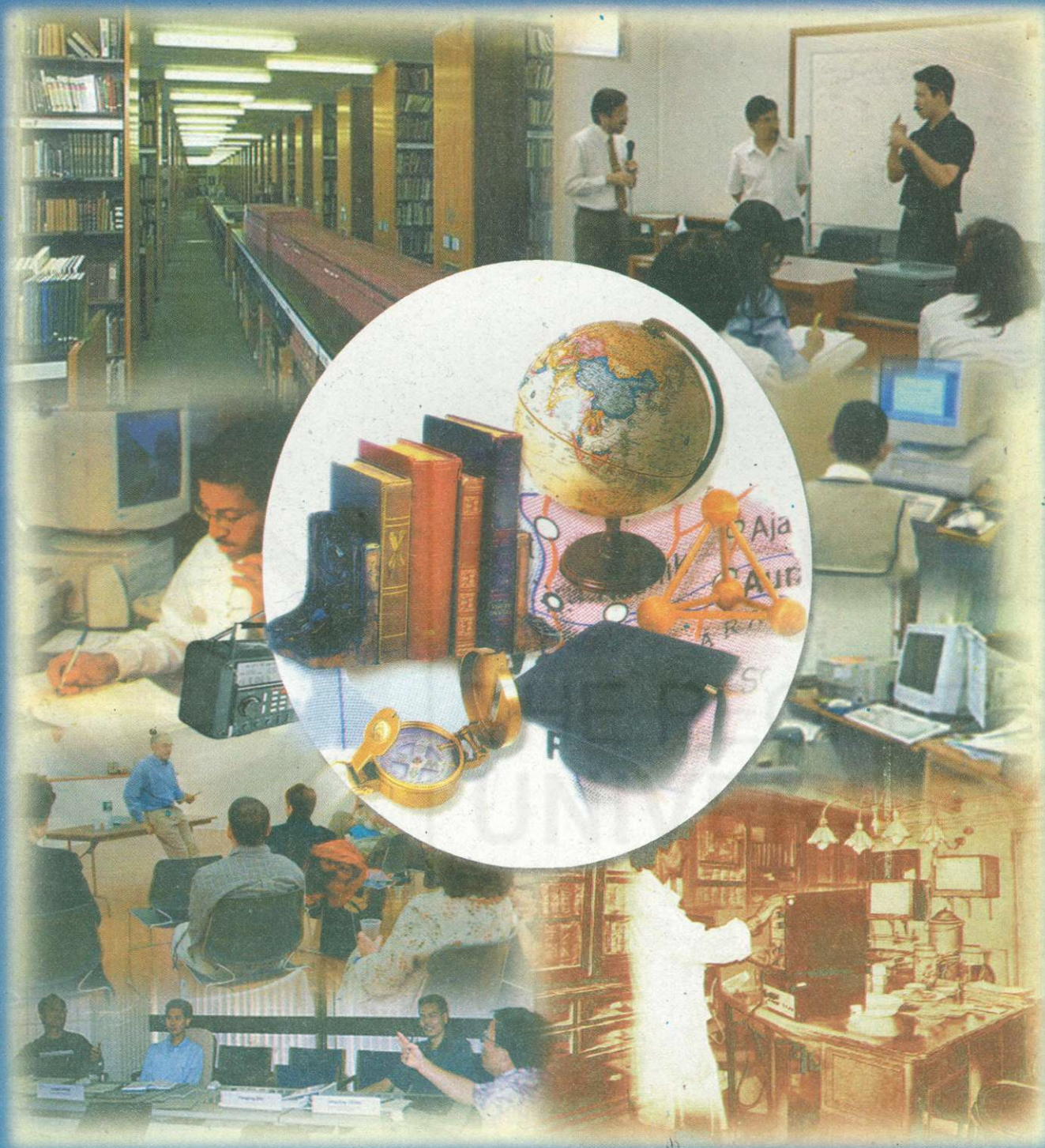




"Education is a liberating force, and in our age it is also a democratising force, cutting across the barriers of caste and class, smoothing out inequalities imposed by birth and other circumstances."

– Indira Gandhi

ignou
THE PEOPLE'S
UNIVERSITY

“शिक्षा मानव को बंधनों से मुक्त करती है और आज के युग में तो यह लोकतंत्र की भावना का आधार भी है। जन्म तथा अन्य कारणों से उत्पन्न जाति एवं वर्गगत विषमताओं को दूर करते हुए मनुष्य को इन सबसे ऊपर उठाती है।”

– इन्दिरा गांधी



Block

4

EVALUATION IN HIGHER EDUCATION: MECHANICS AND PROCESSING

Unit 15

Mechanics of Evaluation-I

5

Unit 16

Mechanics of Evaluation-II

24

Unit 17

Processing Evaluation Data

44

Unit 18

Alternative Evaluation Procedures

70

Unit 19

Online/Web-based Student Assessment

90

COURSE TEAM

Course Contribution

Prof. G.R. Sudame
Formerly with Centre of
Advanced Study in Education
M.S. University
Boroda

Coordination

Dr. N.K. Dash
Course Coordinator
SOE, IGNOU

Prof. S.B. Menon
(Programme Coordinator)
Former Director, SOE, IGNOU
and Dean, CIE
University of Delhi

Prof. B.N. Koul
(Programme Coordinator)
Former Director, SOE, IGNOU

Prof. M.B. Menon
Former Director, SOE, IGNOU
Currently Teacher Education
Specialist, COL, Vancouver

Editing

Prof. S.S. Kulkarni
(Course Editor)
Narayan Ashram
Koloshi, Sindhudurg
Maharashtra

Prof. M.B. Menon
(Associate Editor)

Dr. N.K. Dash
(Content Transformation
and Format Editing)

Ms. Jayshree Raveendran
(Language Editing)
Formerly with IGNOU

COURSE TEAM (REVISED EDITION)

Course Contribution

Dr. S. Kumar
Reader
Centre of Advanced
Study in Education
Faculty of Education
and Psychology
M.S. University, Baroda

Dr. Santosh Sharma
Reader
Department of Teacher
Education, NCERT
New Delhi

Dr. Anurag Saxena
Reader
School of Management Studies
IGNOU

Dr. N.K. Dash
Reader
School of Education
IGNOU

Editing

Prof. Pritam Singh
Former Head, Department
of Measurement and Evaluation
NCERT, New Delhi

Coordination

Dr. N.K. Dash
(Course Coordinator)
SOE, IGNOU

Ms. Poonam Bhushan
(Programme Coordinator)
SOE, IGNOU

Format Concept and Design

Ms. Poonam Bhushan
(Programme Coordinator)
SOE, IGNOU

Prof. S.V.S. Chaudhary
Director, School of Education
IGNOU

Dr. N.K. Dash
SOE, IGNOU

Production

Prof. M.L. Kaul
Director
School of Education
IGNOU, New Delhi

S.S. Venkatechalam
A.R. (P)
School of Education
IGNOU, New Delhi

M. Sumathy Nair
S.O. (P)
School of Education
IGNOU, New Delhi

December, 2008 (Reprint)

© Indira Gandhi National Open University, 2003 (Revised Edition)

ISBN-81-266-0955-9

All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Indira Gandhi National Open University.

Further information on the Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi, New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi, by the Director, School of Education.

Laser typeset by Nath Graphics, 1/21, Sarvapriya Vihar, New Delhi-110 016.

