration:
1. Use of direct-question format is preferable than the incomplete sentence format. It is more natural for the student and is often easier to respond.
2. The questions should be so worded that the response required is as brief as possible. This will facilitate objective scoring.
3. Text-book wording in phrasing should be avoided. Different phrasing will reduce the possibility of correct responses and will encourage students to think.
4. Prepare a scoring key that contains all possible answers. Though there should be only one correct answer, if it is not possible then prepare all anticipated acceptable responses, and follow them consistently.
5. **Avoid unintentional cues to the answer.** You should avoid including any inadvertent clues that might lead student to respond with thinking. Giving cues about the relative length of different answers may be avoided. The use of article ‘a’ or ‘an’ will give clues to the students that whether response starts with vowel or consonant.

6. **Make sure the blank provides enough space for the response.** Make sure that each blank provides adequate space to the students to write the response. The space for short-answer should be fixed by keeping in view the space required for longest response to a short-answer item. Therefore, leave equal space for each item irrespective of the length of the response.

7. **Make sure that there is only one correct answer.** It is important that there only be one correct response. When writing a short-answer item, ask yourself if the student can interpret it in more than one way.

### 4.7 Scoring Essay Tests

Essay tests represent the ultimate in scoring complexity because they permit each student to construct a unique and lengthy response to the question posed. This means that there is not one definite answer key that can be applied uniformly to all responses. Essay tests have been most criticized for the unreliability of scoring. Yet there are scoring procedures that markedly improve grading of essay tests. To enhance the measurement characteristics of essay tests, teachers need to concentrate on developing structured, unbiased scoring procedures. Here are few suggestions to help the scorer of essay tests in a consistent and reliable way.

1. **Develop a scoring rationale for each answer**

   Use of predetermined scoring criteria will reduce the content in-determinacy effects. A written guide that helps the scorer for constructed response items is referred to as a scoring **rubric**. This is specially required for extended response items and items at the higher level of cognitive domain. There are actually two approaches for doing this. **Analytical scoring** rubric identifies different aspects or dimensions of the response and the teacher scores each dimension separately. It is also necessary to specify the value assigned to each characteristic. With a **holistic rubric** the teacher assigns a single score based on the overall quality of the student’s response. Holistic rubrics are often less detailed than analytical rubrics. They are easier to develop and scoring usually proceeds faster. Their primary disadvantage is that they do not provide specific feedback to students about the strength and weaknesses of their responses.

2. **Steps to ensure objectivity**

   Regardless of whether a teacher uses holistic or analytical scoring, certain steps should be followed to ensure that student’s essays are scored objectively.

   i. Define what constitutes a good answer before administering an essay question.
   ii. Decide and tell pupils how hand writing, punctuation, spelling, and organization will be scored.
   iii. Grade each essay anonymously.
   iv. Score all pupils’ answers to the first question before moving to the second question.
v. Read essay answer a second time after initial scoring.
vi. Carry out post test review in order to locate faulty test items and when necessary make scoring adjustments.
vii. Avoid fatigue.

McMillan (2007) gives a detailed example of holistic and analytical scoring. These two dimensions table provide guidelines for both types of scoring (See detail in Chapter 7, pp. 221-222).

Also read following articles in Folder Unit 4 on CD:

Article 4.2 ‘Students Perspective on Rubric-Reference Assessment’,
Article 4.3 ‘Scoring Rubrics’,
Article 4.4 ‘Designing Scoring Rubrics’

<table>
<thead>
<tr>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepare an essay type test consisting of five questions and develop a question-wise key for its evaluation.</td>
</tr>
</tbody>
</table>

3. Factors Affecting Essay Tests Scoring

When scoring essay items, a multitude of factors can contribute to its reliability. These factors include:

i) Expectancy effects

These effects occur when the teacher scoring the test allows irrelevant characteristics of the students to affect scoring. This is also refers as the ‘halo effect’. If a teacher has a favorable impression about the student with a history of academic excellence, the teacher might incline to assign higher score than a student with poor academic record. These effects are not intentional or conscious, but they are often present nevertheless. Similar effects also carry over from one item to the next within a test.

ii) Hand writing, grammar, and spelling effects

It is common observation that good handwriting raises scores, and poor handwriting, misspelling, incorrect punctuation, and poor grammar reduces scores even when content is the only criteria for evaluation. Even the length of the response impacts the score. Teachers tend to give higher scores to lengthy responses, even when the content is not good to that of a shorter response.

iii) Order effects

Order effects the changes in scoring that emerge during the grading process. It is observed that essay tests scored early in the grading process receive better grades than essay tests scored latter. Research has shown that the quality of preceding responses impacts the scores assigned. It means that essay responses tend to receive higher scores when they are preceded by poor quality responses.
iv) **Fatigue effects**

The teacher’s physical and cognitive qualities are likely to degrade if essays scoring continue too long a period. The maximum period of time vary according to complexity of the responses, but reading essays more than two hours without sufficient break will likely produce fatigue effects.

### 4.8 Scoring Short-Answer Items

Some authors classify short-answer items as objective items, which suggest that they can be scored in an objective manner. Although scoring well-written short items in a reliable manner is easier than scoring extended-response essay items, there is still a significant degree of subjectivity. Three guidelines can help teachers to overcome problems of scoring supply items.

1. Prepare an answer key before scoring. Know what you are looking for in pupils’ responses before scoring.
2. Determine how factors such as spelling, grammar, and punctuation will be handled in scoring. Should points be taken off for such factors? Decide before scoring and inform students before testing.
3. If students’ responses are technically correct but not initially considered in the scoring guideline, give credit to each unexpected but correct response.

*For further reading and related topics: e.g. Assessing Interpretive items Assessing deep understanding and Reasoning, Study McMillan (2007), Chapter 7, pp. 193-227.
Also see article 4.1 ‘Preparing and Evaluating Essay Test Questions’ in Web Material Folder Unit-4 on CD.*

### 4.9 Advantages and Limitations of Essay Tests as Achievement Measures

The advantages and limitations of subjective tests can be classified in the following five main components:

1. **Sampling**
   In essay test questions the test developer is limited to the number of questions or topics, to which the test taker can respond to, from a limited content area. When the test universe covers a wide range of topics, content validity for the test usually suffers. It generally takes less time to write essay items. Because most essay test items contain only a fraction of the number of items that an objective test might, teacher have usually fewer items to write.

2. **Construction**
   Subjective test requires fewer items and those are easier to construct and revise. Measurement experts suggest that essay questions are better suited for testing higher skills such as creativity and organization. It has been argued that higher-level cognitive skills can be assessed by multiple-choice items. Nevertheless, some educational objectives are most easily measured with essay items, and these tend to be higher-level objectives. Some objectives such as written skills literally require the use of essay items.
Essay items also have the requiring recall, often denoting stronger mastery of the material than recognition of selected -response items.

3. **Scoring**
   Scoring subjective items require time-consuming judgments by an expert. But it can be more reliable and accurate when there are two independent scorers who have been trained to avoid biased judgments. Scoring keys are important for subjective items, but these are more difficult to construct because they must address as many likely responses as possible.

4. **Guessing**
   The use of essay items largely eliminates blind guessing. Because essay items require the student to produce a response as apposed to simply selecting one, students are not able to guess successfully the desired answer.

5. **Studying essay tests**
   Many teachers believe that students study differently for essay tests than they do for selected response tests. It is possible that students preparing for essay tests spend more time analyzing and synthesizing information rather memorizing facts. Therefore, it is suggested that teachers may combine a few essay items with selected-response items to achieve this potential instructional benefit.

4.10 **Aligning Assessment and Instruction**

   Testing can result in a better teaching, when the aspect of alignment of assessment and instruction is taken into consideration. Meaningful testing is a part of an instructional effort in classroom-based instruction and assessment. In unit 2 we discussed the importance of aligning objectives and instruction and now the alignment of assessment and instruction is described. In fact, the nature of objectives should be manifested in the instruction and both objective and instruction should be manifested in assessment.

1. **Factual Knowledge**
   The main aspect of factual knowledge is remembering different kinds of information such as dates, names, events, etc. Instructional activities related to factual knowledge include direct instruction, drill and practice, repetition strategies, etc. When instruction focuses on remembering, assessment should reflect strategies that assess remembered information.

2. **Conceptual Knowledge**
   Conceptual knowledge focuses on building connection between new and existing knowledge of concepts and categories. Instructional activities related to conceptual knowledge include differentiating between examples and non-examples of concepts, using diagrams, questioning, etc. Conceptual knowledge demands more than rote memorization, therefore, the items used to assess this type of knowledge should require students to do more than just remember when answering the question. To construct such items, the teacher should choose examples that are similar but not identical to what was taught.
3. **Procedural Knowledge**

Procedural knowledge requires carrying out the steps in tasks or problems, for example, writing a story, finding places on a map, drawing pie chart, driving a car, etc. Instructional strategies for this kind of knowledge include flow and sequencing charts, demonstrations, repeated practice and teacher modeling. Assessment of procedural knowledge can involve products such as a completed story or drawing of a pie chart; and processes such as the steps for finding places on a map, steps for writing a story, etc. Teacher may assess product or process according to their instructional and assessment needs.

*Nassif discussed the importance of aligning assessment and instruction for teachers’ assessment in article 4.5 ‘Aligning Assessment and Instruction in Folder 4 on CD.*

**Self-Assessment Questions**

1. What type of educational objectives can best be measured by different types of essay tests? Explain with examples.
2. What are the major considerations to be kept in view while developing an essay type test?
3. Provide a comparison between the major principles for developing restricted response and extended response items.
4. Prepare at least 20 short-answer items for 8th class for measuring the factual knowledge of general science.
5. Compare the advantages and limitations of short- answer and essay tests.
6. Compare the advantages and limitations of essay tests with those of objective tests?
7. How can bias and subjectivity, in evaluating essay tests, can be minimized?
8. Keeping in view the textbook of social studies for 7th class, write 5 extended response questions, 10 restricted response questions and 10 short answer items.

4.12 **Recommended Readings**


4.13 **Web based Material on CD**


APPRAISING THE ACHIEVEMENT TESTS
# CONTENTS

5.1 Introduction of the Unit .................................................................69
5.2 Objectives of the Unit .................................................................69
5.3 Analyzing the Effectiveness of Test Items ......................................69
   1  Item Difficulty Index (or Item Difficulty Level) .........................70
   2  Item Discrimination ....................................................................71
   3  Item Analysis of Speed Tests .....................................................74
   4  Distracter Analysis .....................................................................75
5.4 Application of Item Analysis for Teachers ......................................76
5.5 Using Item Analysis to Improve Items ..........................................78
5.6 Qualitative Item Analysis ............................................................79
5.7 Using Item Analysis to Improve Classroom Instruction ...............79
5.8 Self Assessment Questions ............................................................80
5.9 Suggested Readings .....................................................................81
5.10 Web Based References and Related Web Sites ..............................81
5.1 Introduction of the Unit

A number of quantitative procedures are useful in assessing the quality and measurement characteristics of the individual items that make up tests. Collectively these procedures are referred to as item analysis statistics or procedures. Item analysis statistics are useful in helping test developers, including both professional psychometricians and classroom teachers, decide which items to keep on a test, which to modify, and which to eliminate. In addition to helping test developers improve tests by improving the individual items, they also can provide valuable information regarding the effectiveness of instruction or training. Although quantitative procedures for evaluating the quality of test items will be the focus of this unit, some qualitative procedures may prove useful when evaluating the quality of test items. These qualitative procedures typically involved an evaluation of validity evidence based on the content of the test and an examination of individual items to ensure that they are technically accurate and clearly stated. Although these qualitative procedures have not received as much attention as their quantitative counterparts, it is often beneficial to use a combination of quantitative and qualitative procedures. In this unit you will learn purpose of performing item analysis, calculation of item difficulty level and item discrimination, purpose and calculation of item correlation, distracter analysis, use of qualitative analysis, and use of item analysis information for improving instructions.

5.2 Objectives of the Unit

After studying this unit you will be able to:
1. Explain the importance of the item difficulty index and demonstrate its calculations and interpretation.
2. Describe the importance of the item discrimination and demonstrate its calculation and interpretation.
3. Analyze the relationship between item difficulty and discrimination.
4. Evaluate how item-total correlation can be used to examine item discrimination.
5. Describe how the calculation of item discrimination can be modified for mastery tests.
6. Describe the importance of distracter analysis and demonstrate its calculation and interpretation.
7. Describe how the selection of distracters influences item difficulty and discrimination.
8. Apply practical strategies for item analysis to classroom tests.
9. Apply the item analysis statistics to improve test items.
10. Describe how item analysis procedures can be applied to performance assessments.
11. Use qualitative approaches to improving test items.
12. Explain how information from item analyses can be used to improve instruction.

5.3 Analyzing the Effectiveness of Test Items

Different types of items and different types of tests require different types of item analysis procedures. Items that are scored dichotomously (e.g., either right or wrong) are handled differently than items that are scored on a continuum (e.g., an essay that can receive
scores ranging from 0 to 10). Tests that are designed to maximize the variability of scores (e.g., norm-referenced) are handled differently than mastery tests (i.e., scored pass or fail). Whether a test is standardized or teacher made, ability or personality, a post hoc analysis of results is very important. The item analysis provides answers to questions: Were the time limits adequate? Were the directions understandable? Were the testing conditions appropriate? Were the problems emerged during testing handled adequately? The major aim of item analysis is to help to improve the test by revising or discarding ineffective items. Moreover, to provide diagnostic information on what examinee knows and don’t know. In preceding sections various specific item analysis procedures have been discussed.

5.3.1 Item Difficulty Index (or Item Difficulty Level)

When evaluating items on ability tests, an important consideration is the difficulty level of the items. Item difficulty is defined as the percentage or proportion of test takers who correctly answer the item. The item difficulty level or index is abbreviated as \( p \) and calculated with the following formula:

\[
p = \frac{\text{Number of Examinees Correctly Answering the Item}}{\text{Total Number of Examinees}}
\]

For example, in a class of 50 students, if 30 students get the answer correct and 20 are incorrect, the item difficulty index is 0.60. The calculations are illustrated here.

\[
p = \frac{30}{50} = 0.60
\]

In the same class, if 10 students get the answer correct and 40 are incorrect, the item difficulty index will be 0.20. The item difficulty index has a range of range 0.0 to 1.0. The easy items in the test will show higher decimal values and difficult items will give lower decimal values. An item that is answered correctly by all students give an item difficult of 1.0 whereas an item that is answered incorrectly by all students receives an item difficulty of 0.0.

The optimal item difficulty level (\( p \) value) for an item depends on a number of factors, including the purposes of the test and the number of response options. A test designed to measure ability the optimum \( p \) value is closer to 0.50; indicating that 50% of the test taker answered the item correctly and 50% answered the item incorrectly or not attempted. Based on this statement, you might conclude that it is desirable for all test items to have a difficulty level of 0.50, but this is not necessarily true for several reasons. One reason is that items on a test are often correlated with each other, which means the measurement process may be confounded if all the items have \( p \) values of 0.50. As a result it is often desirable to select some items with \( p \) values below 0.50. Aiken (2000) recommends that there should be approximately a 0.20 range of these \( p \) values around the optimal value. For example, a test developer might select items with difficulty levels ranging from 0.40 to 0.60, with a mean of 0.50.

Another reason why 0.50 is not the optimal difficulty level for every testing situation involves the influence of guessing. On constructed response items (e.g., essay and short answer items), for which guessing is not a major concern, 0.50 is typically considered the optimal difficulty level. However, with selected response items (e.g., multiple choice and true false items) for which test takers might answer the item correctly simply by guessing the optimal difficulty level varies.
Table 5.1: Optimal $p$ values for items with varying numbers of choice

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Number of choices</th>
<th>Optimal Mean $p$ Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>(e.g. True and False)</td>
<td>0.85</td>
</tr>
<tr>
<td>2.</td>
<td>(three choices)</td>
<td>0.77</td>
</tr>
<tr>
<td>3.</td>
<td>(four choices)</td>
<td>0.74</td>
</tr>
<tr>
<td>4.</td>
<td>(five choices)</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Constructed Response (e.g. Essay)</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Adapted from Aiken (2003, p. 66)

A test which covers a broad range of content several very easy and difficult items are included. Though, it involves guessing the answer instead of knowledge level. To take into consideration the effects of guessing, the optimal item difficulty level is set higher than for constructed response items. For example, for multiple choice items with four options the average $p$ should be approximately 0.74. That is, the test developer might select items with difficulty levels ranging from 0.64 to 0.84 with mean of approximately 0.74. Table 5.1 provides information on the optimal mean $P$ value for selected response items with varying numbers of alternatives or choices.