Printed at M/s. Apex Printer, WA-117A, Shakarpur, Delhi-110092

MES-102 *Instruction in Higher Education*

Block 1 **Instruction in a Systemic Perspective**

- Unit 1 Instructional System
 - Unit 2 Input Alternatives–Teacher Controlled
 - Unit 3 Input Alternatives–Learner Controlled
 - Unit 4 Evolving Instructional Strategies
 - Unit 5 Unit and Topic Planning
-

Block 2 **Communication Skills**

- Unit 6 Teacher Competence in Higher Education
 - Unit 7 Skills Associated with a Good Lecture
 - Unit 8 Skills Associated with the Conduct of Interaction Sessions
 - Unit 9 Skills of Using Communication Aids
 - Unit 10 Emerging Communication and Information Technologies
-

Block 3 **Evaluation Perspectives in Higher Education**

- Unit 11 Status of Evaluation in Higher Education-I
 - Unit 12 Status of Evaluation in Higher Education-II
 - Unit 13 Evaluation Situations in Higher Education-I
 - Unit 14 Evaluation Situations in Higher Education-II
-

Block 4 ***Evaluation in Higher Education: Mechanics and Processing***

- Unit 15 *Mechanics of Evaluation-I*
 - Unit 16 *Mechanics of Evaluation-II*
 - Unit 17 *Processing Evaluation Data*
 - Unit 18 *Alternative Evaluation Procedures*
 - Unit 19 *Online/Web-based Student Assessment*
-

Block 4 Evaluation in Higher Education: Mechanics and Processing

Introduction to the Block

We have already introduced you to the meaning, varieties and uses of evaluation in Block-3. In this Block, we shall deal with the various specific issues of evaluation, which every teacher is expected to know. We have here tried to acquaint you on how to prepare a question paper, validate it and develop a question bank which contains a large number of questions for future use. But evaluation is not complete with preparing a question paper and obtaining evaluation data alone. Equally important, is the interpretation of data and their uses. Hence, we have also discussed how to process and interpret both quantitative and qualitative data. We have given alternate evaluation procedures.

In Unit 15, we talk about three basic elements which are combined to prepare a question paper. These are: outlining the subject matter, identifying the desired learning outcomes in specific terms and selecting suitable types of test items. We also discuss preparation of blueprint.

In Unit 16, we identify the features of a test which make the test a valid one. They are validity, reliability, objectivity and usability. We also discuss the basic parameters—item difficulty and item discrimination—which determine the worth of a test. Lastly, We highlight on the importance of question banks in the field of testing and evaluation.

In Unit 17, we point out the merits and demerits of marking and grading, and emphasize the viewpoint that grading is preferably to making. We also provide some basic statistics which would help you in the processing and interpreting of norm-referenced and criterion-referenced evaluation data.

In Unit 18, we provide you with some facts on alternate evaluation procedures—observation, self-reporting modes, peer appraisal—which are used to collect qualitative data. We have also touched upon how to report the results of evaluation to different beneficiaries.

In Unit 19, we have presented how students' performance can be assessed online. In this context, we have given some useful websites which can be accessed to know details about online student assessment.

Mechanics of Evaluation-I

15

Introduction

You have been made aware of the status of evaluation and evaluation situations in higher education in the country, while studying through Block-3. In this Unit, you will learn about the mechanics of evaluation. Framing a question paper systematically involves mainly four steps. These are: (i) outlining the instructional content, (ii) identifying the desired learning outcomes in specific terms, (iii) development and selection of evaluation items and (iv) preparation of a blueprint. In this Unit, you will learn about different types of test items and their merits and demerits. Towards the end of the Unit, you will be acquainted with the preparation of a good question paper. In this context, you will learn how to develop a table of specification (blue-print) for preparing a question paper.

Learning outcomes

By the end of this Unit, you will be able to:

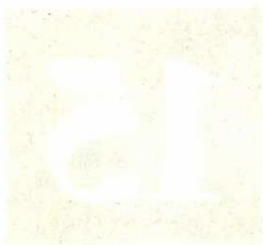
- describe the different steps in framing a question paper;
- distinguish among essay type, short-answer type and objective type items;
- frame different types of test items, keeping in view their characteristics; and
- design and develop a table of specifications (blueprint) for setting a question paper.

Framing test items and question papers

Educational evaluation is an integral part of the activities that are meant to provide education to students. Various activities like classroom lectures, laboratory work, seminars, conducting research activities, practical experiences in the field, internship programs in colleges and universities, games, sports, etc. are organised in educational institutions. These activities are organised so that specific objectives can be achieved. Taken together, these specific objectives reflect the purpose of education for the students. We have called these specific objectives of various activities organised in colleges as "Instructional Objectives". The extent to which such instructional objectives have been achieved by different students is, indeed, the focus of organising different activities in the colleges.

Outlining the subject matter content

The first step in preparing a good question paper is to make an outline of the subject



matter content. Outlining the content means listing the specific points of the instructional content. In this activity you should ensure, that, content details must be adequate enough to represent the entire course. The outline of an 'Educational Evaluation' Unit is illustrated for test preparation.

Outline of the content of educational evaluation unit

- (a) Meaning of evaluation
 - i) Meaning of educational objectives
 - ii) Meaning of Educational evaluation
 - iii) Steps in the process of evaluation
 - iv) Characteristics of evaluation
- (b) Relationship of educational objectives, learning experiences and evaluation
- (c) Types of evaluation
 - i) Formative evaluation
 - ii) Summative evaluation
 - iii) Diagnostic evaluation
- (d) Uses of evaluation
 - i) Uses of evaluation for teachers
 - ii) Uses of evaluation for administrators
 - iii) Uses of evaluation for students

Identifying and stating the desired learning outcomes in specific terms

The second step in preparing a good question paper is identifying and stating the desired learning outcomes in specific terms. You may have noticed that physical training and sports are organised in many colleges. Which specific objectives does a student achieve through these activities? The student's physical condition is expected to be improved through such physical activities. Other activities such as classroom lectures, laboratory work, etc., are organised so that desirable changes in students' knowledge, understanding, skills, interests and attitudes are brought about. Such desirable changes in the students' behaviour-physical, mental, social and emotional- essentially represent the learning that has taken place and therefore, may be called as 'learning outcomes'. We may represent the relationship among the intended instructional objectives, learning experiences (various activities organised) and the actual learning outcomes or objectives achieved as follows:

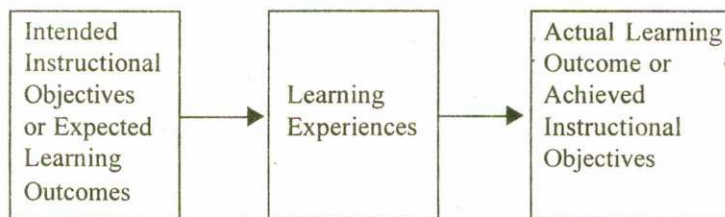


Figure 1: Relationship among intended instructional objectives, learning experiences and actual learning outcomes.

It may be mentioned here, that, the achievement tests, for which we prepare the question paper, measure students' knowledge, understanding, ability to apply, analyse, etc. and also skills related to drawing figures, charts, etc. in different subjects. For evaluating these, a teacher has to identify and state the learning outcomes in students' behaviour. This means, that, statement of objectives will indicate the specific competencies which students will be able to develop after the instruction. For example, the learning outcomes for a unit on 'Educational Evaluation' may be stated as follows:

- (a) The student knows the basic concepts, which means he/she:
- defines different terms like assessment, examination, evaluation, etc.
 - writes the steps involved in the process of evaluation
 - states the characteristics of evaluation process
 - describes different types of evaluation
 - lists the purposes of evaluation
- (b) The student understands the concept and process of evaluation, which means he/she:
- differentiates between evaluation and measurement, assessment and examination,
 - identifies the relationship amongst educational objectives, learning experiences and learning outcomes,
 - relates instructional objectives to learning experiences and evaluation.
 - illustrates how results can be used for improving instruction
- (c) The student applies concepts and principles to new situations, which means he/she:
- states the instructional objectives in behavioural terms or intended competencies.
 - selects the learning experiences suitable for attaining the given instructional objectives.
 - gives reasons why summative evaluation is not suitable for improving instruction.
- (d) The student develops drawing and practical skills, which means he/she:
- draws a figure showing relationship of objectives, learning experiences and learning outcomes.
 - defines practical skills in terms of process and product skills.

You have seen that for framing test items and question paper, the first step is to outline the subject matter content into different units of teaching. The second step is to identify desired learning outcomes and then state them in terms of instructional objectives. The third step is to decide the type of questions to be included and develop the questions. In this context, let us discuss the different types of questions or test items in the coming section.