See following articles in folder Web Material Unit-5 on CD:

Article 5.1 by Matlock-Hetzel (1999) on ‘Basic Concepts in Item and test analysis’

Article 5.2 web reference ‘Item Analysis’ (and solve the given problems)

Application of Item Difficulty

The item difficulty is most applicable to norm referenced tests. For mastery learning, which are criterion-referenced tests, different procedures are used for evaluating the test. The examinee has to pass or fail the test in mastery tests. Therefore the average $p$ values are as high as 0.90. Other tests that are designed for special assessment purposes may vary in terms of what represents desirable item difficulty levels. For example, the purpose of a test is to distinguish between the highest performing examinees, e.g., in testing gifted and talented students, it may also be desirable to include at least some very difficult items. In summary although a mean $p$ of 0.50 is optimal for maximizing variability among test takers, different difficulty levels are desirable in many testing applications.

5.3.2 Item Discrimination

The item-discrimination index (D) of effectiveness of an item is discriminating between high and low scorers on a test. The higher the value of D, the more effective the item is in discriminating between those with high scores and those with low scores on the test. In simple words we can say that it is a measure of Item discrimination to know how well an item can discriminate or differentiate among students who differ on the construct being measured by the test. For example if a test is designed to measure reading comprehension, item discrimination reflects an item’s ability to distinguish between individuals with good reading comprehension skills and those with poor reading skills.
i) **Discrimination Index**

The most popular method of calculating an index of item discrimination is based on the difference in performance between two groups. There are different ways of selecting the two groups. Usually 27% top and bottom 27% examinees in terms of their overall performance on the test are taken for performing the analysis. Whether you choose top and bottom 25%, top and bottom 33% or 50% top and 50% bottom you will obtain the same results. In practice all of these are probably acceptable. The difficulty of the item is computed for each group separately, and these are labeled $P_t$ and $P_b$ ($t$ for top and $b$ for bottom). The difference between $P_t$ and $P_b$ is the discrimination index, designated as $D$ and is calculated with the following formula (Kline, 2005):

$$D = P_t - P_b,$$

Where

- $D = \text{Discrimination index}$
- $P_t = \text{Proportion of examinees in the top group getting the item correct}$
- $P_b = \text{Proportion of examinees in the bottom group getting the item correct}$

The purpose of calculating this index is to design a classroom test to measure the academic achievement in some specific area. If the item is discriminating between students who know the material and those who do not, then students who are more knowledgeable (i.e., students in the top group) should get the item correct more often than students who are less knowledgeable (i.e., students in the bottom group). For example, if $p_t = 0.70$ (70% of the students in the top group answered the item correctly) and $p_b = 0.20$ (20% of the students in the bottom group answered the item correctly), then:

$$D = 0.70 - 0.30 = 0.40$$

Hopkins (1998) provided guidelines for evaluating items in terms of their $D$ values (see Table 5.2). He suggested that $D$ values of 0.40 and above are considered excellent, between 0.30 and 0.39 are good, between 0.11 and 0.29 are fair and between 0.00 and 0.10 are poor. Items with negative $D$ values are problematic. The test developer either should exclude the item from the test or revise it. A thorough analysis of such item is required. Aiken (2003) suggested that “An item is usually considered acceptable if it has a D index of .30 or higher” (p.66).

Read article 5.3 by Michigan University Scoring Office on ‘Interpreting the Index of Discrimination’ in Folder Unit-5 on CD.

Table 5.2: Guidelines for Evaluating $D$ Values

<table>
<thead>
<tr>
<th>Sr.No.</th>
<th>Difficulty</th>
<th>Grading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>0.40 and larger</td>
<td>Excellent</td>
</tr>
<tr>
<td>2.</td>
<td>0.30-0.39</td>
<td>Good</td>
</tr>
<tr>
<td>3.</td>
<td>0.11-0.29</td>
<td>Fair</td>
</tr>
<tr>
<td>4.</td>
<td>0.00-0.10</td>
<td>Poor</td>
</tr>
<tr>
<td>5.</td>
<td>Negative Values</td>
<td>Mis-keyed or other major flaw</td>
</tr>
</tbody>
</table>

*Source based on Hopkins (1998)*
3. **Relationship between Item Difficulty and Discrimination Power**

The value of D and \( p \) are not independent. The variation in D effects the value of \( p \). A D value less than .30 is adequate if minimum and maximum \( p \) is within optimum acceptable range. And item with low D value and high \( p \) does not need to be discarded rather could be considered for modification (Aiken, 2003).

<table>
<thead>
<tr>
<th>Sr.No</th>
<th>Item Difficulty Index ((p))</th>
<th>Discrimination value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2.</td>
<td>0.90</td>
<td>0.20</td>
</tr>
<tr>
<td>3.</td>
<td>0.80</td>
<td>0.40</td>
</tr>
<tr>
<td>4.</td>
<td>0.70</td>
<td>0.60</td>
</tr>
<tr>
<td>5.</td>
<td>0.60</td>
<td>0.80</td>
</tr>
<tr>
<td>6.</td>
<td>0.50</td>
<td>1.00</td>
</tr>
<tr>
<td>7.</td>
<td>0.40</td>
<td>0.80</td>
</tr>
<tr>
<td>8.</td>
<td>0.30</td>
<td>0.60</td>
</tr>
<tr>
<td>9.</td>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>10.</td>
<td>0.10</td>
<td>0.20</td>
</tr>
<tr>
<td>11.</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*Source: Murphy and Davidshofer (1998)*

The D value of an item is related to its \( p \) value (see Table 5.3). Items that all examinees either pass or fail, (i.e., \( p \) values of either 0.0 or 1.0), do not provide any information about individual differences and their D values will always be zero. If half of the examinees correctly answered an item and half answered incorrectly (i.e., \( p \) value of 0.50), then it is possible for the item’s D value to be 1.0. It does not indicate that every item with \( p \) value of 0.50 will have D values of 1.0. The \( p \) value of an item will range between .20-.80 if an item has D value .40. There are situation when tests are either very easy or difficult therefore one can expect low D values. Moreover, such item which measure less important abilities or objectives could exhibit poor discrimination due to their unique focus. In this situation, if the item measures an important ability or learning objective and is free of technical defects it should be retained (Linn & Gronlund, 2000). No doubt constructing a good item is a time-consuming process, it is suggested that the guidelines provided in Table 5.2 with little variation in using these values is acceptable.

**ii) Item-Total Correlation Coefficients**

Item discrimination can be measured after correlating performance of the items (scored as either 0 or 1) with the total test score. This is referred to as an item-total correlation. The total score on a test is generally the total number of items answered correctly. The item-total correlation is calculated while using the point-biserial correlation. The point-biserial is used when one variable is a dichotomous nominal score and the other variable is measured on an interval or ratio scale. Here the dichotomous variable is the score on a single item (e.g., right or wrong) and the variable measured on an interval scale is the total test score. The high correlation value indicates that an item is measuring the same construct as the overall test measures and that the item discriminates between individuals high on that construct and those
low on the construct. The item discrimination index and the item-total correlation will reveal similar item discrimination index and can be interpreted in a similar manner (Hopkins, 1998). Calculation of total-item correlation is tedious tasks but becomes easy as teachers gain more access to computer test scoring programs.

iii) Item Discrimination of Mastery Test Items

The item difficulty indexes on mastery tests tend to be higher (indicating easier items) than on tests designed primarily to produce norm-referenced scores. This is because with mastery testing it is usually assumed that most examinees will be successful. As a result, on mastery tests it is common for items to have average $P$ values as high as 0.90 and the standard approach to interpreting item difficulty levels needs to be modified to accommodate this tendency.

The interpretation of indexes of discrimination is also complicated on mastery tests. Because it is common to obtain high $P$ values for both high and low scoring examinees, it is normal for traditional item discrimination indexes to underestimate an item’s true measurement characteristics. Several different approaches have been suggested for determining item discrimination mastery tests (Aiken, 2003; Popham, 2000). Aiken (2000) suggested that instead of taking 27% top scorers and 27% bottom scorers (or the top and bottom 50%), a cut score between the students who reached the mastery and those we showed no mastery will be decided. The formula is:

$$D = P_{\text{mastery}} - P_{\text{no mastery}}$$

Where

$P_{\text{mastery}} =$ proportion of mastery examinees getting the answer correct

$P_{\text{no-mastery}} =$ proportion of no-mastery examinees getting the answer correct

The advantage of this approach is that it can be calculated based on data from one test administration with one sample. It is common in mastery test that the majority of examinees reach at mastery level, therefore, the proportion of no-mastery group might be small and will result unstable statistics.

5.3.3 Item Analysis of Speed Tests

The interpretation of item analysis results is complicated with speed tests. The items on speed tests are often easy and examinees are required to complete the test with in prescribed time. Due to strict time limits test taker are not able to complete all of the items. Whereas, on power tests the examinee is given sufficient time to complete the test. But the items vary in difficulty with some being so difficult that no examinee will answer them all correctly. In many situations tests incorporate a combination of speed and power, so the speed-power distinction is actually one of degree.

On speed tests, measures of time difficulty and discrimination will largely reflect the location of the item in the test rather than the item’s actual difficulty level or ability to discriminate. Items appearing late on a speed test will be passed by fewer individuals than items that appear earlier simply because the strict time limits prevents students from being able to attempt them. The items appearing later on the tests are probably not actually more difficult than the earlier items; their item difficulty index will suggest that they are more difficult. Similar
complications arise when interpreting indexes of discrimination with speed tests. Because the individuals completing the later items also tend to be the most capable test takers, indexes of discrimination may exaggerate the discriminating ability of these items. Although different procedures have been developed to take into consideration these and related factors, they all have limitations and none have received widespread acceptance (Aiken, 2000; Anastasi & Umina, 1997). Our recommendation is that you should be aware of these issues and take them into consideration when interpreting the item analysis of highly speeded tests.

5.3.4 Distracter Analysis

The last stage in item analysis is the analysis of individual distracters for such items that do not show optimal range of difficulty level and discrimination index. On multiple-choice items, the incorrect alternatives are referred to as distracters because they “distract” examinees who do not actually know the correct response.

Distracter analysis provides the information to the teacher to examine that how many examinees in the top and bottom groups selected each option on a multiple-choices item. If a distracter is not selected by the examinees it is not doing its job. An effective distracter must be selected by some examinees. If a distracter is so obviously incorrect that no examinees select it, it is ineffective and needs to be revised or replaced. The second question involves discrimination. Did the distracter attract more examinees in the bottom group than in the top group to select it (i.e., negative discrimination) or attracted more examinees in top group than examinees in the bottom group (i.e., positive discrimination). Distracters should demonstrate negative discrimination!

Consider the following example:

**Example 1**

<table>
<thead>
<tr>
<th>Item 1</th>
<th>Options</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B*</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>Number in top group</td>
<td>2</td>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Number in bottom group</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

*Correct answer

This item has a $P=0.53$ (moderate difficulty) and a $D=0.40$ (good discrimination power). Based on these values, this item does not require modification but gives an good example for a “good” item. The D value shows that majority of examinees in the top group than the bottom group was correct in selecting the answer (i.e., option B). All distracters A, C, and D were selected by some examinees, which indicate they have distracted the examinees who have little or no knowledge about correct answer. The analysis of distracters further showed that all three distracters were selected more by members of the bottom group than the top group. In summery this is a good item and all of the distracters are performing well.

**Example 2**

<table>
<thead>
<tr>
<th>Item 1</th>
<th>Options</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C*</td>
<td>D</td>
</tr>
<tr>
<td>Number in top group</td>
<td>0</td>
<td>4</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Number in bottom group</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

*Correct answer
This item shows a $P = 0.47$ and a $D = 0.10$. Though this item has moderate difficulty level and Discrimination is also in acceptable range but needs revision. By examining option A you note that no one selected this distracter from both groups. But option B has been selected by more examinees in the top group than in the bottom group. It may be possible that the wording is ambiguous or that the option is similar in some way to the correct answer. Option D have been selected more poor-performing examinees than top performing examinees (i.e., 6 versus 3). It is likely that if the test developer revises options A and B this will be a more effective item.

**Effectiveness of Distracter for item Difficulty and Discrimination**

In the example below you will see the effective of distracters. The selection of distracters impacts both item difficulty and discrimination. Consider the following item:

1. In what year the Pakistan Resolution was passed?
   a. 1930
   b. 1935
   c. 1940
   d. 1945

Even the examiner knows about Pakistan movement this item will be fairly difficult question due to the closeness of distracters:

Now consider this revision:

2. In what year the Pakistan Resolution was passed?
   1. 1690
   2. 1750
   3. 1870
   4. 1940

This is the same question but different distracters. This revised item would likely be a much easier item in a typical high school science class. The point is that the selection of distracters can significantly impact the difficulty of the item and consequently the ability of the item to discriminate. Now study carefully a section on distractor analysis in article 6.4 for further clarification.

*Read Article 5.4 by Kehoe (1995) on ‘Item Analysis for Multiple-Choice Items’ in folder Unit-5 on CD*

**5.4 Application of Item Analysis for Teachers**

Teachers typically have a number of practical options for calculating item analysis statistics for their classroom tests. Many teachers will have access to computer scoring programs that calculate the various item analysis statistics we have described. Numerous commercial companies sell scanners and scoring software that can scan answer sheets and produce item analysis statistics and related printouts. If you do not have access to computer scoring at your school, website Reactions has an excellent Internet site that shows you to compute common item analysis statistics online (www.surveyreaction.com/itemanalysis.asp).
The procedure involves the following steps:

- After scoring the test, arrange them according to score (i.e., lowest score to highest score).
- Take the papers of ten top scorers and the 10 papers with the lowest score. Set these into two piles, one with the 10 highest scores (i.e., top group) and one with the lowest scores (i.e., bottom group). The rest of the papers will be not used in analysis.
- For each item determine how many of the students in the top group correctly answered it and how many in the bottom group correctly answered it. With this information you can calculate the overall item difficulty index (i.e., P) and separate item difficulty indexes for the top group ($P_t$) and bottom group ($P_b$). For example if 8 students in the top group answered the item correctly and 3 in the bottom group answered the item correctly, add these together (8+3=11) and divide by 20 to compute the item difficulty index: $p=11/2=0.55$. Although this item difficulty index is based on only the highest and lowest scores it is usually adequate for use with classroom tests. You can then calculate $P_t$ and $P_b$. In this case:
  
  $$P_t = \frac{8}{10} = 0.80 \text{ and } P_b = \frac{3}{10} = 0.30.$$

- You now have the data needed to calculate the discrimination index for the item. Using the data from our example:
  
  $$D = P_t - P_b = 0.80 - 0.30 = 0.50$$

Using these simple procedures you see that for this item $p = 0.55$ (moderate difficulty) and $D=0.50$ (excellent discrimination). If your items are multiple-choice you can also use these same groups to perform distracter analysis.

Now consider the following example:

**Example 3**

<table>
<thead>
<tr>
<th>Item 1</th>
<th>Options</th>
<th>A</th>
<th>B*</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top group (top 10)</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Bottom group (bottom 10)</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

*Correct answer

As reflected in the item D value (i.e., 0.50), more students in the top group than bottom group selected the correct answer (i.e., option B). By examining the distracters (i.e., options A, C, and D), you see that they each were selected by some students (i.e., they are all distracting as hoped for) and they were all selected by more students in the bottom group than the top group (i.e., demonstrating negative discrimination). In summary, this item is functioning well!

**Activity**

Prepare a tests consisting of multiple choice questions, administer it and assess the quality of items with the help of difficulty and discrimination indexes as well as distracter analysis.
5.5 Using Item Analysis to Improve Items

The following few examples will enable you to improve the quality of test items:
Consider this information:

Example 4

<table>
<thead>
<tr>
<th>Options</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B*</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

| Number in top group | 2 | 10 | 2 | 1 |
| Number in bottom group | 3 | 6 | 4 | 2 |

*Correct answer  \( P = 0.53 \)  \( D = 0.37 \)

Now consider the following questions:
1. Is the item difficulty level appropriate for the testing application? A \( p \) of 0.53 is acceptable for a multiple-choice item on a norm-referenced test. You have already learn that the optimal mean \( p \) value for a multiple-choice item with four choices is 0.74.
2. Does the item discriminate adequately? With a \( D \) of 0.37 this item is a good discriminator since it discriminates between those who have knowledgeable and those who are poor.
3. Are the distracters performing adequately? It is obvious from the example the all distracters are equally attractive for some of the examiners from both groups.
4. Overall evaluation? In summary, this is good item and no revision is necessary.