Self-assessment

1. Why do you think it is necessary to specify the expected learning outcomes while framing a question paper?
2. State two instructional objectives specifying expected learning outcome.
3. Why is the subject matter content to be analysed while framing a question paper?

Different forms of test items or questions

The most widely used tool of evaluation in colleges at present, is the achievement test in the written format. Such a test consists of many questions or items of evaluation.

Questions or test items which are used in achievement testing are generally divided into two broad major categories.

- a) Supply type/Free response type
- b) Selection type/Objective response type

Supply Answer Type Questions

Supply answer type questions can be broadly categorized into three major types.

- i) Long answer type questions
- ii) Short answer type questions
- iii) Very short answer type questions

In this section, we shall discuss essay type questions.

Essay type items / questions

You are familiar with essay questions which require long answers from students giving them considerable freedom of response. Students answering such questions are free to select, organise and present the ideas in their own words. The learning outcomes that may be expected of students through such a question, are related to a complex set of behaviours, which involve selection and organisation of relevant ideas and the ability to express them in written form in their own words. An essay type test, by its very nature, can include only a few questions, say five or six, to be answered in 2 or 3 hours. Generally, the preparation of an essay type question paper is considered to be easy. On the contrary, due to lack of adequate content coverage, the **content validity of the essay test is low**. This is a major limitation of the essay type question.

Essay type questions usually do not clearly specify the exact answers required and the students are free to select, organise and express their ideas. Sometimes, when a student does not know the actual answer, he may still write some matter, which seems to be relevant to the question, but does not represent the answer that is actually required. This is generally called '**bluffing**'. Again, because of the lack of specificity in the answers required, and the variety of expression in the written answers given by the students, scoring gets influenced to a great extent by the opinion of the scorers. This reduces the objectivity in terms of consistency in scoring of the essay type test. This is the second limitation of the essay type item.

In addition to the possibility of bluffing, the writing ability and language proficiency of the student, may interfere with the giving of a precise answer. because of these errors of measurement, the total score of a student on an essay test may not represent the actual amount of learning by him. This makes the **essay type item or question less reliable**, which is its third limitation.

Thus, generally, essay type tests are low in the three essential characteristics of an effective tool of evaluation, namely **validity, reliability and objectivity**. We will discuss in detail about these terms in the next Unit. But, despite these limitations, essay type tests are useful in evaluating certain higher level objectives such as organisation, interpretation, application, etc. if they are properly framed. Can these tests be improved? Is it possible to reduce some of their limitations?

How to improve essay type tests

Consider the following questions:

Describe the four essential characteristics of a State and explain why sovereignty is so important.

Marks: 10

It may be seen that the student is expected to answer two different questions, which are related to two different objectives. The student may find it difficult to decide which

part of the question should be given more importance. To avoid this situation, each essay type question should reflect differential marks allocation. For example, instead of 10 marks, we can write 4+6 which delimits the importance of each part and facilitate scoring of scripts.

Sometimes, essay type questions are used to measure the extent to which students are able to recall certain names, dates, specific tools, etc. For example, consider the following question.

What are the three laws of Newton?

In fact, this is not an essay type but short answer question where answer is expected in three lines or so covering three specific points. Here, the student has to remember just the names of the three types of Newtonian laws. This question is intended to measure the student's learning outcome related to his knowledge of the three types of Newtonian laws. But the question does not specifically indicate whether only the names of the three Newtonian laws are to be stated or some kind of explanation also to be given. If only the memorisation of the names is the expected learning outcome, it can be better tested by a fill-in-the blank type of question as given hereunder.

Three laws of Newton are:

- a) Law of _____
- b) Law of _____
- c) Law of _____

Therefore, the vague language of the question creates difficulty for students to know the expected answer. Essay type questions, often, do not indicate to the student in clear and specific terms what he is expected to answer.

Very often, when essay type questions are scored by teachers, they may have a personal bias toward a particular style of writing, certain content of the answer, or certain other factors such as legibility, spelling, grammatical usage, etc. This (difference in scores) is, in fact, one of the major limitations of essay type questions. Scoring of answers in these tests cannot be objective. On the other hand, stopping of essay question for this reason is neither possible, nor desirable, because, as we have stated earlier, it is through this form that we can test the student's achievement of certain higher level objectives, such as application, analysis, interpretation, logical organisation, etc. Hence, if a standard scoring procedure is evolved, which can be followed by all, the objectivity in scoring essay questions could be increased to a considerable extent.

While scoring essay questions, inconsistency is introduced at the very first stage, because the content which constitutes the correct answer has not been decided in specific terms. Therefore, the first step in formulating a standard procedure is to formulate a model answer to each question. This model answer should guide assessing the answer for its content. External factors like the style of writing, legibility, etc. should not be given undue weightage unless specified explicitly. If necessary, these external factors may be assigned separate marks, which represent the learning outcomes related to the content sampled.

Sometimes, you may come across questions such as:

Q. Describe any two advantages of a presidential form of Government.

Or

Q. 'Describe any two uses of the essay type test'.

It can be seen that several choices are there and each of them will be acceptable as the answer. In a situation like this, all the acceptable alternative choices should be included in the model answer or scheme of marking and considered as correct while marking.

Another inconsistency in scoring arises in awarding marks to different parts or aspects of an answer. For example, consider the students who have been asked to discuss a phenomenon with four illustrations. Students may give a variety of answers to such a question. Some may give only one illustration, others may give two, three or four. The illustrations given may also differ in their degree of relevance to the phenomenon. This suggests that there is a need to formulate a standard scoring key, which divides the total marks for the question into partial credits, thus clearly specifying the marks that can be allotted to different parts or aspects of the answer.

It may be concluded that in order to bring objectivity in scoring answers to essay type questions, a model answer should be formulated for each question, which should reflect detailed marking scheme clearly specifying the credit points that can be assessed as different aspects of the answer.

In addition to providing model answers and marking scheme, examiners can be advised to assess answer sheets question wise. When assessment work is to be done in groups, then, one examiner may be assigned assessment of single question. Limiting the content and repeated evaluation of the same content will bring more objectivity in marking essay type questions.

We have discussed in this section about certain steps which may be taken to improve the essay type tests. Even after these improvements, the essay type tests cannot be made completely objective. Still, it is desirable to use the essay type tests for measuring the learning outcomes related to certain educational objectives, such as selection, organization summarization, integration and expression of ideas. A number of other learning outcomes can be more effectively measured through objective type test. We discuss the objective type tests in the section that follows.

Self-assessment

4. *What are the limitations of essay type questions? How can construction of such type of questions be improved?*

Short-answer type questions

Short - answer questions are those which require answers to be supplied by the students in terms of sentences related to a single or more than one concept. These questions do not require selection and organization of the content and then presenting it in the form of an essay. In these questions both freedom of response and length of the answer are restricted. It may also be noted that these questions do not require answers in one or two words like objective questions.

As stated earlier, these questions have restricted length. It could be a single sentence answer, definitions or descriptions of terms/concepts in 1–4 lines, or sometimes these items require answers to be written just in terms of a few points.

A few examples of short – answer items are given below:

1. Make sentence using the word 'arduous'.
2. Define the following terms in 2-3 lines each.
 - i) Educational evaluation.
 - ii) Validity
 - iii) Objectivity
 - iv) Reliability.

i. List four uses of evaluation for teachers:

- i) _____.
- ii) _____.
- iii) _____.
- iv) _____.

Short-answer questions require definite and restricted answers. With the help of these questions coverage of large content becomes possible. As items require definite and restricted content it increases objectivity of the test. The large coverage of content increases validity of the test.

In short – answer questions students supply the answers, therefore, guessing is completely eliminated. Items also deal with specific concepts, ideas, terms having very – very definite and restricted answers which avoids bluffing on the part of students. Thus we can say that short - answer questions neither admit bluffing nor encourage guessing.

With respect to one of the essential characteristics of the tools, i.e. objectivity; one finds that short answer items are less objective when compared to objective type items. But these items comparatively ensure objectivity in scoring when compared to essay type questions.

Considering the nature of short answer questions, their use along-with objective type items may be desirable in higher education.

Self-assessment

5. Define short – answer question in 2-3 lines.

Very short answer type questions

Very short answer type questions expect students to supply the answers to the question. This type of questions are mainly, in the form of direct question and the answers may be writing a definition, a phrase, a principle, or a fact. The time given to answer such questions varies from two to three minutes.

Very short type items require the student to supply rather than select the answer. Consider the following question.

- i) “What is the chemical formula for sulphuric acid?”

The same question can be asked in another format as given below:

Complete the sentence by filling the blank with suitable words or symbols:

“The chemical formula for sulphuric acid is _____.”

In the first format, the answer is to be supplied by writing a word or a set of symbols. In the second format, words or symbols are to be provided to complete the missing words in the sentence. These two forms of the supply type items are known as direct questions and the completion type respectively.

It is clear that the question given in the two formats in the above illustration requires the student to recall memorized information. Suppose another question formulated through this item is as follows:

- ii) “Suppose the price of a book is U.S. \$ 10, then the price of 20 copies of the book at the rate of \$1 = Rs. 50 will be in Indian Rs. _____.”

iv) Which part of a plant if damaged, would effect absorption of water?

This question requires some problem solving on the part of the student. Thus, supply type items can be used to measure the learning outcomes of the students, related to factual information, solving computational problems and even outcomes related to understanding objectives as in item (iii).

An important advantage of these items is that the student must either recall the information required or make the necessary calculations to solve the problems. This eliminates the possibility of guessing the correct answer. However, in this type of test items, sometimes the answers may be wrongly spelt. What is to be done if the teacher finds that some students have misspelt the answer? This may lead to inconsistency in scoring, once again reducing the objectivity of these items.

Consider a question such as the following:

“Who is the President of India?”

The teacher may be expecting the name of the president.

But a variety of answers are possible, such as ‘a man’, ‘an eminent person’, etc. Suppose the question is stated as:

“What is the name of the President of India?”

Now, because the answer required is stated in specific terms, the teacher may not find significant differences in the students’ answers as in the earlier case.

This indicates that the question or the blank in the incomplete sentence should require a specific answer. Nevertheless here also variation in answer is possible. For example, Dr. A.P.J Abdul Kalam, Dr. Abdul Kalam.

The supply type items can be formulated in two formats, the direct question or an incomplete sentence. Both forms are essentially the same, differing only in the matter of presentation. However, sometimes the direct question format is preferable as it is more familiar to the students. Consider the two forms of a question given below:

A ‘What is the name of the capital of India?’

B _____ is the capital of India.

C. Chandigarh is the capital of Punjab and _____.

You will agree that form A will be easily understood by the students, especially those who have language problem, being a direct question. In form B, they have to first understand the structure of the sentence and then the position of the blank. Thus, direct question form is generally, more efficient than an incomplete sentence, except where it makes the idea more concise. It may also be noted that placing the blank in the beginning of the sentence hinders the understanding of the structure. As far as possible, the blanks should be placed at the end of the statement, as in ‘C’ above.

Merits and demerits of very short answer type questions

Further, some teachers may take a sentence directly from the text book and remove a word to make it incomplete. But sometimes such questions may not make sense if they are taken out of the context. To avoid such mistakes one should formulate incomplete statements that make complete sense and require definite answers. It should also be kept in mind that a blank is kept for the key term.

Considering the merits and limitations of the supply type items, these should be used where they are necessary, for example, cases where students have to recall information,

solve computational problems, differentiate between concepts and the expected response is fixed which can be scored objectively.

Selection type or fixed response type items or questions

In selection type items, students have to recognize and select the correct answer from a given set of alternatives. The given set of alternatives can range from two to any number. Such items, where the correct answer is to be selected and not supplied, are called **objective type questions or items**.

(a) True-false or alternate response items

Where only two choices are provided, the item is known as true-false or alternate response item. Alternate response items have only two possible answers. Consider the following examples.

- | | | |
|---|---|---|
| i) 'Sugar is soluble in water'. | T | F |
| ii) 'The Sun revolves round the earth'. | T | F |
| iii) 'Particles of negative electricity are called neutrons'. | T | F |

In the above examples, items i and ii are statements of facts, and item iii is the definition of 'neutrons', which may be correct or incorrect. The student has to select the answer by recognizing the statement to be true or false. So, true-false items measure the ability of students to identify the correctness of statements of facts, principles, definitions, etc. It is possible to measure certain simple learning outcomes of students, which involve recall of learnt material, through such true-false items. Middle order competencies can also be tested provided the item writer has the needed technical know how.

It may be observed that a student will know that the statement in item (iii) above is false, which indicates that the given definition of neutrons is wrong. But, he may not know the correct definition of neutrons. This type of test items does not help us to check whether he knows the correct definition. Since only two alternatives are provided, right-wrong, true-false, yes-no, etc. possibility of guessing is quite high, may be 50% if equal number of true and false items are given in the test.

Because the answer is specific and pre-determined, these items can be assessed objectively. Scoring such tests is quite easy. But, there is wide scope for guessing. If a student has answered correctly 25 items on a test which consists of 50 items, one cannot know whether the student's score is due to guessing or it represents his actual learning outcomes.

Thus, true-false items can be said to possess **high objectivity** but **low validity**. They can be used to measure the student's learning outcomes pertaining to facts, definitions, concepts, etc. These items are also suitable for covering a large content area within a limited time span.

ii) Multiple-choice item.

In the multiple-choice item, the student is provided with more than two choices from which he has to select the correct answer.

The capital of Gujarat is _____.

- a) Baroda
- b) Ahmedabad
- c) Gandhinagar
- d) Rajkot

In this type of item, just as in the true-false type, the student is not given any freedom to write his answer. But he selects the correct or the best answer from the given set of alternatives. Since there is complete control over the student's response. This type of test item ensures 100% scoring objectivity. Possibility of guessing the correct answer in this type is also considerably reduced, because there are several choices from which the student has to select the correct answer. This reduces the chance of getting the correct choice by guessing. Therefore, it can be said that multiple-choice items are better than completion and true-false items, with respect to objectivity and reliability of scores.

A multiple-choice item consists of an incomplete statement as in the example above. It can also be given in the form of a direct question. The direct question or the incomplete statement is called the stem of the item, whereas the suggested answers are called alternatives. One of the alternatives is the correct or the best answer, called the key, other alternatives are provided to distract the attention of a student, who does not know the answer. These alternatives other than the correct answer are known as the distracters.

Consider the following multiple-choice item;

'India is _____.'

- a) ruled by the Britishers.
- b) situated in Africa.
- c) a Peninsula.
- d) in the southern hemisphere.'

Here, the stem of the item does not present a meaningful task. The distracters refer to widely separated ideas. The student is not presented with a clear problem. It is essential that in a multiple-choice the stem should always present a meaningful problem.

Further, in this kind of problem, since the alternatives are not homogenous in meaning, they may not really distract the student at all. A student who has only partial knowledge may be able to exclude those distracters which are not related to the idea and choose the correct answer. Thus, it may be noted that distracters should be homogenous in a multiple-choice item.

Multiple-choice items, besides being more objective and reliable in comparison with the true-false items, are useful for measuring even higher level objectives, such as, application analysis, synthesis, and evaluation etc. As an example consider the following item:

Constitutionally the Indian democracy can be described as a _____.

- a) one party system
- b) two party system
- c) multi party system
- d) dynasty ruled system

Here a student's understanding (not just memory) of the constitution of India is being measured.

v) Matching items

Another category of the selection type items is the matching type. In this type, a number of statements, words or phrases are given in one column, and a set of several alternatives are given in another column. Thus, all the items will have several alternatives, and the student has to match the words, phrases or statements in the two sets. The following is an example of matching type test item. Note that one list (i.e. Col. B) is longer than the other list (i.e. Col. A).