Now consider a problematic item

Example 5

<table>
<thead>
<tr>
<th>Items</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
</tbody>
</table>

| Number in top group | 20 | 4 | 4 | 2 |
| Number in bottom group | 11 | 6 | 8 | 5 |

*Correct answer  \( P = 0.20 \)  \( D = -0.13 \)

1. Is the item difficulty level appropriate for the testing application? A \( p \) of 0.20 suggests that this item is too difficult for most applications. Unless there is some reason for including item that is this difficult, this is cause for concern.
2. Does the item discriminate adequately? A \( D \) of -0.13 suggests major problems with this item. It may be miskeyed or some other major flaw is present.
3. Are the distracters performing adequately? Option A, attracted most of the examinees in the top group and a large number of examinees in the bottom group. The options B, C, and D were negative discriminators because these were selected by more examinees in the bottom group than the top group.
4. Overall evaluation? five times as many examinees in the top group selected option A than option C, which is keyed as correct, we need to verify that option C actually is the correct response. Assuming item has been miskeyed and correct answer is A not C then the item is acceptable. Since the item is not miskeyed therefore there is some major problem with the item, should be deleted.

5.6 Qualitative Item Analysis

In addition to the quantitative item analysis procedures, there are qualitative item analysis procedures to improve the tests. Popham (2000) provided some useful suggestions (p. 314).
1. He suggested that after writing the test items set the test aside for a few days taking a break by distancing from the test. Immediately reviewing test items will leave spelling and grammar unnoticed. Review after a break of some time will enable you detect such errors that you would not be able to catch in immediate reviewing. The time you spend proofing a test is well spent and can help you avoid problems once the test is administered and scored.
2. Popham (2000) also suggests that ask a non-student reviewers most probably a colleague who is familiar with the content of the test to review the test items. For example, a math teacher might have another math teacher review the test. In addition to checking for clerical errors, clarity, and accuracy, the reviewer should determine: (P.316). 1) Is the item congruent with its assessment domain? 2) Are there violations of standard item-writing guidelines? 3) Is the content of the item accurate? 4) Is the item ethically, socioeconomically, or otherwise bias?
3. Ask the examinees to provide feedback on the test. For example, after completing the test you might have the examinees complete a brief questionnaire asking whether the directions were clear and if any of the questions were confusing.

Ideally both quantitative and qualitative approaches may be used to improve tests. Often a combination of quantitative and qualitative procedures will result in the optimal enhancement of classroom tests.

Popham (2000) argues that efficiency of an instrument can be judged by detecting the differences among students on the basis of their performance on the test. Generally quantitative item analysis procedures have been applied for norm-referenced score interpretations. Popham says “thus, for tests aimed at norm-referenced interpretation, there is usually great alliance on empirical methods in the improvements of test items” (p.314). He suggest that in many cases irrespective of test’s focus on norm-referenced or criterion-referenced interpretations, were equally useful and beneficial for improving the tests.

5.7 Using Item Analysis to Improve Classroom Instruction

Developing Item Banks

Developing a test item bank is very helpful for teacher. Teacher can select new item every time when they develop new tests. It will allow teacher to avoid those items that have been included in the test previously. Consider the following example.
Example:

Learning Objective: Describe the measures of variability and their appropriate use.

- If the variance of a set of scores is 16, the standard deviation will be equal to:
  a. 3
  b. 4
  c. 5
  d. 6

Administration date: March, 2006

<table>
<thead>
<tr>
<th>Group</th>
<th>Options</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number in top group</td>
<td>A</td>
<td>B*</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>28</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Number in bottom group</td>
<td>8</td>
<td>16</td>
<td>6</td>
<td>10</td>
</tr>
</tbody>
</table>

Correct answer  
P = 0.55  
D = 0.30

Administration date: March, 2007

<table>
<thead>
<tr>
<th>Group</th>
<th>Options</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number in top group</td>
<td>A</td>
<td>B*</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>30</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Number in bottom group</td>
<td>10</td>
<td>20</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Correct answer  
P = 0.62  
D = 0.25

This indicates that this item has been administrated on two different occasions. By including information from multiple administrations, you will have a better idea of how the item is likely to perform on a future test. (see www.assess.com for assessing the item).

Read Article 5.5 on folder Unit-5 by Haladyna, Dowing, & Rodriguez (2002) on ‘How can the use of item analysis benefit your students including those with special needs?’

5.8 Self-Assessment Questions

1. Describe different techniques of analyzing effectiveness of test items.
2. What practical strategies teachers can use to improve test items?
3. Discuss some qualitative procedures of item analysis.
4. What is the difficulty ($p$) and discrimination (D) index of a test item administered to 80 students if 18 of the upper group (upper 25%) and 12 of the lower group (lower 25%) get the item right?
5. Compute the difficulty ($p$) and discrimination (D) indexes for an item on a mastery test administered on 50 students, 30 of whom reached the mastery level on total test
score. Of the 30 who reached mastery level, 20 get the item right; of the 20 who do not reach the mastery level, 10 get the item right.

6. Calculate difficulty level and discrimination index on the following item. Also perform distracter analysis for a multiple choice item on a test of secondary school students’ computation skills. Examine the effectiveness of each distracter and suggest changes for the improvement of item.

<table>
<thead>
<tr>
<th>Item 5</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper 27%</td>
<td>1</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Lower 27%</td>
<td>3</td>
<td>0</td>
<td>12</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

5.9 Suggested Readings


5.10 Web-Based Material

Unit–6

TEST SCORES AND NORMS
## CONTENTS

6.1 Introduction of the Unit .................................................................87
6.2 Objectives of the Unit ..................................................................87
6.3 What is a Raw Score? ....................................................................87
6.4 What are Norms? .......................................................................88
6.5 Norm-Referenced Tests (NRT) and Criterion-Referenced Tests (CRT) ..........88
6.6 Norm-Referenced Interpretations ..................................................89
6.7 The Normal Curve ......................................................................90
6.8 Derived Scores ...........................................................................93
6.9 Types of Norms ..........................................................................96
6.10 Criterion-Referenced Interpretations .........................................97
6.11 Norm-Referenced VS Criterion-Referenced Tests ......................98
6.12 Qualitative Description of Scores ..............................................99
6.13 Self Assessment Questions .........................................................99
6.14 Suggested Readings ................................................................100
6.15 Web Based Material ................................................................100
6.1 Introduction of the Unit

Test scores reflect the performance or rating of the individuals taking a test. Test scores are the keys to interpreting and understanding the examinees’ performance, their meaning and interpretation are extremely important topics and deserve careful attention. Evaluation is a process in which we put a value or assign a worth to something As you have read in previous chapters that the essential characteristics of evaluation is judgment. Whereas measurement is quantitative which, describes something in terms of specific numbers or percentage. In evaluation a judgment is made in attaching a value or qualitative description to a measurement derived from a test. A typical raw score for the sake of interpretation are converted into standard scores. There are number of procedure for converting raw scores in standard score. These derived scores can be classified as criterion-referenced or norm-referenced. There is a wide variety of scores available for use in education and each has its own unique characteristics. This unit and related web based allied material and further references will explain the following concepts: Raw scores, norm and types of norms, norm-referenced vs criterion-referenced tests and their interpretations, transforming and derivation of raw scores into standard scores, and relationship between different standard scores.

6.2 Objectives of the Unit

After completing this unit you should be able to:
1. Differentiate between raw scores and norms.
2. Describe Norm-referenced and Criterion-referenced score.
3. Differentiate between Norm-referenced and Criterion-referenced score interpretations.
4. Describe the characteristics of Normal Curve.
5. Transform raw scores into different standard score formats.
6. Describe relationship between different standard score formats.
7. Apply Norm-referenced and Criterion-referenced score interpretations in appropriate situations.
8. Evaluate the importance of Norm-referenced and Criterion-referenced score interpretations.

6.3 What is a Raw Score?

The simplest type of achievement score is a ‘raw score’. Generally a raw score is the number of items marked or coded such as yes/no, correct/incorrect, true/false, and so on. For example, the raw score on a classroom test might be the number of items a student has answered correctly. The calculation of raw score is very simple. The teacher adds the marks assigned to correct answers. But raw scores have limited usefulness in interpreting the test results because they offer very little useful information. Let’s say a student’s score on a math test is 70 out of 100. Does a raw score of 70 represent is poor, average or high performance? The answer to this question depends on a number of factors such as how many items are on the test, how difficult the items are, how many
students scored above 70 and how many students scored below 70. If the test contained only 70 items and the student’s raw score were 70, then the student demonstrated perfect performance. If the test contained 100 items and the student’s raw score was 70, then we can say that he or she answered more than half of the items correctly. However, we still do not know what that really means. If the test contained 100 extremely difficult items and a raw score of 70 were the highest score in the class this would likely reflect very good performance on the other hand if the test included more easy items then 70 will most probably be considered poor performance.

6.4 What are Norms?

“Norms” is the plural form of norm. In a psychometric context, norms are the test performance data of a particular group of test takers that are designed for use as a reference for evaluating or interpreting individual test scores. The technical manual of all standardized tests contain norms for the test. The particular group of test takers may be defined broadly (for example a sample (representative of adult population of Pakistan) or narrowly (twelve years old students of elementary schools of community school). Whether broad or narrow in scope, members of the group will be typical with respect to some characteristic(s) of the people for whom the particular test was designed. A test administered to this representative sample of test takers yields a distribution or distributions of scores. These data constitute the norms for the test and typically are used as a reference source for evaluating and placing into context test scores obtained by individual test takers. The data may be in the form of raw scores or converted score. You will learn about different types of converted scores in preceding sections.

6.5 Norm-Referenced Tests (NRT) and Criterion-Referenced Tests (CRT)

Standardized tests are norm-referenced for that is, the performance of sample population has been established and serves as a basis of interpreting a student’s relative performance. A norm-referenced measure allows us to compare one individual with other individuals. The norms are based on a large population which is used to compare the score of a student on a test with students from others schools. For example scores on intelligence which is a standardized test is norm-referenced. If a person IQ score is 110 this indicates that he/she has scored higher than or 50% more than the people taken from a sample. This comparison is called norm-referenced interpretation.

Norm-referenced tests tend to have high estimates of reliability and validity because the norms are based on large populations. The test manual reports reliability and validity data obtained for various sample populations. A test manual provides comprehensive descriptions of procedures in establishing normative data. Norms are also reported with reference to age, gender, ethnicity, geographical setting, etc.

Sometimes an educator is not concerned with how well a student performs compared to other students, but whether the student exhibits progress in learning. The educator establishes a set of objective with corresponding proficiency or achievement levels and then determines whether the student can achieve an acceptable proficiency or achievement level. Rather than comparing the student with other students, the teacher assesses the student only on the basis
of predetermined standard. The criterion-reference interpretations emphasizes on what the examinees know or what they know about a specific body of knowledge or skill. Criterion-reference tests are usually locally developed and sometimes teacher made. The criterion-referenced tests allow the teachers to judge the students’ proficiency in specific content areas, and therefore they have better curricular validity than norm-referenced tests. The most common example of criterion-referenced score is the percentage of correct response on a classroom examination. If you report that a student correctly answered 90% of the items on a classroom test, this is a criterion-referenced interpretation. Norm-referenced interpretations are relative (i.e. relative to the performance of other examinees) whereas criterion-referenced interpretations are absolute (i.e. compared to an absolute standard). Criterion-referenced tests may be practical in areas of achievements that focus on the acquisition of specific knowledge. It can be noted that it is difficult to develop reliable criterion-referenced tests, since most instructional units deal with specific curriculum and instruction of information.

See following articles in web Material folder Unit-6 on CD for further explanation and comparison of Norm-referenced and Criterion-referenced Tests:

Article 6.1  Comparison of NRM and CRM (web reference),
Article 6.2 by Huitt (1996) on ‘Criterion Vs Norm Referenced Testing’

6.6 Norm-Referenced Interpretations

As we have discussed in section 6.5 that norm-referenced test measure a student’s level of achievement at a given period of time compared to other students elsewhere. Scores from criterion-referenced test do not indicate a relative level of achievement or produce standards because no comparisons are made. The test indicates that how efficient a student is in terms of a specific body of learning. It can measure changes in learning overtime. According to researchers the norm-referenced test is valuable measuring higher and abstract levels of cognitive domain, whereas, the criterion-referenced test is valuable for measuring lower and concrete levels of learning. The norm reference is valuable or heterogeneous groups in which the range of abilities is wide and a test is intended to measure a wide range of performance. One of the characteristics of norm-referenced tests is that it emphasizes discrimination among students in terms of relative level of learning or achievement. It also favors average difficulty and omits easy and difficult items. Due to this characteristic criterion test is used mastery or specific test situation. Interpretation of a norm-referenced score is based on a defined group. For norm-referenced interpretations to be meaningful, teachers need to compare the examinee’s performance to that of a relevant reference group or sample. Therefore, the first step in developing good normative data is to define clearly the population for whom the test is designed.

Reynolds, Livingstone, and Willson (2006, p.61), write that:
1. After defining the population clearly and appropriately, a stratified random sample is selected from target population and is tested. The randomly selected reference is used to derive scores and is called standardization sample.
2. After selecting the standardization sample, it is tested, scores are derived and tables of derived scores are developed. These tables are based on the performance of the standardized sample and are referred to as normative tables or “norms”.

Reynolds, Livingstone, and Willson (2006) have made the following considerations:

i. Is the standardization sample representative of the examinees to whom test will be administered? Are demographic characteristics of the sample (e.g., age, race, sex, education, geographical location, etc.) similar to those who will take the test?

ii. Is the sample current? Participants in samples from twenty years ago may have responded quite differently from a contemporary sample. Attitudes, beliefs, behaviors and even cognitive abilities change over time, and to be relevant the normative data need to be current.

iii. Is the sample size large enough to provide appropriate statistical information? Although there is no magic number, if a test covers a broad age range it is common for standardization samples to exceed 1,000 participants. Otherwise, the number of participants at each age or grade level may be too small to produce stable estimation of means, standard deviations and more general distribution of scores.

A final consideration regarding norm-referenced interpretations is the importance of standardized administration. The normative sample should be administered under the same conditions and with the same administrative procedures that will be used in actual practice. Accordingly, when the test is administered in educational settings, it is important that test user follow the administrative procedures precisely. For example, if you are administering standardized tests, you need to make sure that you are reading the direction verbatim and closely adhering to time limits. It obviously would not be reasonable to compare your students’ performance on a time mathematics test to the performance of standardization sample that was given either more or less time to complete the items. (The need to follow standard administration in scoring procedures actually applies to all standardized tests, both norm-referenced and criterion-referenced).

Many types of derived scores or unit of measurement may be reported “norms tables” in selection of which derived score to employ can influence the interpretation of scores. Before starting our discussion of common norm-referenced derived scores we need to introduce the concept of a normal distribution.

6.7 The Normal Curve

Theoretically, the normal curve is a bell-shaped, smooth, mathematically defined curve highest at the center and gradually extends on both sides approaching the x-axis asymmetrically. The curve is perfectly symmetrical, with no skewness, so if you folded it in half at the center one side it would lie exactly on top of the other. Figure 6.1 depicts a normal distribution.
Referring figure 6.1, you can see that a large number of scores tend to “pile-up” around the middle of distribution. However, for a relatively small number of individuals a unique combination of factor results in them being either much shorter or much taller than the average. This accounts for the distribution tailing off at both the low and high ends.

Although the previous discussion addressed only observable characteristics of the normal distribution, certain mathematical properties make it particularly useful when interpreting scores. For example, the normal distribution is a symmetrical, uni-modal distribution in which the mean, median, and mode are all equal. It is also symmetrical, meaning that if you divide the distribution into two halves, they will mirror each other. Probably the most useful characteristic of the normal distribution is that predictable proportions of scores occur at specific points in the distribution.

Referring to Fig. 6.1 you find a normal distribution with the mean 10 and standard deviation, 5. You must know that the mean and median are the same in a normal distribution, we know that an individual who scores at the mean score is better than 50% of the sample examinees see fig.6.1 (a) (remember, the median is the score that divides the distribution in half). In normal distribution approximately 34% of the scores fall between the mean and 1 standard deviation above the mean (in figure 6.1 the unit of one standard deviation is 5), an individual whose score falls 1 standard deviation above the mean (10+5=15) and one standard deviation below the means cover 68% area of the curve (figure 6.1 (b). Those who perform one standard deviation above mean exceeding approximately 84 percent (i.e., 50%+34%) of the population. A score 2 standard deviation above the mean (i.e 10+5+5= 20) in figure 6.1 (c) and 2 standard deviation below mean will be about 95% of population.
Because the distribution is symmetrical, the relationship is the same in the inverse
below the mean. A score one standard deviation below the mean indicates that the
individuals exceeds only about 16% (i.e., 50%-34% in figure 6.1 10-5=5) of the
population with score within 1 standard deviation above and below the mean on normally
distributed variable.

For further clarification see figure 6.1 (d)

Many variables in educational settings such as achievements and intelligence
scores are normally distributed. Not all educational, psychological, or behavioral
variables are normally distributed. Some negative behavior like, aggressive behavior and
psychotic behavior, hate, cruelty, and negative attitudes are the variables distinctly
different from the normal curve in their distributions. Such variable will each has its own
unique distribution.
6.8 Derived Scores

Because raw score in most situations have little interpretative meaning, we need to transform or convert them into another format to facilitate their interpretation and give them meaning. These transformed scores, referred to as derived scores, standard scores, or scaled scores, are pivotal in helping us interpret test results. There are a number of different derived scores, but they all can be classified as either norm-referenced or criterion referenced. We will begin our discussion of scores and their interpretation by introducing you to these two different approaches to deriving and interpreting test scores.