This is to reduce the influence of guessing (selecting) the last answer only by elimination. Consider the following example:

Directions: Column A contains a list of countries (Premises). Column B (alternatives) describes a form of government which was during the decade 1970-80. Match this description to the countries name. Each description in column B may be used more than once or not at all. Write the letter indicating a description (a or b to f) next to the number indicating a country in the bracket.

Match Column A with Column B

Column A	Column B
(Name of the countries)	(Rules by the form of Govt. during 1970-80)
1. United States of America	a) Monarchy
2. United Kingdom	b) President elected by the people
3. Pakistan	c) Constitutional Monarchy through elected cabinet Ministers
4. Soviet Russia	d) President elected by one political party.
5. Nepal	e) Military dictator
	f) Parliament

The matching item can be considered as a modification of the multiple-choice item. Instead of writing each item separately, a series of items are listed in one column and the responses are listed in the other column. The matching item is, therefore, the more compact format for measuring the same knowledge or understanding particularly when differentiation between concepts/procedures is to be tested. But, it must be remembered that unless all the responses in a matching item serve as plausible alternatives for each premise, the matching format remains ineffective.

We have seen that under the selection type item there are three specific types of test items. These are:

- i) True-false items
- ii) Multiple-choice items
- iii) Matching items

Essay type and objective type items compared

Objective type items are by definition easy to score. There has to be only one correct or the best answer. This objective scorability increases their reliability to a considerable extent when compared with essay type items. One should, however, keep in mind that a test may be highly reliable but not valid i.e., it may not be measuring the instructional objectives that it claims to measure.

Hence both essay type and objective type items should be judged in terms of reliability and validity. It is generally believed that essay type item measures originality, organisational ability, etc. But does it really help in doing so in the particular case of the specific exams? When should such an item be used? May be in that exam the essay item is measuring only the memorization. the same holds true in the case of objective type items. It is true that in most of the university exams, the objective type items are used to assess mainly knowledge, memory, etc. But this is more a fault of the items constructor than an inherent defect of the item type. We shall now see how the objective type items can be used to assess higher cognitive outcomes.

Self-assessment

6. Which one of the following types of items admit maximum guess in answering a question?
 - a) Essay items
 - b) Multiple-choice items
 - c) Short answer items
 - d) True-false items
7. In objective type items, the term "objective", refers to the method of
 - a) objective tested
 - b) form of question
 - c) method of scoring
 - d) mode of indicating answers
8. Systematic organization and logical presentation can best be measured through
 - a) Multiple choice type items
 - b) True-false items
 - c) Short-answer items
 - d) Essay type items.

Structured questions

These are sometimes referred to as interpretative exercises. We have discussed earlier that the essay type test serves the purpose of measuring certain higher level learning outcomes like analysis, application, etc. The objective type item, known as the interpretative exercise, can also be used for measuring a variety of complex learning outcomes. The interpretative exercise consists of a series of objective items based on a common set of data. The data may be presented in a variety of formats such as written materials, tables, charts, graphs, maps, etc. The interpretative exercises can be formulated in various objective item formats as discussed earlier. However, the most commonly used format is the multiple-choice type and also short answer questions. The students are presented a common set of data, and asked to identify relationships in the data, to recognize valid conclusions, to test assumptions and inferences, and to detect proper applications, of the data. An example of the interpretative exercise type items is given below:

Directions: Read the following comments which a teacher made about testing. Then answer the questions that follow the comments by encircling the letter relating to the best answer.

"Students go to school to learn, not to write tests. In addition, tests cannot be used to indicate students' absolute levels of learning. All that the tests can do, is to rank students in their order of achievement, and this relative ranking is influenced by guessing, bluffing, and are subject to the opinions of the teacher who does the scoring. The teaching-learning process would benefit only if we do away with such tests and depend on students' self-evaluation."

1. Which one of the following unstated assumptions is the teacher making?
 - a) Students go to school to learn.
 - b) Teachers primarily use essay tests.
 - c) Tests make no contribution to learning
 - d) Tests do not indicate a students' level of learning.

2. Which one of the following types of tests is this teacher primarily talking about?
 - a) Diagnostic test
 - b) Mastery test
 - c) Pretest
 - d) Essay test.
3. Which one of the following propositions is most essential to the final conclusion?
 - a) Effective self-evaluation does not require the use of tests.
 - b) Tests place students in rank order only.
 - c) Tests scores are influenced by factors other than achievement.
 - d) Students do not go to school to take tests.

(Best answer to questions 1,2 and 3 are c, d and a respectively).

It may be observed that in the above examples, item 1 measures one's ability to recognize unstated assumptions, item 2, the ability to interpret, and item 3, the ability to identify relationship.

Thus, objective type tests items as well as short answer questions can be used for measuring a variety of learning outcomes ranging from the sample to the complex.

Rearrangement exercise

There are many events which have occurred in chronological order. These could be historical events, scientific events, social events, etc. There are several specific processes having definite steps occurring in definite sequential order. There are structures where arrangement of different elements is in definite order. For all structures of these kind, you can develop items requiring rearrangement of given events, elements, etc. These items measure students' knowledge of the definite sequence in which the event or processes occur.

Some examples of rearrangement exercises are given below:

1. Rearrange the following events in their order of occurrence.
 - i) The civil disobedience movement
 - ii) The Jalianwala bagh Massacre
 - iii) The non-co-operation movement.
 - iv) The khilafat movement
 - i) _____
 - ii) _____
 - iii) _____
 - iv) _____
2. Rearrange the steps involved in the process of education evaluation in order.
 - i) Measure the learning outcomes.
 - ii) Use the technique of evaluation
 - iii) Compare the learning outcomes with objectives
 - iv) Select the evaluation technique
 - v) Judge whether achievement is satisfactory or not.
 - i) _____
 - ii) _____
 - iii) _____
 - iv) _____
 - v) _____

Self-assessment

9. The following items have been constructed on the basis of the types of items, discussed in the preceding sections, namely, essay items, multiple-choice items, matching items, and true-false items, multiple-choice items-interpretative. Now, go through all these items.

- Item: i) Critically discuss the examination system in Indian education.
- ii) Define and distinguish between the terms "selection and supply type items" and give one illustration of each.
- iii) Which one of the following statements is true?
- a) The reliability of essay type test is low because two different examiners may give different scores to the same answer.
- b) The validity of the essay type test is high because it is the type of test used in the past for centuries.
- iv) Select the most appropriate answer. The criterion-referenced test is:
- a) any test which measures achievement of the school subject.
- b) a test which helps a teacher to diagnose the difficulties of individuals.
- c) a test constructed on the basis of a detailed analysis of content and abilities.
- d) used to give ranks to students.
- v) Match the learning outcomes in col. A with those of domains of behaviour in col. B.
- (Each category in col. B can be used more than once or not at all).

Col. A

Learning Outcomes

- a) Ability to apply a concept to a situation
- b) Co-operative attitude
- c) Sporting spirit
- d) Skill of swimming
- e) Ability to frame hypothesis
- f) Typewriting

Col. B

Domains of Behaviour

- I) Cognitive
- II) Affective (emotional)
- III) Psycho-motor (Conative)

By now, you have gone through the above items. Based on these items, attempt the following question.

Match the number of item (Col. A) with the type provided in (Col. B). The same label could be used more than once or not at all.

Col. A

- Item No. i
- Item No. ii
- Item No. iii
- Item No. iv
- Item No. v

Col. B

- a. True-false
- b. Multiple-choice
- d. Essay type (Vague)
- c. Essay type (Specific)
- e. Rearrangement exercise
- f. Matching type
- g. short type

10. Compare items i and ii. Which format would you prefer and why?

11. Compare items iii and iv. Which format is less influenced by guessing and why?

Preparing a good question paper

In the preceding sections, we had discussed the different types of questions and their suitability with regard to specific learning outcomes. It can, therefore, be concluded that a good question paper, which is intended to evaluate the achievement of students with respect to a variety of specific learning outcomes, may include different types of test items that are relevant to the specific instructional objectives.

Let us consider the necessary steps in preparing a good question paper. The first step for preparing a good question paper is to develop a design or framework. For this, the test developer has to:

- analyze the course content into different content units and decide the weightage that is to be given to each in the test.
- decide the weightage to be given to different objectives being tested.
- decide the weightage to be given to different forms of questions to be used in preparing a question paper.
- decide the weightage to be given to time and marks for different forms of questions.
- decide the weightage to be given to the difficulty level in the test.

The second step is to prepare a table of specification, i.e. the blueprint, which reflects distribution of various types of question to be set on different content units testing particular objectives, such as knowledge, understanding, application, skill, etc. You will most likely need essay type, short answer type as well as objective type items to evaluate the learning outcomes of students related to objective at varying levels. Let us see how a table of specifications (blueprint) is developed.

Preparing a Table of Specifications (Blueprint)

As discussed earlier, the first step in preparation of a Table of Specifications is to give the weightage to different content units. Let us take the example of the blueprint given in Unit 11 of Block 3. There are six units and the question paper will be based on these units. The weightage given to each unit is presented in Table 1.

Table 1: Weightage given to different content units

<i>Content Units</i>	<i>Weightage given</i>
Unit-1	10
Unit-2	20
Unit-3	15
Unit-4	20
Unit-5	20
Unit-6	15
Total	100

The second step in the preparation of blueprint is to give weightage to different objectives. The weightage given to different objectives is presented in Table 2.

Table 2: Weightage given to different instructional objectives

<i>Instructional Objectives</i>	<i>Weightage given</i>
Knowledge	40
Understanding	30
Application	20
Skill	10
Total	100

The third step in the preparation of blueprint is to give weightage to different forms of questions. The weightage given to different forms of questions is presented in Table 3.

Table 3: Weightage given to different forms of questions

<i>Forms of Questions</i>	<i>Weightage given</i>
Essay type questions	30
Short answer type questions	45
Objective type questions	25
Total	100

The fourth step in the preparation of question paper is to give weightage to marks and time for different forms of questions. The allotment of marks and time to different forms of questions is presented in Table 4.

Table 4: Weightage given to marks for different forms of questions

<i>Forms of questions</i>	<i>Marks per question</i>	<i>Total no. of questions</i>	<i>Total marks</i>
Essay type	10	03	30
Short answer type	03	15	45
Objective type	01	25	100
		43	100

The weightage given to time for different forms of questions is presented in Table 5.

Table 5: Weightage given to time for different forms of questions

<i>Forms of questions</i>	<i>Time per question (in minutes)</i>	<i>Total no. of questions</i>	<i>Total time (in minutes)</i>
Essay type	20	03	60
Short answer type	05	15	75
Objective type	01	25	25
		43	160

The next step is to give weightage to difficulty levels of the items, which is presented in Table 6.

Table 6: Weightage to difficulty levels of the items

<i>Difficulty levels</i>	<i>Weightage given</i>
Difficult items	25
Average items	50
Easy items	25
Total	100

Based on the above steps of preparation of a blueprint, the final blueprint is developed. With the help of such a table of specifications, you will be able to ensure the needed validity of adequate coverage of syllabus units and assessment objectives. The final blueprint is presented in Table 7.

Table 7: Blueprint (Table of Specifications)

Objective→ Form of → Question ↓ Content	Knowledge			Understanding			Application			Skill			Total
	E wise	S	O	E	S	O	E	S	O	E	S	O	Unit
Unit/ Theme-1	-	3(1)	2(2)	-	-	2(2)	-	3(1)	-	-	-	-	10
Unit/ Theme-2	-	3(1)	1(1)	10(1)	-	2(2)	-	-	1(1)	-	3(1)	-	20
Unit/ Theme-3	-	3(1)	1(1)	-	6(2)	1(1)	-	3(1)	-	-	-	-	15
Unit/ Theme-4	10(1) 10(1) (a)	-	1(1)	-	3(1)	1(1)	-	-	2(2)	-	3(1)	-	20
Unit/ Theme-5	7(1)	-	4(4)	-	-	2(2)	-	3(1)	1(1)	3(-)	-	-	20
Unit/ Theme-6	-	3(1)	2(2)	-	3(1)	-	-	6(2)	-	-1(1)	-	-	15
Sub-total	17 (2)	12 (4)	11 (11)	10 (1)	12 (4)	8 (8)	-	15 (5)	5 (5)	3 (-)	6 (2)	1 (1)	100 (43)
Total		40			30				20		10		100

Summary: Internal option (a) in Essay type question (E) in Unit 4 under knowledge objective column 1E.

Entries made are only to illustrate. Paper setters have to decide about these while preparing the blueprint. However, it must confirm to weightages indicated in the **design** to the various objectives (40, 30, 20, 10), content units (10, 20, 15, 20, 20, 15) and form of questions E = 30%, S.A. = 45%, O.T. 25%) as reflected in this table of specification or blueprint.

Self-assessment

12. Which one of the following steps is relevant in the preparation of a table of specifications?
- Identifying the learning outcomes and content areas to be measured by the test.
 - Giving weightage to the learning outcomes and content areas in terms of their relative importance.
 - Deciding the forms of question to be included and giving weightage to each type.
 - Building the table in accordance with these relative weights by distributing the test items proportionately among the relevant cells of the table.
 - All of the above.

Summary

In this unit, we discussed framing of questions of different types in setting a question paper. In this context, we pointed out that before framing a question paper, it is necessary to analyze the subject matter into content units and give weightage to each unit and instructional objectives under each unit. The third aspect to keep in mind while framing a question paper is the development and selection of a variety of question to be included. Under "types of question", the major types are essay, short-answer and objective type questions. Objective type items are of different types. We discussed the nature, merits and demerits of these types of evaluation items. We also highlighted the functions of interpretive and rearrangement exercises. Having discussed the three main

steps in the designing of a question paper, we suggested a table of specifications which is also called a Blueprint. This Blueprint is based on a design which talks about the weightage given to learning outcomes to be achieved and the subject matter to be covered and the weightage to be given to various types of questions. However, a question paper, in order to get the final shape, must pass through other steps too. In the next unit of this Block, we will have more discussion on the Mechanics of Evaluation.

Unit-end activities

1. Write three items each of the following objective type items on a topic of your choice, testing different instructional objectives.
 - a) True-false type
 - b) Multiple-choice type
 - c) Matching type
2. Construct three essay type questions and prepare a model answer and scoring key for each question.
3. Construct three short-answer type items testing different objectives.
4. Construct one item of rearrangement type and one of interpretive type.

Points for discussions

1. If you decide to introduce more objective type test in your college/in the subjects you are teaching, what precaution you would take and the arrangements you would make so as to make a question bank of such items more useful?
2. The large selection agencies like U.P.S.C. and Banking Services Recruitment Boards use objective type items in their selection tests. Why?

Suggested readings

- Conningham, G.K. (1986) *Educational and Psychological Measurement*. New York: Macmillan.
- Ebel, R.L. (1979) *Essentials of Educational Measurement*. (3rd ed.), Englewood Cliffs, N.J.: Prentice-Hall.
- Gronlund, Norman E. (1990) *Measurement and Evaluation in Teaching*. (6th ed.), New York: Macmillan, and London: Collier Macmillan.
- Groulund, Norman E. (1968) *Constructing Achievement Tests*. Englewood Cliffs, N.J.: Prentice-Hall.
- Mehrens, W.A. and Lehman, I.J. (1978) *Measurement and Evaluation in Education and Psychology*. (2nd ed.), New York: Holt, Rinehart and Winston.
- Harper, A.E. and Haper, E.S. (1990) *Preparing Objective Examinations — A Handbook for Teachers, Students and Examiners*. New Delhi: Prentice-Hall of India.
- Weimen, M. (1996) *Improving Your Classroom Teaching*. Newbury Park: London, New Delhi: SAGE Publications.
- Yadav, M.S. and Govind, R. (1977) *Educational Evaluation A Package of Auto Instructional Material*. Ahmedabad: Sahitya Mudranalya, Kankaria Road.
- Freeman, Richard and Lewis, Roger (1998) *Planning and Implementing Assessment*. London: Kogan Page.