1. Standard scores

Simply stated, a standard score is a raw score that has been converted from one scale to another scale is most widely used and interpretable. Different types of systems for standard scores exist, each unique as regards its respective mean and standard deviations. You will learn in preceding sections about number of different standard score formats.

- Z scores

A z score is equal to the difference between particular score and the mean divided by the standard deviation. In other terms a z score expresses a score in terms of the number of standard deviation units the raw score is below or above the mean. Using an example from the normally distributed “GAT” conducted by NTS having a mean 50 and standard deviation 15 we can convert raw score 65 to a z score using the following formula:

\[ Z = \frac{X - \mu}{\sigma} \]

\[ = \frac{65 - 50}{15} = 15/15 = 1 \]

In this test a raw score of 65 is equal to a z score of +1. Knowing simply that someone obtained a raw score on a test gives no useable information. However knowing that someone obtained a z score of 1 provides context and meaning to the score. It further gives the knowledge about the area under the normal curve, for example we would know that only 16% of the other testtakers obtained higher scores. The scales used in normal curve are called ‘Zero plus or minus one’, because the mean of a theoretical normal curve is zero and standard deviation is one.

- T-scores

The scale used in the computation of T-scores is called a “fifty plus or minus ten” (Cohen and Swerdlik, 2002, p.95). The T-scores have a mean of 50 and a standard deviation of 10. Relative to z-scores they have the advantage of all scores being positive and without decimals. For example, a score 66 is 1.6 standard deviation above the mean (i.e., exceeding 95% of the scores in the distribution) and a score of 34 is 1.6 standard deviation below the mean.
deviation below the mean (i.e., exceeding only 5% of the scores in the distribution). This standard score system is composed of a score that ranges from 5 standard deviations below the mean to 5 standard deviations above the mean. If you wanted to convert individual’s score to a T-score, you would use the following formula:

You can easily convert standard scores from one format to another using the following formula;

\[ T\text{-score} = 50 + 10 \times (z\text{ score}) \]

In our previous example calculated z score is 1 therefore the T score will be:

\[ = 50 + 10 \times (1) \]
\[ = 50 + 10 \]
\[ = 60 \]

♦ Wechsler IQs (and many others)

The Wechsler intelligence scales use a standard score format with a mean of 100 and a standard deviation of 15. Like T-scores the Wechsler IQ format avoids decimal and negative values. For example a score of 124 is 1.6 standard deviations above the mean (i.e., exceeding 95% of the score in distribution). This format has become very popular, and most aptitude and individually administered achievement tests report standard scores with mean of 100 and standard deviation of 15.

♦ Stanford-Binet IQs

The Stanford-Binet intelligence scales use a standard score format with a mean of 100 and a standard deviation of 15. This is similar to the format adopted by the Wechsler scales.

Table below illustrates a simple formula for converting standard scores from one format to another (e.g., z-scores to T-scores).

2. Normalized Standard Score:

Normalized standard scores are standard scores based on underlying distributions that were not originally normal, but were transformed into normal distributions. The transformations applied in these situations are often nonlinear transformations. Whereas standard scores calculated with linear transformations retain a direct relationship with the original raw scores and the distribution retains its original shape, this is not necessarily so with normalized standard scores based on nonlinear transformations. This does not mean that normalized standard scores are undesirable. In situations in which the obtained distributions are not normal because the variable is not normally distributed, normalization is not generally useful and indeed may be misleading. However, in situations in which the obtained distribution is not normal because of sampling error or choice of subjects, normalization can enhance the usefulness and interpretability of the scores. Nevertheless, it is desirable to know what type of score you are working with and how they are calculated.

In most situations normalized standard scores are interpreted in a manner similar to other standard scores. These include:
Stanine scores

Stanine (i.e., standard nine) scores divide the distribution into nine bands (1 through 9). Stanine Scores have a mean of 5 and a standard deviation of 2. Because Stanine scores use only nine values to represent the full range of scores, they are not particularly precise score format. As a result some professionals avoid their use. However certain professionals prefer them because of their imprecision. These professionals, concerned with the imprecision inherent in all psychological and educational measurement, chose stanine scores because they do not misrepresent the precision of measurement (Popham, 2000).

Relationship of Different Standard Score Formats

<table>
<thead>
<tr>
<th>z-scores X = 0 SD = 1</th>
<th>T-scores X = 50 SD = 10</th>
<th>Wechsler IQ X = 100 Sd = 15</th>
<th>Percentile Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4</td>
<td>74</td>
<td>136</td>
<td>99</td>
</tr>
<tr>
<td>2.2</td>
<td>72</td>
<td>133</td>
<td>99</td>
</tr>
<tr>
<td>2.0</td>
<td>70</td>
<td>130</td>
<td>98</td>
</tr>
<tr>
<td>1.8</td>
<td>68</td>
<td>127</td>
<td>96</td>
</tr>
<tr>
<td>1.6</td>
<td>66</td>
<td>124</td>
<td>95</td>
</tr>
<tr>
<td>1.4</td>
<td>64</td>
<td>121</td>
<td>92</td>
</tr>
<tr>
<td>1.2</td>
<td>62</td>
<td>118</td>
<td>88</td>
</tr>
<tr>
<td>1.0</td>
<td>60</td>
<td>115</td>
<td>84</td>
</tr>
<tr>
<td>0.8</td>
<td>58</td>
<td>112</td>
<td>79</td>
</tr>
<tr>
<td>0.6</td>
<td>56</td>
<td>109</td>
<td>73</td>
</tr>
<tr>
<td>0.4</td>
<td>54</td>
<td>106</td>
<td>66</td>
</tr>
<tr>
<td>0.2</td>
<td>52</td>
<td>103</td>
<td>58</td>
</tr>
<tr>
<td>0.0</td>
<td>50</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>-0.2</td>
<td>48</td>
<td>97</td>
<td>42</td>
</tr>
<tr>
<td>-0.4</td>
<td>46</td>
<td>94</td>
<td>34</td>
</tr>
<tr>
<td>-0.6</td>
<td>44</td>
<td>91</td>
<td>27</td>
</tr>
<tr>
<td>-0.8</td>
<td>42</td>
<td>88</td>
<td>21</td>
</tr>
<tr>
<td>-1.0</td>
<td>40</td>
<td>85</td>
<td>16</td>
</tr>
<tr>
<td>-1.2</td>
<td>38</td>
<td>82</td>
<td>12</td>
</tr>
<tr>
<td>-1.4</td>
<td>36</td>
<td>79</td>
<td>8</td>
</tr>
<tr>
<td>-1.6</td>
<td>34</td>
<td>76</td>
<td>5</td>
</tr>
<tr>
<td>-1.8</td>
<td>32</td>
<td>73</td>
<td>4</td>
</tr>
<tr>
<td>-2.0</td>
<td>30</td>
<td>70</td>
<td>2</td>
</tr>
<tr>
<td>-2.2</td>
<td>28</td>
<td>67</td>
<td>1</td>
</tr>
<tr>
<td>-2.4</td>
<td>26</td>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>-2.6</td>
<td>24</td>
<td>61</td>
<td>1</td>
</tr>
</tbody>
</table>

X = mean, SD = standard deviation
Source: Adapted from Reynolds, Livingston, and Willson (2006).
Activity
Take a test of a group of students and convert their raw scores into z and T scores.

6.9 Types of Norms:

We can classify norms as: grade norms, age norms, national norms, subgroup norms, local norms, norms from a fixed reference group, and percentile norms. In this section the percentile norms has been presented first because the norms for many tests re expressed as percentile norms.

(i) Percentiles

We could divided a distribution into 100 equal parts——100 percentiles. A percentile is an expression of the percentage of people whose score on a test or a measure falls below a particular raw score. A more familiar description of test performance, the concept of percent correct, must be distinguished from the concept of a percentile. A percentile is a converted score that refers to a percentage of test takers and the percentage correct refers to the distribution of raw scores, specially the number of items that were answered correctly multiplied by 100 and dived by total number of items. In such a distribution, the xth percentile is equal to the score at below which x% of scores fall. Thus 20th percentile is the score below which 20% of the scores in the distribution fall. Because percentiles are easily calculated they are popular way of organizing test data.

(ii) Grade Norms

Grade norms are designed to indicate the average test performance developed by administering the test to representative sample of children over a range of consecutive grade levels. Next, the median or mean score for children at each grade level is computed. A common misperception is that children should receive instruction at the level suggested by their grade equivalents. Parents may ask, “Naila is only in the 4th grade but has a grade equivalent of 6 in math. Doesn’t that mean she is ready for 6th grade math instruction? The answer is clearly “No!” Although Naila correctly answered the same number of items as an average 6th grader, this does not indicate that she has mastered the necessary prerequisites to succeed at the 6th grade level.

(iii) Age Norms

Age equivalents are another derived score format that indicates the age, typically in years and months, at which a raw score is the mean or median. Age equivalents have the same limitations as grade equivalents and we again recommend that you avoid using them. Many test publishers report grade and age equivalents and occasionally you will find a testing expert that favors them (at least at the lower grade levels). Nevertheless, they are subject to misinterpretation and should be avoided when possible.

(iv) National Norms

National norms are derived from a standardized representative sample of population. In the field of education and psychology national norms are obtained though the testing of large number of students representative of different variables like age,
gender, socioeconomic status, or geographical location etc. Norms would typically be obtained for every grade to which the test sought to be applicable, and other factors related to the representativeness of the school itself might be criteria for inclusion or exclusion from the representative sample.

(v) **Local Norms**

Local norms provide normative information with respect to the local population’s performance on some test. A nationally standardized test may select a representative sample from every strata of the society. That could be used for separating local norms from national norms. Individual high schools may wish to develop their own school norms for student scores own some examination that is administered in the country.

(vi) **Subgroup Norms**

Subgroup norms for narrowly defined groups can be developed. For example if we are interested to developed a standardized ‘Reading Comprehension’ test for children of different age, educational level, gender, geographical region, community type, then the manual will include normative information with respect to all variables of interest.

*See article 6.3 on web folder Unit-6 by D’Onofrio (2002) on ‘Standardized Scores’.*

**6.10 Criterion-Referenced Interpretations**

In Criterion-referenced interpretations the examinee’s performance is not compared to that of other people, but to a specified level of performance (i.e., a criterion). Criterion-referenced interpretations emphasize what the examinees know or what they can do, not their standing relative to other test takers. For example, educators evaluate their students’ performance in terms of “percentage correct” or letter grades to reflect mastery (i.e., A, B, C, D, and F).

1. **Percent Correct**

The most common example of a criterion-referenced score is percent correct. For example, when a teacher reports that a student correctly answered 85% of the problems on a classroom test assessing the student’s ability to multiply double digits; this is a criterion-referenced interpretation. Although there is a variety of criterion referenced scoring systems, they all involve an absolute evaluation of examinees’ performance of others (a: a relative interpretation), a criterion referenced interpretation attempts to describe what they know or capable of doing---the absolute level of performance.

2. **Mastery Testing**

In addition to percent correct, another type of criterion-referenced interpretation is referred to as mastery testing. Mastery testing involves determining whether the examinee has achieved a specific level of mastery of knowledge or skills domain and is usually reported in all-or-none such as a pass/fail designation. A cut score had been previously established and all scores equal to or above this score are reported as “pass” whereas scores below it are reported as “fail.” If the cut score requires correctly answering 85% of the items, all examinees with scores of 84% or below fail and all with 85% and above pass.
3. **Standard-Based Interpretations**

Another common criterion-referenced interpretative approach is referred to as “standard-based interpretations.” Whereas mastery testing typically results in all-or-none interpretation (i.e., the student either passes or fails), standard-based interpretations usually involve three to five performance categories (Reynolds, Livingston, and Willson, 2006). For example, the result of an achievement test might be reported as not proficient, partially proficient, proficient, or advanced performance (Linn & Gronlund, 2000). An old variant of this approach is the assignment of letter grades to reflect performance on classroom achievement tests. For example, many teachers assign letter grades based on the percentage of items correct on a test, which is another type of criterion-referenced interpretation. For example, ‘A’ might be assigned for percentage correct scores between 90% and 100%, ‘B’ for scores between 80% and 89%, ‘C’ for scores between 70% and 79%, ‘D’ for scores between 60% and 69% and ‘F’ for scores below 60%. Note that with this system a student with a score of 95% receives an ‘A’ regardless of how other students scored. If all of the students in class correctly answered 90% or more of the items correctly, they would all receive ‘A’ on the test.

For criterion-referenced interpretations to provide useful information about what students know or what skills they possess, it is important that the knowledge or skill domain assessed by the test be clearly defined. Although criterion-referenced interpretations are most applicable to narrowly defined domains, they are often applied to broader, less clearly defined domains. For example, most tests used for licensing professionals such as physicians, lawyers, teachers, or psychologists involve criterion reference interpretations. Dunn, Parry and Morgan (2002) pointed out some difficulties in the practical use of criterion-referenced assessment. You are advised to read the following article.

Dunn, Parry and Morgan (2002) explored some of the difficulties with implementing criterion-referenced assessment. See article 6.4 ‘Seeking Quality in Criterion Referenced Assessment’ in Folder 6 on CD.

### 6.11 Norm-Referenced Vs Criterion-Referenced

Bond describes the differences between these two types of assessments and explains the most appropriate uses of each. See article 6.5 ‘Norm and Criterion Referenced Testing’ in Folder 6 on CD.

<table>
<thead>
<tr>
<th>Characteristics of Norm-Referenced and Criterion-Referenced Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Norm-Referenced Interpretations</strong></td>
</tr>
<tr>
<td>Compare performance to a specific group—a relative interpretation</td>
</tr>
<tr>
<td>Useful interpretations require a relevant reference group</td>
</tr>
<tr>
<td>Usually assess a fairly broad range of knowledge or skills</td>
</tr>
<tr>
<td>Items are selected that are of medium difficulty and maximize variance: very difficult and very easy items are usually deleted</td>
</tr>
<tr>
<td>Example: percentile rank—a percentile rank of 80 indicates that the examinee scored better than 80% of the subject in the reference group</td>
</tr>
</tbody>
</table>
6.12 Qualitative Description of Scores

Test developers commonly provide qualitative description of the scores produced by their tests. These qualitative descriptors help professionals communicate results in written reports and other formats. For example, the Stanford-Binet Intelligence Scales, Fifth Edition (SB5) (Roid, 2003), provides the following qualitative descriptions:

TABLE 6.4: Qualitative Descriptions of IQ Scores

<table>
<thead>
<tr>
<th>IQ</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>145 and above</td>
<td>Very Gifted or Highly Advanced</td>
</tr>
<tr>
<td>130-144</td>
<td>Gifted or Very Advanced</td>
</tr>
<tr>
<td>120-129</td>
<td>Superior</td>
</tr>
<tr>
<td>110-119</td>
<td>High Average</td>
</tr>
<tr>
<td>90-109</td>
<td>Average</td>
</tr>
<tr>
<td>80-89</td>
<td>Low Average</td>
</tr>
<tr>
<td>70-79</td>
<td>Borderline Impaired or Delayed</td>
</tr>
<tr>
<td>55-69</td>
<td>Middle Impaired or Delayed</td>
</tr>
<tr>
<td>40-54</td>
<td>Moderately Impaired or Delayed</td>
</tr>
</tbody>
</table>

These qualitative descriptors help professionals communicate information about an examinee’s performance in an accurate and consistent manner.

Internet sites of interest

www.teachersandfamilies.com/open/parent/scores1/cfm
Understanding test scores: A Premier for Parents is a user-friendly discussion of tests that is accurate and readable.
Another good resource for parents is:
http://childparenting.miningco.com/cs/learningproblems/a/wisciii.htm

6.13 Self-Assessment Questions

1. Compare and contrast norm-referenced and criterion-referenced score interpretations.
2. Describe main characteristics of normal curve.
3. In which situations the use of norm-referenced or criterion-referenced interpretations will be most appropriate? Explain with examples.
4. Transform the following raw scores to the specified standard score formats. The raw scores distribution has a mean of 70 and a standard deviation of 10.
   a. Raw score = 80  z-score =  T-score =
   b. Raw score = 70  z-score =  T-score =
c. Raw score = 55  z-score = T-score =

5. Convert the following z-scores to T-scores.
   a. z-score = 1.5  T-scores =
   b. z-score = -1.5  T-scores =
   c. z-score = 2.5  T-scores =
   d. z-score = -2.0  T-scores =
   e. z-score = -1.70  T-scores =

6. A student makes a raw score of 65 on a Math test having a mean of 50 and standard deviation of 10, but he makes raw score of 80 on an English test having a mean of 75 and standard deviation of 15. What are his z scores on the two tests? Is he better in Math or English?