Answers to self-assessment

1. It is necessary to specify the expected learning outcomes while framing a question paper, because through this the teacher knows whether the instructional objectives set by him/her have been achieved by students.
2. i) Students will be able to define the concept of democracy.
ii) Students will be able to explain the causes of earthquake.
3. Ideally, a question paper should represent all the content areas of the course. If this is not done, the student's assessment is hampered. Because, if some content areas are left out in the question paper, we fail to measure the behavioural changes of the students in those specific areas. Hence, prior to framing a question paper, the subject content should be completely outlined and duly represented.
4. Essay type items suffer from the following demerits:

- a) low content validity
- b) low objectivity
- c) low reliability

Such type of items can be improved, if the following aspects are taken care of:

- a) An essay item evaluates one specific educational objective.
- b) It indicates the required answer.
- c) Standard scoring key specifying various aspects of the answer is prepared.
- d) Standard answer is developed.

5. Short answer questions expect students to supply answers in terms of sentences related to a single or more than one concept.
6. b)
7. c)
8. d)

9. Col. A

- Item No. – i
- Item No. – ii
- Item No. – iii
- Item No. – iv
- Item No. – v

- Col. B

- d – Essay type-vague
- e – Essay type specific
- a – True-false
- b – Multiple-choice
- f – Matching type

10. The specific format of the essay type is better, as the student knows precisely what is expected of him.
11. Multiple-choice format is less prone to guessing, while in true-false type, the guessing probability is fifty-fifty, in multiple-choices, of 4 alternatives, this is reduced to 25%.
12. All of them.

Introduction

In the preceding unit, we have discussed the process of framing test items and question papers. In that context, we discussed how to outline the subject matter content, how to prepare different types of test items and how to develop a blueprint or a table specifications. Before a test item or the question paper is given the final touch, it has to fulfill certain other requirements. Any tool of evaluation, particularly achievement tests, to be effective, should satisfy certain characteristics and each item in that tool should be made valid by certain parameters. In the present unit, we will study about essential characteristics of evaluation tools, i.e., validity, reliability, objectivity and usability. While discussing the parameters concerning a test item, we will discuss item analysis and item indices. Towards the end of the unit, we will discuss question banks and their role in framing the question paper and improving the teaching learning process.

Learning outcomes

By the end of this Unit, you will be able to:

- identify the characteristics of different types of evaluation items and differentiate between them;
- describe the procedures to find out validity, reliability, objectivity and usability of a given test;
- calculate the parameters of items, such as difficulty level and discrimination index and also interpret them;
- describe the importance and uses of question banks; and
describe the qualities of a good question paper.

Different types of evaluation tools and their characteristics

Teachers need a variety of techniques and tools for evaluating students. Teacher may like to diagnose learning difficulties of students through diagnostic test, assess students learning through achievement tests, mental abilities through intelligence tests, special abilities through attitude scale and interest through inventories. Teachers also assess students' performance in seminars, interactive sessions and practicals. In the faculties of performing arts, performance of students in music, dance, etc. is assessed through rating scales and checklists. It may be pointed out here, that intelligence, aptitude, personality, attitude, interest, etc. are assessed by the teachers rarely and whenever

assessed, they use ready made standardized tools available for these purposes. However, achievement of students is evaluated in different subjects by all the teachers in higher education. For this purpose, all the teachers use achievement tests. In addition to this, checklists and rating scales also are used by large number of teachers for evaluating written work, reports and performance of students in various activities like dance, music, seminar, etc.

It may also be noted that achievement tests have to be prepared by all the teachers in higher education. In addition to this, rating scales and checklists also have to be prepared by a large number of teachers in higher education. In order to be effective, the tools of evaluation used by teachers, particularly the achievement tests and rating scales, should satisfy certain essential characteristics. These essential characteristics are discussed in the following section in more details.

Essential characteristics of an effective tool of evaluation

It is known that results of evaluation are based on information obtained from the tools of evaluation. Therefore, to get accurate results of evaluation, one should select a proper tool. The essential characteristics to be looked for, while choosing such a tool of evaluation are validity, reliability, objectivity and usability. Let us discuss all these.

Validity

We know that tools of evaluation are used to collect information about students' achievement. The collected information is assessed and used for the purpose of determining the extent to which instructional objectives have been achieved. So the tool of evaluation is used for a specific purpose. The efficiency of a tool in providing accurate information which would serve the purpose is an important characteristic. If the tool has this efficiency, that, it serves the purpose for which it has been developed then we say that the tool is valid or it possesses the characteristic validity. Thus validity of a test can be defined as:

'The extent to which a test serves the purpose for which it has been prepared is called its validity.'

Example

- An intelligence test prepared by you should measure intelligence only (not achievement or aptitude) and a student's performance on it should indicate high scores of highly intelligent students and low scores for less intelligent scores.
- Railway pass serves the purpose for the period it is issued. As soon as date is over, railway authorities check and tell that it is not valid now.
- If dictation is used for the purpose of testing spellings of students it is valid upto the extent it provides evidence on spellings, but its validity is affected or lowered since purpose of speed testing also creeps in.

The tests are used for various purposes and accordingly validity is also categorized into different types which are hereinafter discussed.

Types of validity

i) Content validity: Since tools are used for different purposes in educational institutions, one comes across different types of validity. Some tools are used by teachers to obtain information which indicates the extent to which students have learnt the specified content material in different subjects in terms of instructional objectives. When a tool serves this particular purpose, it may be said to possess content validity.

This type of validity is related to the content area that is being tested. Thus, a tool, such as a written test which is generally used by teachers, is said to possess content validity to the extent the information obtained through it indicates the student's achievement with respect to specific areas of learning. One can assess whether a test has content validity by checking with the help of a table of specifications, whether all the important cells of the matrix are represented by one or more questions in the test related to content units and instructional objectives. To the extent the test is representative of all the cells in the matrix, it is valid. You may refer to the table of specifications given in unit 15.

ii) Predictive validity: Test performance of students is sometimes used for predicting their future performance in the particular area that is tested. When a test serves this purpose of predicting future performance on the basis of information obtained through it, it is considered to have predictive validity. A test may be said to possess predictive validity to the extent the information obtained through it serves the purpose of predicting the future performance of students in learning Mathematics. Usually this is studied by correlating the score on the test with some criterion scores (performance) which is likely to occur in future. For example, admission tests for medical, engineering, etc. are used for admitting students who will perform better in future when they actually study in medical or engineering courses. These admission tests and other selection tests have predictive validity. The extent of predictive validity depends on how well the criterion test (achievement test, admission test, selection test, aptitude tests) measures these abilities which are pre-requisite for success in the professional courses like engineering, medical, etc.

iii) Concurrent validity: As in predictive validity, here also the test scores are correlated with another set of criterion scores which are presently (concurrently) available. For example, in order to validate a test of scholastic ability one may study its correlation with the final examination marks (already available) obtained by the same student. However, it must be ensured that the test which is presently in use is valid. For predictive validity, one would predict on the basis of the test scores attained by the students on the scholastic ability test given now with their performance in the college final examination in future.