6.14 Suggested Readings

6.15 Web-based Material
Unit–7

THE RELIABILITY OF MEASUREMENT METHODS
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Introduction of the Unit</td>
<td>105</td>
</tr>
<tr>
<td>7.2</td>
<td>Objectives of the Unit</td>
<td>105</td>
</tr>
<tr>
<td>7.3</td>
<td>Reliability</td>
<td>106</td>
</tr>
<tr>
<td>7.4</td>
<td>Errors of Measurement</td>
<td>107</td>
</tr>
<tr>
<td>7.5</td>
<td>Sources of Measurement Error</td>
<td>108</td>
</tr>
<tr>
<td>7.6</td>
<td>Forms of Reliability</td>
<td>109</td>
</tr>
<tr>
<td>7.7</td>
<td>Factors Affecting the Reliability Estimates</td>
<td>111</td>
</tr>
<tr>
<td>7.8</td>
<td>Improving the Test Reliability</td>
<td>112</td>
</tr>
<tr>
<td>7.9</td>
<td>The Standard Error of Measurement</td>
<td>112</td>
</tr>
<tr>
<td>7.10</td>
<td>Self Assessment Questions</td>
<td>113</td>
</tr>
<tr>
<td>7.11</td>
<td>Additional Readings</td>
<td>114</td>
</tr>
<tr>
<td>7.12</td>
<td>Web Based References</td>
<td>114</td>
</tr>
</tbody>
</table>
7.1 Introduction of the Unit

This unit will provide an overview of different types of reliability evidence. Reliability refers to the consistency in test scores. If a test or assessment procedure produces consistent measurements, its scores are reliable. For making reliable and good decisions, we need reliable information. In everyday conversation, reliability is a synonym for consistency or dependability. It is important for us as users of test and consumers of information about tests to know how reliable tests and other measurement procedures are. By estimating the reliability of assessment results, we get an indication of how much confidence we can place in them. If we have highly reliable and valid information, it is probable that we can use that information to make better decisions. If the results are unreliable they are of little value to us. A reliability coefficient is an index of reliability. It is a proportion that indicates the ratio between the true score variance on a test and the total variance. Although reliability coefficients are useful when comparing the reliability of different tests, the standard error of measurement (SEM) is more useful when interpreting test scores. The SEM is an index of the amount of error in test scores and is used in calculating confidence interval within which we expect the true scores to fall. An advantage of the SEM and the use of confidence intervals is that they serve to remind us that the measurement is present in all scores. In this we will explore different kinds of reliability coefficients, including those for measuring test-retest reliability, alternate-form reliability, internal-consistency reliability and inter-rater reliability. There are numerous factors that affect reliability. The time the test is administered, the specific set of questions included on the test, destruction due to external or internal events, and the person grading the test are just a few of these factors. In this unit you will learn to take many of the sources of unreliability into account when selecting or developing assessment and evaluating scores.

This unit will also provide some guidelines for selecting the type of reliability estimate most appropriate for specific assessment procedures, some guidelines for evaluating reliability coefficient, and some suggestions on improving the reliability of measurement.

7.2 Objectives of the Unit

1. Define and explain the importance of reliability in educational assessment.
2. Define and explain the concept of measurement error.
3. Explain classical test theory and its importance to educational assessment.
4. Describe the major sources of measurement error.
5. Identify the major methods for estimating reliability and describe how these analyses are performed.
6. Analyze the factors that should be considered when evaluating the magnitude of reliability coefficient.
7. Evaluate the standard error of measurement (SEM) and explain its importance.
8. Explain how confidence intervals are calculated and used in educational and psychological assessment.
9. Describe and apply procedures for estimating the reliability of classroom tests.
7.3 Reliability

Most dictionaries define reliability in terms of dependability, trustworthiness, or having a high degree of confidence in something. By degree of confidence we mean that the test yields similar results when it is repeated over a short period of time or when a different form is used. It is also extended to such concepts as stability and consistency. Consistency is the most important attribute of a measurement instrument. One of the important things to be remembered is that consistency is considered to be a characteristic of scores or assessment results, not tests themselves. Figure 7.1 shows conceptual understanding of reliability. It is not merely a word rather the understanding of reliability based on systematic body of knowledge. The following example will clarify the concept: a teacher administers a 20-item math test to assess students’ skill in multiplying three digits numbers. If the same test had been administered to the same class after three hours, would Naila’s score on the test have been the same? Would the students have received the same scores if another teacher had graded the test? Did the loud noise outside disturbed Naila when she was re-administered the same test and she could not attempt the item correctly? All of these questions involve issues of reliability. Researchers are concerned about the consistency of their measuring devices when they attempt to measure such complex traits as performance,
satisfaction, motivation, attitude, and the like. A test is reliable to the extent that the scores made by an individual remain nearly the same in repeated measurements.

Test reliability can be expressed numerically. A coefficient of .80 or higher indicates high reliability, .40 to .79 fair reliability, and less than .40 low reliability.

7.4 Errors of Measurement

Some degree of error is inherent in all measurement. Variations in measurement can be referred to as measurement error. A score on ability test is presumed to reflect both testtaker’s true score on the ability being measured as well as error. Even in some situations we believe that measurement is exact, but yet some error is present. In its broadest sense, ‘error’ refers to the component of observed score on an ability test that does have to do with testtaker’s ability. If we use X to represent an observed score, T to represent a true score, and E to represent error, then the fact that an observed score equals the true score then the fact that an observed score equals the true score plus error may be expressed as follows:

In mathematical terms:

\[ X_i = T + E \]

Obtained score = True score + Error score

\[ T = X_i - E \]

The true score (T) is a score that would be obtained if there were no errors or the error score (E).

Here \( X_i \) is used to represent the obtained or observed score of an individual. It is the score the student received on the test. The symbol T is used to represent an individual’s true score and reflects the student’s true skills, abilities, knowledge, attitude, or whatever the test measures. Finally, E represents measurement error. Measurement error limits the extent to which test results can be generalized, and reduces the confidence we have in test results. Due to the presence of measurement error we can never know with absolute confidence what the true score is.

There are two types of errors that appear in test scores: (Reynolds, Livingstone, and Willson (2006):

- Random Error
- Systematic Error

Random Error is the unexplained difference between the true score (T) and the obtained score (\( X_i \)). Theoretically if the test could be given an infinite number of times and an average score calculated from those administrations, the average test score would equal the true test score. Furthermore, if average error from all those administrations were calculated, it would be equal to zero. One way to reduce random error is to find out what is causing the error and eliminate or control it. When a single source of error can be identified, we can often calculate the amount and direction of the error and take its source into account when we interpret test scores. Such known error is identified as ‘Systematic Error’. If you know the scale in your kitchen regularly adds one pound then you can
simply subtract one pound whatever the scale says. The error your scale makes is predictable and systematic.

There is also example of few thermometers that give two or more degrees higher than the actual temperature, then the error is systematic and predictable and can be taken into account. Systematic error though sometimes difficult to identify, however, taking test again and again and order effects can add systematic as well as random error to test scores. If students learn the answer to a question in the first administration (practice effect) and they can derive the answer from the previous question (order effect), the every one will get the question right. Such occurrences systematically raise test scores. In such cases the error can be eliminated by removing the items from the test or replacing it with another that will be unaffected by practice or order.

As a matter of fact random error lowers the reliability of a test whereas systematic error has no effect on reliability of test score. However, if we have information about the reliability of measurement, we can establish intervals around an obtained score and calculate the probability that the true score will fall within the interval specified. This topic has been discussed in detail in section 7.8. Before proceeding to the sources of error of measurement read the provided material.

*See Article 7.1 in folder web material Unit-7 ‘Measures of Reliability’ on CD for explanation of error of measurement.*

### 7.5 Sources of Measurement Error

So far you have learned about reliability, errors of measurement, sources of error of measurement and now you as an educational professional you will be able to identify not all but few sources of error of measurement. Identification of measurement errors will enable you to minimize their effect on test scores to some extent. Reliable assessment results are free from measurement errors whereas less reliable results are influenced to larger degree by measurement error. A number of factors may introduce error into the test scores.

Since we estimate reliability by correlating the scores obtained by the same individuals on different occasions or with different sets of equivalent items therefore, these procedures require two administrations of a test, namely the test-retest reliability and the equivalent-forms reliability procedures. Another form of reliability is known as ‘internal consistency’ and as the label suggests, it focuses on the consistency of test’s internal elements, its test items. This form of reliability requires a single administration of a test. The test is split into two halves and correlates the individual’s scores on the two forms. Often internal-consistency measures of reliability do not require splitting the test into halves and scoring each half separately. These procedures assess the inter-items consistency, or homogeneity of the items. Two commonly used indexes of homogeneity are the Kuder-Richardson formula 20 and Cronbach’s coefficient alpha. Reliability of a test is a function of (1) length of the test, (2) group heterogeneity, (3) procedure used for its estimation, and (4) nature of the variable being measured.
7.6 Forms of Reliability

Now you have a basic understanding of reliability and how it is measured, let’s discuss five different types of reliability: test-retest reliability, alternate-forms reliability, split-half reliability, and scorer or inter-rater reliability. Each type provides a measure of consistency, but the various types of reliability are used in different situations.

1. **Test-Retest Reliability**
   One of the most commonly used methods of establishing reliability is to administer the same test to the subjects. The correlation between two sets of scores is calculated. The resulting correlation coefficient exhibits the reliability of two sets of scores. If the test is reliable, the two scores for each individual will be same, and the resulting correlation coefficient will be high i.e. close to +1.00. This measure of reliability assesses the stability of test overtime. Naturally some error will be present in each measurement therefore correlation coefficient cannot be perfect i.e. +1.00 but could be .80 or higher. A number of objections have been raised to the test-retest method. If the same items are used on both tests, the respondents’ answers on the second test may be influenced by their memory of first test or by discussions about the items with classmates or teachers between tests. If the interval between tests is too short, memorization is a factor. If the interval is too long, the score may change as a result of learning. The two tests conditions might also differ. Lack of interest on student’s part during one of the test situations, a change in student’s health and diet, a change in the mood of the student, or test administrator may affect the scores.

2. **Alternate-Forms (Parallel forms) Reliability**
   To overcome the problem of administering the same test twice to sample in test-retest reliability method, the alternate may be used. In the alternate-forms (also called equivalent or parallel forms) method of establishing reliability, two identical equivalent tests are constructed. These tests are similar in kind of content, mental processes required, number of items, difficulty and all other respects. The two tests are administered at the same time to same group one after the other the time interval between the two tests should be as little as possible. The correlation between scores on the two sets provides a good estimate of reliability. One drawback to this method is that parallel forms are not always available, especially with teacher made tests. The two forms are not always equivalent and may differ in difficulty.

3. **Internal Consistency Coefficients**
   Developing a parallel form of test is often difficult and time consuming and also expensive. Therefore, the experts in this field have devised other methods. These include Spearman’s split-half method, the Kuder-Richardson formulas, and Cronbach’s coefficient alpha.

   (i) **Split-Half Reliability**
   The difficulties associated with test-retest and parallel forms methods have lead to the development of the split-half reliability method. A single test is split into reasonably equivalent halves, and these two subsets are used to determine the reliability coefficients. One common method of splitting a test is to score the even-numbered and odd-numbered items separately. Splitting a test in two half means that the reliability scores are
determined by half the number of items. Too few items in calculation can lead to greater
distortion and more chance effects. In order to calculate the reliability of the total test
Spearman-Brown Prophecy Formula is applied:
Reliability on full test = $2 \times (\text{reliability of half test})$
1+ (reliability of half test)

Then the estimated reliability of the entire test in mathematical term is:

$$r_{11} = \frac{2r_{12}}{1 + r_{12}}$$

To demonstrate the use of formula, assume that the correlation between total scores on
the odd-numbered items and total scores on even-numbered items of a test is .85.

$$r_{11} = \frac{2 \times .85}{1 + .85} = .91$$

(ii) Kuder-Richardson Method.
A test can be divided in many different ways into two halves containing equal
numbers of items. Because each way may result in a some what different value $r_{11}$,
therefore, best strategy for estimation may not be cleared. One solution to overcome this
problem is to calculate the average of the reliability coefficients obtained from all half
splits as the overall reliability coefficient. This can be done by the following formula:

$$r_{11} = \frac{k \left[ 1 - \sum p_i (1 - p_i) / s^2 \right]}{k - 1}$$

(iii) Kuder-Richardson (K-R) formulas 20 and 21, respectively. Formulas (iii) is
based on the assumption that all items are of equal difficulty and it is easier to calculate.
To demonstrate application of formula (iii), assume that a test containing 100 items has a
mean of 60 and a variance of 20. Then $r_{11} = \frac{100 - 60}{100 - 60}/20 = .80$

Coefficient Alpha.

Coefficient alpha is defined as :

$$\alpha = \frac{k \left(1 - \sum s_i^2 / s^2 \right)}{k - 1}$$
Where $k$ is the number of items, $s^2_i$ is the variance of scores on item $i$, and $s^2_t$ is the variance of total test scores. The Kuder-Richardson formulas are applicable when test items are scored 0 or 1, but coefficient alpha is a general formula for estimating the reliability of a test consisting of items on which different scoring weights may be assigned to different responses.

4. **Inter Rater Reliability**

Inter-rater reliability refers to the concern that a student's score may vary from rater to rater. In evaluating scores involving subjective scorer judgment, it is important to know the extent to which different scorers agree on ratings or other numerical values given to the responses of different examinees and items. Students often criticize exams in which their score appears to be based on the subjective judgment of their instructor. For example, one manner in which to analyze an essay exam is to read through the students' responses and make judgments as to the quality of the students' written products. Without set criteria to guide the rating process, two independent raters may not assign the same score to a given response. Each rater has his or her own evaluation criteria. Scoring rubrics respond to this concern by formalizing the criteria at each score level. The descriptions of the score levels are used to guide the evaluation process. Although scoring rubrics do not completely eliminate variations between raters, a well-designed scoring rubric can reduce the occurrence of these discrepancies.

*Read following articles in web material folder Unit-7 on CD for further explanation and calculation of different forms of reliability:*

*Article 7.2 ‘Reliability’*

*Article 7.3 ‘Reliability 1’*

*Article 7.4 ‘Basic Concepts in Reliability’*

<table>
<thead>
<tr>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepare two equivalent forms of tests of 30 items for 10th class, administer them and compute reliability. Give your comments on results.</td>
</tr>
</tbody>
</table>

7.7 **Factors Affecting the Reliability Estimates**

The reliability coefficient may be affected by the following factors:

1. **Number of Items in the Test**

The number of items in an assessment is an important factor that influences reliability. The greater the number of items, the greater is the reliability. This is because a large number of items provide a more adequate sample of the behavior being measured and the scores are apt to be less distorted by chance factors such as guessing.

2. **Number of Students**

The number of students also makes a difference, i.e. the higher the number of students, the stronger the reliability. The range of talent, achievement or ability of the pupils on whom the reliability is based has a direct effect on the reliability coefficient. The larger variability in the group increases the reliability coefficient.
3. **Spread of Scores**
   Reliability coefficient is directly influenced by the spread of scores in the group tested. Other things being equal, the larger the scores, the higher the estimate of reliability will be.

4. **Testing Conditions**
   The conditions of testing and scoring the test may raise or lower the reliability of the given test.

5. **Difficulty of Test**
   Tests that are too easy or too difficult will tend to reduce the reliability. This is because both easy and difficult tests result in restricted spread of scores. For easy test scores are close together at the top while for difficult tests, scores are grouped at the bottom end. Carefully constructed items will improve the reliability of the test.

   *Read Article 7.4 in web material folder unit-7 on CD for clarification of factors affecting reliability.*

7.8 **Improving the Test Reliability**

   The test reliability can be improved by the following factors:
   
   1. *Increased number of test items.* Reliability is higher when the numbers of items is increased, because the test involves a large sample of material covered.
   2. *Homogeneity of student group.* Reliability is higher when test scores are spread over a range of abilities. Measurement errors are smaller than from a group that is more homogeneous in ability.
   3. *Moderate item difficulty.* Reliability is increased when the test items are of moderate difficulty because this spreads the scores over a greater range than a test composed mainly of difficult or easy items.
   4. *Objective scoring.* Reliability is greater when tests can be scored objectively. With subjective scoring the same responses can be scored differently on different occasions, even if the scorer is the same person.
   5. *Limited Time.* A test in which speed is a factor is more reliable than a test that all students can complete in the time available.

7.9 **The Standard Error of Measurement**

   As you have learned in section 7.5 of this unit that every test has some degree of error; the higher the value of measurement error the lower the reliability will be. Though we measure the reliability directly from the scores of the students but we cannot know what type error has affected a student’s score. Therefore, we estimate the degree of error that is probable, given the reliability of the test. This degree of error is estimated mathematically and is reported as the **standard error of measurement (SEM).**

   SEM is determined by the following formula:
   
   \[
   SEM = s_t \sqrt{1 - r_{tt}}
   \]
Where

\( s_t \) - standard deviation of test

\( r_{tt} \) - reliability of test

If a student took a test many times, the resulting scores would look like a normal distribution. That is, sometimes student would get higher scores and sometimes lower scores. If we assume that student’s true score is the mean of this hypothetical distribution and the standard error of measurement corresponds to the standard deviation of this distribution, then we can use it as a starting point for estimating the actual true score. From what we know about normal curve and standard deviation, 68 percent of the time the true score would be between one standard deviation of the student’s normal curve of many testings, and 96 percent of the time the true score would fall within two standard deviations of this distribution.

For example, if a student’s score on a test is 70 and the standard error of measurement is 3, then we would interpret the student’s true performance, with 68 percent confidence, to be 70 + 3; and with 96 percent confidence, to be 70 + 6. In other words SEM creates an interval, and it is within this interval that we can be confident that the student’s true score lies.

For further explanation see article 7.5 ‘Reliability 2’ in web material folder-7 on CD.

### 7.10 Self-Assessment Questions

1. What is the importance of reliability in educational measurement and assessment?
2. What is measurement error? What are some major sources of error in measurement?
3. How can different types of reliability be used in various educational settings?
4. Which factors can affect the reliability of a test?
5. Set the 95% confidence interval around the true score of a student who obtained a raw score of 50 on a test that has a reliability of 0.90 and a standard deviation of 10.
6. Calculate both split-half (odd-even) and Kuder-Richardson reliability (formulas 20 and 21) coefficients on the following achievement test, where 1 indicates a right answer and 0 a wrong answer.

<table>
<thead>
<tr>
<th>Items</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

The (\( \bar{X} \)) of the total score is 5.30, and the variance is (\( s^2 \)) is 5.21
7. If standard deviation of a test is 12 and reliability coefficient is 0.84, what is the standard error of measurement of the test? If a student has raw score of 36, he will be placed, with 96% confidence, in the group of students whose true score fall in what limits?

7.11 Additional Readings


7.12 Web based References on CD


114
THE VALIDITY OF MEASUREMENT METHODS
CONTENTS

8.1 Introduction of the Unit ................................................................. 119
8.2 Objectives of the Unit .................................................................. 119
8.3 Validity ....................................................................................... 119
8.4 Types of Validity ........................................................................ 120
8.5 Factors Affecting Validity ........................................................... 122
8.6 Relationship between Reliability and Validity ......................... 123
8.7 Unitary Concept of Validity .......................................................... 124
8.8 Interpreting Validity Coefficients .............................................. 126
8.9 Convergent and Discriminant Evidence of validity .............. 127
8.10 Practical Strategies for Teachers ............................................... 128
8.11 Self Assessment Questions ......................................................... 130
8.12 Suggested Books ....................................................................... 130
8.13 Web Based References ............................................................... 130
8.1 Introduction of the Unit

The second most important characteristic of good assessment is its ability to help the teacher to make a correct decision. This characteristic is called validity and is the key to obtaining high-quality and meaningful assessment decisions. Without validity, assessment data will not lead to correct decisions. Validity is, therefore, the most fundamental consideration in developing and evaluating tests.

In the previous unit you have learned the concept of the reliability of measurement of educational tools. In this context, reliability refers to accuracy and consistency in test scores. Now this unit turns your attention to validity, another important psychometric property. In simpler terms validity refers to the appropriateness and accuracy of the interpretations of test scores. Traditionally validity has been defined as the extent to which a test measures what it was designed to measure. The methods by which validity may be determined include: (i) analyzing the content of the test, (ii) computing the correlation between scores on the test and those on the criterion of interest, (iii) investigating the particular psychological characteristics or constructs measured by the tests. You will also learn about the characteristics, types, uses of validity and major threats to validity. One of the major concerns of educationist is to predict the future academic and non academic behaviour of student from a number of factors. Therefore one objective of this unit deals with predictive validity. And what are the practical guidelines for the application of validity in classroom perspectives is also an important topic to discuss.

8.2 Objectives of the Unit

After completing this unit, allied material and related research articles you will be able to:
1. Define validity and explain the importance of validity of educational measures.
2. Explain the type of validity to be used for a specific measure.
3. Identify the validity evidence required for each type of validity.
4. Explain relationship between reliability and validity.
5. Explain the factors affecting validity.
6. Explain the unitary concept of validity.
7. Describe and explain standard error of estimate.

8.3 Validity

McMillan (2007) defined validity in technical terms as “the soundness, trustworthiness, or legitimacy of the claims or inferences that are made on the basis of obtained scores” (p.64).

This technical term describes the usefulness of the interpretations of test scores. In simple words valid interpretations help us to make better decisions while invalid interpretations do not. It implies that for making valid decisions a valid and reasonable interpretation of scores is very important.
In general when we ask whether a test is valid, we are asking does the test measure what it claims to measure. It is crucial to know the answer to this question since in education, classroom tests are used to test a student’s knowledge of the content and to determine grades. If a teacher does not know whether a test really measures what it claims to measure then he/she cannot know whether the decision made on the basis of test scores are right. Central to the validity of assessment information is the question, does the evidence I have gathered tell me about the characteristic I wish to judge.

8.4 Types of Validity

1. Content validity
   Content validity is typically based on professional judgments about the appropriateness of the test content. It involves how adequately the test samples the content area of the identified construct. A classroom teacher can determine the content validity by carefully examining a test to determine the questions/items are representative of the material that has been or should be covered by the test. In other words, is the content of the test relevant and representative of the content domain? When we speak about representativeness we are actually asking the questions that have been chosen to sample and represent the full domain of questions, because, practically it is not possible to include every possible question that could be asked. If a test is made up of questions that represent all aspects of the attribute being measured, then the test is said to show evidence of content validity. The content validity of an achievement test is evaluated by analyzing the composition of the test to determine the extent to which it represents the objectives of instruction. One way of establishing content validity is to compare the test’s content with an outline or table of specification concerning the subject matter to be covered by the test. If subject matter experts agree that a test looks and acts like an instrument that was designed to measure whatever it is supposed to, then it is said to possess content validity. Experts’ judgement concerning what items to include is necessary from the very beginning of the test construction process.

2. Criterion-related validity
   Criterion validity involves examining the relationships between the test and external variables that are thought to be direct measures of the construct. In its typical meaning criterion-related validity is a judgment regarding how adequately a test score can be used to infer an individual’s standing on some measure of interest the measure of interest being the criterion. A criterion may be broadly defined “as the standard against which a test or test score is evaluated. Operationally a criterion can almost be anything” (Cohen and Swerdlik, 2002; p.161). No specific rules have been defined for setting a criterion. The individuals or groups behaviour, a test score, attitudes, amount of time, rating, learning environment, and so on could serve as criterion. Cohen and Swerdlik state: “although a criterion can almost be anything, it should be reliable, relevant, valid, and uncontaminated” (p.160).

We determine the criterion related validity of a test by correlating the scores on a test with scores on another measure of performance, the test shows evidence of criterion related validity. There are two types of criterion related validity.
(i) Predictive
(ii) Concurrent

(i) **Predictive Validity**: If test scores are obtained at one time and criterion measure obtained at a future time or has been taken place after some training, experience, therapy, medication, etc. In simple words for determining the predictive validity intervention of time is required for the criterion to take place. The criterion measure obtained at future time shows the predictive validity of the test. Measure of the relationship between scores of admission test conducted by National Testing Service (NTS) for admission in MPhil program and first semester GPA may provide evidence of **predictive validity** of the admission test. An employment test has **predictive validity** if high scores are achieved by those who later do well on the job.

The judgment of criterion related validity (both predictive and concurrent) is based on ‘validity of coefficient’ which is a statistical evidence. The validity coefficient is a correlation coefficient that provides a measure of relationship between test score and scores on the criterion measure. Pearson correlation coefficient formula is used for calculating the validity between two sets of scores. In cases where the data is not interval or ratio, or sample size is small, and scores are not distributed normally then Spearman rho Rank-order correlation coefficient would be employed.

(ii) **Concurrent Validity** An employment test has **concurrent validity** if the scores on the test correlate closely, at the same point in time, with the score on some well established measure of job performance. It implies that is score are obtained at about the same time the criterion measures are obtained, measures of the relationship between the test scores and the criterion provide evidence of concurrent validity. Studies of criterion-related validity empirically examine the relationships between test scores and criterion scores using correlation or regression analyses.

3. **Construct validity**
   
   Construct validity involves an integration of evidence that relates to the meaning or interpretation of test scores. The greater concern with respect to personality tests is construct validity. The construct validity of a psychological instrument refers to the extent to which the instrument measures a particular construct or psychological concept such as, anxiety, extroversion-introversion, self-concept, self efficacy, achievement motivation, emotional adjustment, job satisfaction, creativity, etc. Construct validity is not determined in a single way. It involves a network of investigation and other procedures designed to determine whether an assessment instrument that presumably measures a certain personality variable is doing so. This evidence can be collected using a wide variety of research strategies and designs. There are several different ways of providing evidence of construct validity. We determine validity of test by examining whether the test’s relationship to other information coincides with some theory. One way to evaluate validity is to examine the scores on a test with some expected behaviour. Among the sources of evidence for the construct validity of a test are the following: (Aiken, 2003)
   
   (a) Expert’s judgment that the content of the test pertains to the construct of interest.
   (b) Analysis of internal consistency of the test.
Studies, in both experimentally contrived and naturally accruing groups, of the relationships between test scores and other variables on which the groups differ.

(d) Correlation of scores on the test with scores on other tests and variables with which it is expected to have a certain relationship, followed by factor analysis of these correlations.

(e) Questioning examinees or raters in detail about their responses to a test or rating scale in order to reveal the specific mental processes involved in responding to the items.

The process of construct involves hypothesizing variable, indicators or conditions responsible for explaining the behaviour. From these hypotheses arises a tentative theory about the nature of the construct the test was designed to measure. Constructs are directly unobservable, presupposed underlying traits that a test developer may invoke to describe test behaviour or criterion performance.

4. **Face Validity**

   It is not recognised as one of the primary types of validity, test developers and users often speak of face validity in the extent to which the test taker perceives that the test measures what it is supposed to measure.

   See following articles in web material folder unit-8 on CD for further detail:
   - Article 8.1 ‘Validity1’ (web reference)
   - Article 8.2 ‘Validity 2’
   - Article 8.3 ‘Measurement of Validity Types’.

8.5 **Factors Affecting Validity**

   In addition to the characteristics of the test itself, factors external to the test can impact the validity of the interpretations of the results. Linn and Gronlund (2000) identified numerous factors external to the test that can influence validity. They highlighted the following factors:

1. **Instructional procedures**

   In educational tests, in addition to the content of the test influencing validity, the way the material is presented can influence validity. For example, consider a test of critical thinking skills; if the students were coached and given solutions to the particular problems included on a test, validity would be compromised. This is a potential problem when teachers “teach the test.”

2. **Test administration and scoring procedures**

   Deviations from the standard administrative scoring procedures can undermine validity. In terms of administration, failure to provide the appropriate instructions or follow strict time limits can lower validity. In terms of scoring, unreliable or biased scoring can lower validity.

3. **Student characteristics:**

   Any personal factors that restrict or alter the examinees’ responses in the testing situation can undermine validity. For example, if an examinee experiences high levels of
test anxiety or is not motivated to put forth a reasonable effort, the results may be distorted.

Additionally, the validity of norm-referenced interpretations of performance on a test is influenced by the appropriateness of the reference group (AERA et al., 1999). As these examples illustrate, a multitude of factors can influence the validity of assessment-based interpretations. Due to the cumulative influence of these factors, validity is not an all-or-none concept. Rather, it exists on a continuum and we usually refer to degrees of validity or to the relative validity of the interpretation(s) given to a particular measurement.

8.6 Relationship between Reliability and Validity

In the preceding chapter we addressed the issue of the reliability of measurement. Reliability refers to the stability or consistency of test scores and reflects the amount of random measurement error present. Reliability is a necessary but insufficient condition for validity. A test that does not produce reliable scores cannot produce valid interpretations. However, no matter how reliable measurement is, it is not a guarantee of validity. From our discussion of reliability you will remember that obtained score variance is composed of two components: true score variance and error variance. Only true score variance is reliable, and only true score variance can be systematically related to any construct the test is designed to measure. If reliability is equal to zero, then the true score variance component must also be equal to zero, leaving our obtained score to be composed only of error, that is, random variations in responses. Thus, without reliability there can be no validity.

Although low reliability limits validity, high reliability does not ensure validity. It is entirely possible that a test can produce reliable scores but inferences based on the test scores can be completely invalid. Consider the following rather silly example involving head circumference. If we use some care we can measure the circumference of our students’ heads in a reliable and consistent manner. In other words, the measurement is reliable. However, if we considered head circumference to be an index of intelligence, our inferences would not be valid. The measurement of head circumference is still reliable, but when interpreted as a measure of intelligence it would result in invalid inferences.

A more relevant example can be seen in the various Wechsler intelligence scales. These scales have been shown to produce highly reliable scores on a Verbal Scale and a Performance Scale. There is also a rather substantial body of research demonstrating these scores are interpreted appropriately as reflecting types of intelligence. However, some psychologists have drawn the inference that score differences between the Verbal Scale and the Performance Scale indicate some fundamental information about personality and even forms of psychopathology. For example, one author argued that a person who, on the Wechsler scales, scores higher on the Verbal Scale relative to the Performance Scales is highly likely to have an obsessive-compulsive disorder! There is no evidence or research to support such an interpretation and, in fact, a large percentage of the population of the United States score higher on the Verbal Scale relative to the Performance Scale on each of the various Wechsler scales. Thus, when the scores are
themselves highly reliable and *some* interpretations are highly valid (the Wechsler scales measure intelligence), other interpretations wholly lack validity despite the presence if high reliability.

### 8.7 Unitary Concept of Validity

The above discussed traditional types of validity has been widely accepted and used by researchers, authors, teachers and students. Measurement professionals in the 1970s and 1980s, re-conceptualized validity and introduced it as a **unitary concept**. The traditional types of validity were named as content-related evidence of validity, criterion-related evidence of validity, and construct-related evidence of validity. Reynolds, Livingston, and Wilson (2006) reported following evidences for measuring validity:

(i) **Evidence based in test content**
   This category includes evidence derived from an analysis if the test content includes the type of questions or tasks included in the test and administration and scoring guidelines.

(ii) **Evidence based on relations to other variables**
   This category includes evidence based in an examination of the relationships between test performance and external variables or criteria.

(iii) **Evidence based on internal structure**
   This category includes evidence regarding relationships among test items and components.

(iv) **Evidence based on response processes**
   This category of evidence is derived from an analysis of the processes engaged in by the examinee or examiner.

(v) **Evidence based on consequences of testing**
   This category includes evidence based on an examination of the intended and unintended consequences of testing.
   Reynolds (2002) reported that currently test manuals and other test-related documents are adopting this new nomenclature.

(vi) **Face validity and content validity**
   Students are generally confused while differentiating between content validity and face validity. At this stage it is appropriate to highlight the distinction between content and face validity. **Face validity** is technically not a form of validity at all as has been believed but instead refers to a test “appearing” to measure what it is designed to measure. That is, does the test appear valid to untrained individuals who take, administer, or examine the test? Face validity really has nothing to do with what a test actually measures, just what it *appears to measure*. For example, does a test of achievement look like general public expects an achievement test to look like? Does a test of intelligence
look like the general public expects and intelligence test to look like? Naturally, the face validity of a test is closely tied to the content of a test. In terms of face validity, when untrained individuals inspect a test they are typically looking to see whether the items on the test are what they expect. For example, are the items on an achievement test of the type they expect to find on an achievement test? Are the items on an intelligence test of the type they expect to find on an intelligence test? Whereas content based evidence of validity is acquired through systematic appearance of a test. A test can appear “face valid” to the general public, but not hold up under the systematic scrutiny involved in a technical analysis of the test content.

<table>
<thead>
<tr>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collect some old question papers of examination boards and find the evidence of validity based on test content.</td>
</tr>
</tbody>
</table>

(vii) Evidence Based on Relations to Other Variables

Validity evidence can also be collected by finding the relationships between test scores and other factors. The followings are the two approaches:

(a) Test-Criterion Evidence

Many tests are designed to predict performance on some variable that is typically referred to as criterion. The criterion can be academic performance as reflected by the grade point average (GPA), job performance as measured by a supervisor’s ratings or anything else that is of importance to the user of the test, attitude, etc. as has been mentioned earlier that there are two types of validity concerns for collecting criterion evidence: (i) predictive studies and (ii) concurrent studies. In a predictive study the test is administered, there is an intervening time interval, and then the criterion is measured. In a concurrent study the test is administered and the criterion is measured at the same time.

To illustrate these two approaches we will consider the GAT Test. The GAT is designed to measure general ability for admission into MPhil and PhD programmes. To complete a predictive study, if we administer the GAT before admission into the programme, wait until the students have completed the programme, and then find out the relationship between the GAT scores and the GPA. The GAT in this reference is predictor and GPA criterion. Correlation coefficient is often used to examine the relationship between a predictor and a criterion, and in this context the correlation coefficient is referred to as a validity coefficient. In predictive studies there is a time interval between the predictor test and the criterion; in a concurrent study there is no time interval. But in concurrent study the relationship between the GAT and GPA, the researcher might administer the GAT to a group of students completing the programme and then correlate their GAT scores with their GPAs.

Student may ask: “Which type of study, predictive or concurrent, is best?” In education and other settings educationists are interested in making predictions about the students’ performance. In the example of GAT the answer to the question that which students will do well in MPhil and which will not. Inherent in this question is the passage
of time. We want to administer a test before students complete coursework that will help predict the likelihood of their success in the programme.

In concurrent study the intervention of time between one and other characteristics in not involved. Educationist, researchers, psychologists prefer prediction studies as the ultimate goal of assessment. However, predictive studies take considerable time to complete and can be extremely expensive. As a result, although predictive studies might be preferable from a technical perspective, for practical reasons test developers and researchers might adopt a concurrent strategy to save time and/or money. In some situations this is less than optimal and you should be cautious when evaluating the results. However, in certain situations concurrent studies are the preferred approach. Concurrent studies clearly are appropriate when the goal of the test is to determine current status of the examinee as opposed to predicting future outcome.

(b) How to select a criterion

In both predictive and concurrent studies, it is important that the criterion itself be reliable and valid. As noted earlier, reliability is a prerequisite for validity. If a measure is not reliable, whether it is a predictor test or a criterion measure, it cannot be valid. At the same time, reliability does not ensure validity. Therefore, we need to select criterion measures that are also valid. In our example of using SAT to predict freshman GPA, we consider our criterion, GPA, to be valid measure of success in college. In a concurrent study examining the ability of a test to diagnose psychopathology, the criterion might be the diagnosis provided by an extensive clinical assessment involving a combination of clinical interviews, behavioural observations, and psychometric testing. Optimally the criterion should be viewed as the “gold standard”, the best existing measure of the construct of interest.

8.8 Interpreting Validity Coefficients

Predictive and concurrent validity studies examine the relationship between a test and a criterion and the results are often reported in terms of a validity coefficient. At this point it is reasonable to ask, “How large should validity coefficients be?” For example, should we expect validity coefficients .80 is acceptable for showing the relationship between GAT and GPA for MPhil in our example. If a test provides information that helps predict criterion performance better than any other existing predictor, the test may be useful even if its validity coefficients are relatively small. As a result, testing experts avoid specifying a minimum coefficient size that is acceptable for validity coefficients (Reynolds, Livingstone, and Willson, 2006)

Though there is no minimum size for acceptable validity coefficients, but certain techniques can be used to evaluate the usefulness of test scores for prediction purposes. Linear regression is a mathematical procedure that allows you to predict values of one variable with given information on another variable. In the context of validity analysis, linear regression allows you to predict criterion performance based on predictor test scores. When using linear regression, a statistic called the standard error of estimate is used to describe the amount of prediction error due to the imperfect validity of the test. The standard error of estimate is the standard deviation of prediction errors around the predicted score. The formula
for the standard error of estimate is similar to standard error of measurement that you have learned in unit-7. The formula for standard error of estimate is:

\[ s_{est} = s \sqrt{1 - r^2} \]

Where \( s \) is the standard deviation of the criterion scores and \( r \) is the product-moment correlation coefficient between the predictor and criterion. Assume for example, that the standard deviation of achievement score (criterion measure) is 10 and the correlation between achievement score and time spent on studying (predictor measure) is .80, then:

\[ s_{est} = 10 \sqrt{1 - (.80)^2} \]

\[ = 10 \sqrt{3.6} \approx 6.0 \]

If a student’s predicted criterion score is 70, the chances are 68 out of 100 that the student will obtain criterion score between 64 and 76 (\( y_{pred} + s_{est} \)), and 95 out of 100 that he or she will obtain score between 58 and 82.

See articles 8.2 ‘Validity 2’ and article 8.3 ‘Measures of Validity’ and article 8.4 ‘Measurement of validity types’ In web material folder Unit-8 on CD.

### 8.9 Convergent and Discriminant Evidence of Validity

Convergent and discriminant evidence of validity have traditionally been incorporated under the category of construct validity. A construct-validated instrument should have high correlations with other measures or methods of measuring the same construct (convergent validity), but low correlations with measures of different constructs (discriminant validity). Aiken (2003,p.100) suggested that evidence for the convergent and discriminant of a psychometric instrument can be obtained by comparing correlations with measures of:

i. The same construct using the same method.

ii. Different constructs using the same method.

iii. The same construct using different method.

iv. Different constructs using different methods

You are supposed to have an understanding of these validity evidence and learn how they can be obtained with the help of examples:

**Convergent evidence** of validity is obtained when you correlate a test with existing tests that measure similar constructs. In simple words evidence for the construct validity of a particular test may converge from a number of sources, such a other tests or measures designed to assess the same construct. If scores on the test undergoing validation correlate highly in the predicted direction with scores on older, published and already validated tests designed to measure the same or similar construct, this would be an example of convergent validity. Though convergent evidence for validity comes by correlating the scores of one measure with the scores on similar measure, it also can be obtained from correlations with measures purporting to measure related constructs. For measuring the evidence of convergent validity one might expect positively high correlation between new test and older test. For example, if you are developing a new
intelligence test you might elect to correlate scores on your new test with scores on the
Because the WISC – IV is a well-respected test of intelligence with considerable validity
evidence, a strong correlation between the WISC – IV and your intelligence test would
provide evidence that your test is actually measuring the construct of intelligence.

**Discriminant Evidence** of validity is obtained when you correlate a test with
existing tests that measure dissimilar constructs. A validity coefficient showing little
relationship between test scores or other variables with which scores on the test being
construct validated should not theoretically be correlated provides discriminant validity.
For example, if you were validating a test designed to measure anxiety, you might
correlate your anxiety scores with a measure of sensation-seeking. Because anxious
individuals do not typically engage in sensation-seeking behaviours, you would expect a
negative correlation between the measures. If your analyses produce the expected
negative correlations, this would support your hypothesis.

There is related, relatively sophisticated validation technique referred to as the **multitrait-
multimethod matrix** (Cohen, 2002). Here multitrait means ‘two or more traits’ and
multimethod means ‘two or more methods’ that results from correlating variables traits)
between and within methods. In our example mentioned in discriminant evidence, you examine
two or more traits (e.g. anxiety and sensation seeking) using two or more measurement methods
(e.g. self report and teacher rating). The researcher then examines the resulting correlation
matrix, comparing the actual relationships with **a priori** (i.e. pre-existing) predictions about the
relationships. In addition to revealing information about convergent and discriminant
relationships, this technique provides information about the influence of **common method
variance**. When two measures show an unexpected correlation due to similarity in their method
of measurement, we refer to this as method variance. Thus, the multitrait-multimethod matrix
allows one to determine what the test correlates with, what it does not correlate with and how
the method of measurement influences these relationships. This approach has considerable
technical and theoretical appeal, yet difficulty with implementation and interpretation has
limited its application to date.

**Factor Analysis:** The validity can be examining the structure of the construct it
measures. Factor analysis is a sophisticated statistical procedure used to determine the number
of conceptually distinct factors or dimensions underlying a test or battery of tests. Because
factor analysis is a fairly complicated technique, we will not go into detail about its calculation.

### 8.10 Validity: Practical Strategies for Teachers

Validity refers to the appropriateness or accuracy of the interpretation of
assessment results. The results of classroom assessments are used in many different ways
in today’s schools, and teachers need to consider the validity of all of these applications.
One of the most prominent uses of classroom assessment results is the summative
evaluation of students’ knowledge and skills in a specified content area (e.g. evaluating
mastery and assigning grades). In this context, Nitko (2001) developed a set of guidelines
for teacher for evaluating and improving the validity of the results of classroom
assessments. These guidelines include the examination of the followings:
1. **Content of the test**

   The examination of test content starts from the objectives and its relationship with content. As we discussed earlier in this unit, that the examination of content involves an examination of item relevance with objectives and content. Analysis of item relevance involves examining the individual test items and determining whether they reflect essential elements of the content domain. Content coverage involves examining the overall test and determining the degree to which the items cover the specified domain.

2. **Specific learning activities**

   What specific cognitive and behavioural activities were designed for students to learn? Simply we can ask, do the assessments require the students to engage in the types of cognitive process and behavioural activities that are specified in the learning objectives? For example, if your learning objectives involve problem solving, does the assessment require the students to engage in problem solving? In the previous unit you have studied about the taxonomy of student abilities. The broad range of these abilities is very helpful in determining the abilities and skills that are necessary to be learned and measured.

3. **Relations to other assessments**

   A teacher can find out the validity of his/her classroom test by examining the relationship between a given assessment and other sources of information about the students. These sources of information can be: observational data, classroom project, results of other tests etc. If the results are consistent, then the assessment is valid. If they are inconsistent, then the validity of the assessment results may be questionable.

4. **Reliability**

   This guideline encourages you to consider evidence regarding the reliability of assessment results. As we noted earlier the reliability of test scores sets an upper limit on the validity of their interpretation. Although reliability does not assure validity, you cannot have valid score interpretations if those scores are not reliable. As a result, efforts to increase the reliability of assessment results can enhance validity.

5. **Examination of test fairness**

   This guideline suggests that teachers should examine their classroom assessments to ensure that they are fair to all students. For example, tests and other assessments should be fair to students from diverse ethnic cultural backgrounds.

6. **Examination of practical features**

   This guideline encourages you to examine your classroom assessments to ensure that they are practical and efficient. This involves a consideration of the amount of time required to develop, administer and score the assessment. Although assessment plays an extremely important role in today’s schools, their development and use should not consume an inordinate amount of time.

   Moskal and Leydens (2000) explained that how the issues of validity and reliability may be addressed in the development of scoring rubrics. See article 8.5 ‘Validity in Scoring Rubrics’ in web material folder unit-8 on CD.
Self Assessment Questions

1. Explain the importance of validity in educational measurement. What are the factors that influence validity?
2. What types of validity should be taken into consideration in various types of measurements and assessments?
3. What type of evidence is required for different types of validity?
4. Describe three types of validity. For what kind of tests and situations each type of validity most appropriate?
5. Discuss in detail the unitary concept of validity? What is the difference between new and old nomenclature?
6. How can teachers improve the validity of their classroom assessment results?
7. Is a reliable test always valid? Why?
8. What is the difference between standard error of measurement and standard error of estimate? How are these two statistics related to the reliability and validity coefficients of a test?
9. What is the standard error made in estimating grade-point averages from scores on an aptitude test if the standard deviation of the criterion is .50 and correlation between test and criterion is .60? If a student’s predicted grade point average is 2.5, what is the probability that the student’s obtained grade-point average will fall between 2.1 and 2.9? Between 1.72 and 3.28?

Suggested Books


Web Based References

Unit–9

APITUDE TEST

131
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1</td>
<td>Introduction of the Unit</td>
<td>135</td>
</tr>
<tr>
<td>9.2</td>
<td>Objectives of the Unit</td>
<td>135</td>
</tr>
<tr>
<td>9.3</td>
<td>The Nature of Human Abilities</td>
<td>135</td>
</tr>
<tr>
<td>9.4</td>
<td>Intelligence</td>
<td>136</td>
</tr>
<tr>
<td>9.5</td>
<td>Aptitude Tests</td>
<td>137</td>
</tr>
<tr>
<td>9.6</td>
<td>Intelligence Tests</td>
<td>139</td>
</tr>
<tr>
<td>9.7</td>
<td>Diagnostic Tests</td>
<td>140</td>
</tr>
<tr>
<td>9.8</td>
<td>The Differential Ability Scales</td>
<td>141</td>
</tr>
<tr>
<td>9.9</td>
<td>Creativity as Special Ability</td>
<td>142</td>
</tr>
<tr>
<td>9.10</td>
<td>Self Assessment Questions</td>
<td>144</td>
</tr>
<tr>
<td>9.11</td>
<td>Suggested Readings</td>
<td>145</td>
</tr>
<tr>
<td>9.12</td>
<td>Web Based References</td>
<td>145</td>
</tr>
</tbody>
</table>
9.1 Introduction of the Unit

Aptitude reflects a student’s potential or capability for learning. Aptitude tests are like achievement tests, however, the difference is that an aptitude test measures the ability to learn while achievement test measures what a student has already learnt. Aptitude tests are not supposed to measure what has been taught in the schools like achievement tests. These tests rather measure level general ability of a student to perform a specific task. In a broader context we can say that aptitude tests measure a student’s overall performance across a broad range of mental capabilities. In addition to this aptitude tests often include items which measure more special abilities—like verbal and numerical skills—that can be used to predict scholastic performance in educational programs. Tests used for entry in a professional program include items that measure knowledge level for performing a specific task for as well as verbal and numerical ability. By doing this they try to ensure the successfulness of their program and institute.

For classroom teachers the application of aptitude tests may help them to design instructional activities and make groups of student for learning activities after matching the aptitude. Research suggests that students’ achievement is maximized when the method of instruction or learning activity matches the aptitude. In the light of importance of aptitude test this unit has been designed besides introducing you with the fundamental concepts about aptitudes their types and uses as diagnostic, differential, and special ability tests. The web based material and articles on human abilities, creativity in education, diagnostic tests and measuring aptitude have been included to give you a full understanding about the concept. You are expected to read these articles carefully.

9.2 Objectives of the Unit

After studying this unit and related web based material you will be able to:
1. Describe the nature of human abilities and intelligence.
2. Define and explain aptitude test.
3. Describe the major uses of aptitude and intelligence tests in schools.
4. Compare and contrasts the construct of achievement and aptitude.
5. Analyze the rationale for the analysis of aptitude-achievement discrepancies.
6. Explain the importance of aptitude tests in education.
7. Describe and evaluate major group aptitude/intelligence tests.
8. Evaluate diagnostic tests and their use in education.
9. Explain the differential ability scales.
10. Define and explain creativity as a special ability.
11. Develop a simple aptitude test for school subjects.

9.3 The Nature of Human Abilities

People differ in limitless way. Knowing something about a person’s interest, aptitudes, and past achievements can be of value both to the individual and organizations with which he or she comes in to contact such as colleges making admissions decisions or business making hiring decisions. Of particular interest to educators and psychologists who are concerned
with identifying individuals of all levels of ability and designing programmes to train, to treat, and educate them are children and adults who have very high and very low abilities. The differences among abilities within individuals may be as great as the differences between individuals. The term "ability," like most ordinary language expressions, has no single and preferred meaning. Generally it may stand for almost any kind of natural capacity or disposition, for example, someone's ability to express happiness, experience fear, or any other physical activity, etc. In this sense of the term, "ability" does not involve any sort of proficiency or skill. On the other side "ability" is synonymous with "skill". We might speak of a person's ability to write essay or poetry, to play cricket, or to make sketches. Abilities in this sense involve a mastery or conscious development of some set of more crucial attributes.

Lohman (1997) has discussed four ways in which the concept of ability has been defined in differential psychology:

1. Ability as Trait
2. Ability as Task Performance
3. Ability as Process
4. Ability as situated

See Lohman’s article 9.1 on ‘Human Abilities’ for detailed discussion in Web material folder Unit-9 on CD.

9.4 Intelligence

Intelligence is an umbrella term used to describe many related abilities, such as the capacities to reason, to plan, to solve problems, to think abstractly, to comprehend ideas, to use language, and to learn. There are several ways to define intelligence. Most people believe that intelligence is something that can be recognized when expressed in observable behavior. Despite the variety of concepts of intelligence, the most influential approach to understanding intelligence is based on psychometric testing. As a result of great efforts and researches the experts and non psychologists conceived intelligence as practical problem solving ability and social competence. The measurement of intelligence entails sampling an examinees performance on different types of tests and tasks as a function of developmental level. The scores obtained on an intelligence test is expressed as Intelligence Quotient (IQ). The ratio IQ is the ratio of the test taker’s mental age divided by his or her chronological age, multiplied by 100 to eliminate decimals. Such intelligence quotient (IQ) tests include the Stanford-Binet, Raven's Progressive Matrices, the Wechsler Adult Intelligence Scale and the Kaufman Assessment Battery for Children. All forms of IQ tests correlate highly with one another. The traditional view is that these tests measure g or "general intelligence factor". Intelligence, as measured by IQ and other aptitude tests, is widely used in educational, business, and military settings because it is an effective predictor of behavior. Intelligence is significantly correlated with successful training and performance outcomes. However, some alternative theories suggest that intelligence is the result of a number of independent abilities that uniquely contribute to human performance. For example, Louis Thurstone proposed a theory of multiple "primary abilities" in the early 20th Century. Howard Gardner's theory of multiple intelligences breaks intelligence down into at least eight different components: logical, linguistic, spatial, musical, kinesthetic, naturalist,
intrapersonal and interpersonal intelligences. He argues that psychometric tests address only few of these aspects of intelligence.

Robert Sternberg’s triarchic theory of intelligence proposes three fundamental aspects of intelligence: analytic, creative, and practical—of which only the first is measured to any significant extent by mainstream tests. His investigations suggest the need for a balance between analytic intelligence, on the one hand, and creative and especially practical intelligence on the other.

9.5 Aptitude Tests

As has been explained in previous units that achievement tests assess student’s knowledge or skills in content domain in which they have received instruction. Whereas aptitude tests are designed to measure the cognitive skills, abilities and knowledge that individuals have accumulated as the result of their overall life experiences. Then we can say that an aptitude test is a psychometric device, more broader in scope than achievement test predict performance in a special skill. Though Aptitude tests are similar in many respects to achievement tests, but the difference is that in aptitude tests emphasis is always on general intelligence. Aptitude tests measure a student’s overall performance across a broad range of mental capabilities.

An aptitude test measures the ability to learn while an achievement test measures what a student has already learnt. Compared to achievement tests, aptitude tests cover a broader area and look at a wider range of experiences. Achievement tests, on the other hand are limited to recent learning and particular school subjects.

Aptitude tests tell us what a student can do apart from the specific curriculum that the student has already experienced. Sometimes, aptitude and achievement tests look similar. At the higher levels of education, the content of aptitude tests more resembles achievement tests, because at advanced levels the knowledge that a student has already accumulated is a good predictor of success.

(i) The value of aptitude tests

According to Macklem (1990) individually administered aptitude tests have the following qualities, for being used in educational situations:

- They are predictor of future scholastic achievement of students.
- They are an excellent way of comparing students’ performance.
- They provide a profile of strengths and weaknesses of students.
- Aptitude tests provide assessment of individual differences.
- Aptitude tests uncover hidden talents in some children, thus improving their educational opportunities.
- They are effective tools for working with gifted children.
(ii) Uses of aptitude test results
Macklem (1990) identified three major uses of aptitude tests in educational settings:

i) Instructional
Teachers can use aptitude test results to adjust their curricula according to the level of their students, or to design assignments for students keeping in mind their individual differences.

ii) Administrative
Aptitude tests can help in making administrative decisions. For example, aptitude test scores can be used in admitting candidates for various types of professional education.

iii) 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.

See article 9.2 by Macklem (1990) ‘Measuring Aptitude’ based Material folder Unit-9 on CD.

9.5.1 Scholastic aptitude test (SAT)
The SAT was introduced as an objective exam in 1926. Until 1995, the SAT was a three-hour test divided into two parts: verbal and Mathematics. The verbal part consisted of sections that included Analogies, Reading Comprehension, Antonyms, and Sentence Completion. Test items for the SAT are constructed by experts in the field. Many group intelligence tests have been designed specifically to measure aptitude for scholastic work and are referred as academic ability tests. Some group intelligence tests have a broader focus than this, but they are still similar in content to academic ability measures. Now a days they are heavily loaded with verbal, numerical, and school type items. A number of different tests have been used over the years for college and university purposes. In short we can say that scholastic aptitude test measures performance, based on the learning abilities. It does not measure the capacity or learning potential but the same is inferred from its results.

Many colleges, institutions, and universities require scores on standardized tests such as SAT or the Graduate Record Examination (GRE) as part of the undergraduate or graduate admission process. Raw Scores on SAT and GRE are converted to standard score such that the resulting distribution has a mean of 500 and a standard deviation of 100. Scholastic aptitude tests are further divided into the following kinds:

(a) Group tests
The majority of tests administered in schools are group tests. Some of these tests yield a single score while others produce more than one scores based on the test items. Single score tests are designed to measure general mental ability of students. Items of different kinds are mixed together and then arranged in the order of difficulty. These tests help in predicting future success in school work. Separate score tests yield both verbal and non-verbal scores. Verbal scores are best indicator of future success in school programs.
(b) **Individual tests**

Individual tests are administered to one person at a time. These tests are generally called intelligence tests, but as the term ‘intelligence’ is losing ground in favor of new terms like general ability, cognitive ability, etc. they are now included in scholastic aptitude tests. These tests also measure the learned ability which is helpful in predicting the school performance in the future years.

<table>
<thead>
<tr>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locate an aptitude test developed by some researcher in Pakistan. Select a few items and apply to a class/group of students and check its validity.</td>
</tr>
</tbody>
</table>

### 9.6 Intelligence Tests

Despite different explanation of intelligence in literature, it is generally agreed that intelligence tests assess general mental ability. The assumption, underlying these tests, is that some individuals, having higher level of general mental ability, can learn better from experience. Many different intelligence tests exist. Some intelligence were constructed on the basis of a theory. An intelligence test may be developed on the basis of one theory, but then be reconceptualized in terms of another theory. For example much has been written about a theory of intelligence that contains features of the Cattell model and the Carroll three stratum model, and Thurstone conceived a model of Primary Mental Ability. Two major tests are discussed below that have been developed according to a strong theoretical framework.

1. **Stanford-Binet Intelligence Scale**

Alfred Binet developed this test in 1905 in France. Binet and Theodore Simon revised this test in 1908 and again in 1911. They created some tasks and arranged them in order of least difficult to most difficult and grouped them into age level tasks. The mental age of the test taker was calculated on the basis of number of tasks completed. The intelligence quotient (IQ) was calculated by dividing mental age (MA) with chronological age (CA) and multiplying by 100.

\[
IQ = \frac{MA}{CA} \times 100
\]

In 1916, Lewis Terman at Stanford University revised the 1911 version, and this version was called Stanford-Binet Intelligence Scale. The latest editions of this test have been published in 1986. This test consists of series of tests arranged by age levels. The lowest age level considered for the test is two years.

2. **Wechsler Intelligence Scale**

The first Wechsler scale was published in 1939. The Wechsler intelligence for Children (WISC) was devised in 1949 and Wechsler Adult Intelligence Scale (WAIS) in 1955. The first Wechsler scale was meant for the age group ranging from 10 years to 60 years; WISC for children between 5 to 15 years of age; and WAIS for age 16 and upward. These tests contain both verbal and performance items. The score on the scale is
arrived at by converting raw scores into standard scores (with mean 100 and standard deviation 15).

For detailed study Kline (2005), pp. 318-322

3. **Culture Fair Testing**

Tests are not true measures of a person’s general ability. They are usually culture biased. A test developed for white majority of Americans cannot be a good measure of intelligence of blacks living in slum areas. Similarly, the tests developed in America or UK cannot be applied in Pakistani situations. Therefore, the need for culture fair tests was badly felt in educational circles. The first bias is that of the language, and then there are issues of motivation for the test, competitiveness, speed, practice, etc. Culture free testing is an effort to develop a measure of a person’s potential that is free of all types of differences. The following steps are taken for this purpose:

1. All material is non-verbal and consists of pictures and diagrams familiar to the group of people for whom the test is developed.
2. Material is interesting for students.
3. There is no time limit.
4. The test procedures are very simple.
5. Test skills are common to the specific culture group.

Cattel’s culture fair intelligence tests are the best example of this kind of testing.

9.7 **Diagnostic Tests**

Diagnostic tests have the diagnostic function of identifying specific difficulties in learning a subject. Psychologists defined diagnostic tests as the tests systematically designed to provide information about skills that students have or have not mastered. According to them these tests provide profile of strengths and weaknesses for specific students on a limited number of clearly defined instructional objectives. In educational context the term ‘diagnostic’ is typically applied to tests or test data that are used to pinpoint a student’s difficulty for the purpose of remediating that difficulty. The narrow focus and curricular relevance of these tests make them a useful tool for instructional planning.

At the start of instructions, teachers may assess the prerequisite skills and pinpoint individual and class needs so that they can target the instruction at appropriate level. During instructions teachers may use diagnostic tests to make grouping decisions and identify students who need extra help or more challenging assignments. At the end of particular instructional period, teacher may use diagnostic tests to judge the effectiveness of their instructional materials and methods. A diagnostic reading test, for example, may contain a number of subtests, each designed to analyze a specific knowledge or skill required to read and to bring into full relief the specific problems, if any that need to be addressed. On the basis of child’s performance on a diagnostic reading test, a teacher or administrator might make a class placement decision. In general, diagnostic tests are administered to students who have already demonstrated their problem with a particular subject area through their poor performance either in the classroom or on some achievement test. It is therefore
understandable that diagnostic tests tend to contain simpler items than do achievement tests designed for use with members of the same grade.

(I) Diagnostic Tests Development Process
Herman and Winters (1985) described following five steps in the process of diagnostic test development:

1. **Specifying Skill Map:** The skill map should be specified in the light of instructional objectives. It must assess skills related to the mastery of particular instructional objectives.

2. **Creating Test Description:** This step requires hard thought about the nature of skills, item content and item format that will appear on the test. Test descriptions provide detailed guideline for item content, level of difficulty, answer choices and how the test will be scored. In short, the test descriptions provide instructions about how to write the test.

3. **Writing Test Items:** After completing test descriptions, test items should be written according to the given descriptions. Once the items are written they should be reviewed against the rules given in test descriptions for consistency.

4. **Reviewing Test Items:** The formal item review should be done by a person other than the test writer. The purpose of the item review is to screen out poorly written questions and the ones that don’t assess the skills as intended.

5. **Tryout Items:** Tryout items, also called pilot test, is the final step in the process. Items are tried out with a small number of students who are representative of the students with whom the test will be used. Item tryout provides information about item difficulty and homogeneity, and other qualities of the test.

*For detail see article 9.3 Guidelines for developing ‘Diagnostic Tests’ in web material folder Folder Unit-9 on CD.*

9.8 The Differential Ability Scales

The Differential Ability Scales is an individually administered test battery that measures different levels of interest and intelligence of an individual in different fields. These scales provide a profile of individual’s strengths and weaknesses, thus, can be used for classification and diagnostic purposes. The theory behind differential ability scale is that different individuals have varying levels of interest and intelligence in different fields. Some may be good at math but bad in verbal reasoning. Some may write language excellently but may be very bad in calculations. Total score of general aptitude tests can’t make a true calculation of different kinds of aptitudes; however, differential ability scales measure all kinds of aptitudes separately.

A brief description of different kinds of differential ability scales is given below:

1. **Verbal Differential Ability Scale**
This scale measures the ability to find relations amongst words. How do you view a concept? How do you manipulate abstract ideas? etc., are the topics on which various verbal analogies are framed.
2. **Numerical Differential Ability Scale**
   This scale measures one’s capability to interpret numerical relationships between different figures and also tests the mathematical reasoning.

3. **Abstract Reasoning Differential Ability Scale**
   This differential scale measures reasoning skills while solving problems in terms of size, shape, position, quantity or other geometric figures and shapes. Logical thinking is involved.

4. **Mechanical Reasoning Differential Ability Scale**
   This scale measures the ability to understand mechanical principles of machines, motion and devices.

5. **Space Relations Differential Ability Scale**
   This scale measures the capability to analyze three dimensional figures. How do you see 2D or 3D drawings?

6. **Spelling Differential Ability Scale**
   This scale measures the capability to recognize correct or incorrect spelled common words.

7. **Language Differential Ability Scale**
   This scale is used to measure the ability to detect grammatical and other language errors (using present/past tense, direct/indirect speech, capitalization etc).

8. **Speed and Accuracy Differential Ability Scales**
   This scale measures the quickness and accuracy in performing different tasks.

<table>
<thead>
<tr>
<th><strong>Activity</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Develop a specific ability test for Urdu/English language learning, try it out, revise it and see whether it is valid. Submit to your tutor for evaluation.</td>
</tr>
</tbody>
</table>

9.9 **Creativity as Special Ability**

   Psychologists believe that inner drive is the main cause of creativity in most of the creative peoples. “Other affective and cognitive traits said to be the characteristics of creative people are ideational fluency, flexibility, unconventionalism, social sensitivity, nondefensiveness, a greater willingness to concede faults, and close ties with parents” (Aiken, 2003). From this definition we can say that creativity is a mental process involving the generation of new ideas, or new associations between existing ideas or concepts. From a scientific point of view, the products of creative thought (sometimes referred to as divergent thought) are usually considered to have both originality and appropriateness. A more everyday conception of creativity is that it is simply the act of making something new. Unlike many phenomena in science, there is no single,
authoritative perspective or definition of creativity. Unlike many phenomena in psychology, there is no standardized measurement technique.

There has been debate in the psychological literature about whether intelligence and creativity are part of the same process or represent distinct mental processes. Evidence from attempts to look at correlations between intelligence and creativity from the 1950s onwards, by authors such as Barron, Guilford or Wallach and Kogan, regularly suggested that correlations between these concepts were low enough to justify treating them as distinct concepts. Some researchers believe that creativity is the outcome of the same cognitive processes as intelligence, and is only judged as creativity in terms of its consequences, i.e. when the outcome of cognitive processes happens to produce something novel.

Measuring Creativity

It is sometime maintained that above-average intelligence is necessary for creative productivity. Efforts have been made to develop measures of divergent thinking with open ended problems. Though, it creates difficulty in scoring and determining the reliability and validity of these tests. Several attempts have been made to develop a creativity quotient of an individual similar to the Intelligence Quotient (IQ) however these have been unsuccessful. Most measures of creativity are dependent on the personal judgment of the tester, so a standardized measure is difficult to develop.

i) Psychometric approach

J. P. Guilford’s group, which pioneered the modern psychometric study of creativity, constructed several tests to measure creativity in 1967:

- Plot Titles, where participants are given the plot of a story and asked to write original titles.
- Quick Responses is a word-association test scored for uncommonness.
- Figure Concepts, where participants were given simple drawings of objects and individuals were asked to find qualities or features that are common by two or more drawings; these were scored for uncommonness.
- Unusual uses means finding unusual uses for common everyday objects such as bricks.
- Remote Associations, where participants are asked to find a word between two given words (e.g. Hand Call)
- Remote Consequences, where participants are asked to generate a list of consequences of unexpected events (e.g. loss of gravity)

Building on Guilford’s work, Torrance developed the Torrance Tests of Creative Thinking in 1974. They involved simple tests of divergent thinking and other problem-solving skills, which were scored on:

- Fluency. The total number of interpretable, meaningful, and relevant ideas generated in response to the stimulus.
- Flexibility. The number of different categories of relevant responses.
- Originality. The statistical originality of the responses among the test subjects.
- Elaboration. The amount of detail in the responses.
The Creativity Achievement Questionnaire, a self-report test that measures creative achievement across 10 domains, was described in 2005 and shown to be reliable and valid when compared to other measures of creativity and to independent evaluation of creative output.

ii) Social-personality approach

Some researchers have taken a social-personality approach to the measurement of creativity. In these studies, personality traits such as independence of judgment, self-confidence, attraction to complexity, aesthetic orientation and risk-taking are used as measures of the creativity of individuals. Other researchers have related creativity to the trait, openness to experience.

iii) Other approaches to measurement

Genrich Altshuller in the 1950s introduced approaching creativity as an exact science with TRIZ (Theory of inventive problem solving) and a Level-of-Invention measure. You will read a detailed discussion about creativity in relation with education in Craft (2001) report.

Craft (2001) discussed the topic of creativity in relation with education.
See article 9.4 ‘Creativity in Education’ in Folder 9 on CD.

9.10 Self Assessment Questions

1. What is the difference between ability and intelligence?
2. Are intelligence and creativity two different concepts? How?
3. Explain how achievement and aptitude can be conceptualized as different aspects of a continuum. Provide examples to illustrate this continuum.
4. How can we use different aptitude tests in various educational settings?
5. What is the importance of diagnostic tests in education?
6. Explain the use of differential ability scales in education.
7. What is the IQ ratio of a child who is 10 years, 10 months old if his or her score on the Stanford-Binet Intelligence Scale is equal to a mental age of 8 years, 6 months?
8. Describe the major uses of aptitude and intelligence tests in schools.
9. To test the creative ability of 10th class students, try the following exercises:
   (a) Ask the students that how many uses can they think for a paper clip, rubber ball, brick, wire clothes hanger, foot ruler, or tooth pick?
   (b) Try to imagine how things would changed if:
       Every one had three arms.
       Every one had six fingers and no thumb on each hand.
       It did not rain at all for six months.
       Compare their responses.
9.11 Suggested Readings


9.12 Web Based References