As both concurrent and predictive validity are established on the basis of empirical evidences they are called **empirical validity** and involve some correlational-statistical investigation to establish a relationship which is expressed in quantitative terms.

iv) Construct validity: Some tests are used to measure certain psychological characteristics or constructs of students, such as their intelligence, creativity, motivation to learn, etc. These characteristics are abstract in nature and are defined in terms of certain psychological concepts or principles. For example, consider the construct 'intelligence'. The test of intelligence is not like a test of English Literature or History of medieval period or electricity in Physics, etc. which are achievement tests. On the contrary, the constructs like intelligence, anxiety, etc., cannot be defined by a specific content or operation from a subject area. Then how do you define these constructs? This is a complex problem. Usually psychologists define it by the relationships such a construct has—positive/negative correlations with various behavioural indicators. Various hypotheses of such correlation are postulated that intelligence possessed by a child increases as his age increases. Therefore, if an intelligence test is administered to a boy at different age levels and the results are noted, these test results should indicate that intelligence level increases as his age level increases up to a certain stage. Results of an intelligence test should be explainable in terms of such correlations with, say, scholastic achievement, leadership and other phenomena. A test can be considered as valid to the extent results obtained through the test can be interpreted in terms of the known psychological concepts and principles about the particular construct. This type of validity is known as the **construct validity** of the test.

Each of the four types of validity has been discussed with reference to a particular purpose of evaluation. Thus, whenever the validity of a tool is discussed, it must be with reference to a specific purpose for which the information obtained through it has to be used. You will agree that the teacher-made classroom tests and examinations, which are conducted to measure a particular achievement of students, should possess content validity. The other measures of validity are more a matter of research and are more useful in the selection situation. As a teacher you may take up such a research. But the minimum you should ensure is the content validity in terms of the table of specifications in terms of the instructional content and objectives that you have stated in your Lesson Plan or Topic Plan. (See Unit 5 of Block 1).

Self-assessment

1. Select students for a new course of study by administering a test to them. What type of validity is essential in such a test?

Reliability

Suppose an achievement test in Mathematics is administered to students and it is found that one student gets a score of 60. This score is considered as an indicator of his actual performance. But, can this score be considered as a reliable indicator of the student's actual performance or learning outcome? Suppose the test contains several questions that the student could answer simply by guessing and he got almost all his guesses correct. In such a situation, the student's score is not a reliable indicator of his actual performance, because a part of the score is due to the error made during measurement by giving marks to answers which do not represent actual learning in the student, but was obtained through guesswork. In another situation, a student knows the mathematical concepts and answers involved in all questions in that test, but the language used in the questions is such that he cannot understand what he exactly has to do. He writes the wrong answers to the questions that he could not understand and gets a low score on the test. Still in another case a student has studied only a part and does not have questions from the other parts. A student may get high marks in that paper; but on another occasion, he may even fail. In all such cases, the scores on the test do not represent the actual learning outcome of the students, because of the errors involved in the measurement are brought about by the nature of the test. Scores obtained on such tests cannot be considered as reliable indicators of the students' actual performance or achievement. Remember, due to the errors of measurement, test is not measuring actual performance of students. You will notice that performance of students will change if such a test is administered to students second time. If performance is actual, it should remain the same or consistent. Therefore, a test which gives consistent results over a period of time may be considered reliable.

Thus, we can say that reliability is the extent to which results obtained through a test remains constant over a period of time or in different situations for the same sample.

i) Test-retest reliability

How to check the reliability of a test? It has to be checked whether there is any error of measurement involved in obtaining the scores or if they represent the actual amount of learning in the students. Suppose a test involving certain mathematical operation was given to a student. You would want to check if the student has learnt the mathematical operations or not. One way of doing this is to ask the student to do the mathematical operation (through the test) again. You may consider that there was no error involved in doing the mathematical operation if the scores obtained again are the same as those obtained the first time. You may check whether a test is reliable or not by giving the test again. The test is considered to be reliable or not by giving the test

again. The test is considered to be reliable if the scores obtained through the test and retest situations are the same (or similar). In other words, reliability of a test is specified by the consistency (stability) of scores obtained through it over a period time. As the reliability in this situation is checked by comparing the test and retest situation, it is called re-test reliability.

ii) Equivalent-forms of reliability

A test has been developed to evaluate the learning outcomes of students with respect to the specific educational objective given below:

“Students solve simultaneous equations correctly”.

Suppose another test is developed with an alternate set of problems equivalent to each problem on the first test with respect to the same specific instructional objective. The second test can be considered as an equivalent form or parallel form of the first test. One could expect the scores of students on the two equivalent forms of the test to be the same. In other words, if a student's score on an equivalent form of the test represents his actual learning, he should then get a similar score on an equivalent form of the test. Thus, a test may be considered reliable, if the scores obtained on it are consistent over the two equivalent forms. This is called as **equivalent-forms method** of estimating reliability.

Support, Form-A and Form-B are two equivalent forms of the same test. When the two are admitted to the same examinees, their scores on both should be same or similar. The extent to which these scores on the two equivalent forms of the test correlate determines the reliability of equivalent forms, showing consistency of scores on the parallel forms of the test.

iii) Split-half reliability

A test is made up of many questions. It can be said that all the questions in a test together try to measure learning related to a particular aspect. Therefore, if the test is split into two halves, they would represent two equivalent forms, and each of them would still measure the same aspect. One would expect that the total scores of students on each half should be consistent. Of course, here it is assumed that the splitting of the test into two halves is done in a way that each half is equivalent in terms of content elements, objectives, or competency tested, form of question and the time required in attempting each half. Here, the reliability of the test is specified in terms of the consistency of the scores over the two halves of the test. As reliability is checked by splitting the test into two and considering the core on each half, it is known as **split-half** reliability.

It may be noted that in all the three approaches of checking reliability, a test's reliability is specified in terms of the consistency of scores obtained through it; it could be consistency over time (Test-Retest), or consistency over tasks (items) as in equivalent forms or split half procedure.

Self-assessment

2. What type of relationship is there between validity and reliability of a test?

.....
.....

3. What will happen to the reliability of the scores for a multiple-choice test, when the number of alternatives of each item is changed from three to four?

- a) It will increase
- b) It will decrease
- c) It will remain the same

Suppose a teacher has conducted a test and he has checked the answers and assigned marks. A student's marks on the test should depend upon the answers that he has given and not the subjectivity of the scorer who is scoring it. When a student's score on a test is based only on the student's answers and not on the personal biases of the scorer (or teacher), then the scoring is considered to be objective. This depends on the procedure of the scoring (and the type of item to be scored). If the procedure is such that even when different scorers (rators/examiners) score the item independently give the same score, such a test is called objective test. Thus, objectivity in scoring may be considered as **consistency in scoring** by different scorers. When the question is such that the response is not fixed but some freedom of response is there for the examinees, then some subjectivity in scoring creeps in, and the complete uniformity in scoring cannot be ensured.

How can this consistency in scoring by different scorers (examiners) be ensured?

Let us consider the following question used in testing:

“Describe in detail the meaning of the term ‘reliability of a test’.”

Suppose several competent teachers have independently scored a student's answer, which contains several important points about the meaning of reliability of a test. Experience suggests that one cannot expect different teachers to give the same marks for the answer given by the students to such a question, because teachers differ in their opinion about the correctness or suitability of the answer. Thus, scoring in this case is not likely to be objective, because it is based not only on the student's answer but also on the opinion of the teacher who is scoring it.

Let us consider what happens if the question is stated differently as follows:

“Explain in about eight to ten lines, the meaning of the term ‘Reliability of a test’.”

You will agree that the scores assigned by different teachers to the above question will not differ to a great extent now. This has happened because the reworded question clearly indicates expected scope of the answer. Objectivity in scoring also refers to the consistency in the scores assigned by different examiners (teachers), i.e. uniformity in scoring.

But here also the nature of expected responses is not clear. Students' answers may vary in terms of definition, factors affecting reliability, methods of estimating the test reliability, reliability related to validity and objectivity, etc. It will again bring in subjectivity of score. What is needed is precise working of the question. Note the following questions:

1. List four factors that improve reliability of a test.
2. List three methods of estimating reliability of a test.

In these two questions, better consistency in scoring the answer is ensured, which is not possible in the earlier two questions where personal bias of scorer does creep in.

What would happen if the item is a multiple-choice? For example:

Test-Retest reliability indicates consistency of scores:

- a) over time
- b) in sampling of items
- c) with different rators
- d) in each half of the test.

In this type of item, there is no possibility of different examiners scoring it differently. All of them shall have to give marks only if the student has checked (a) as his correct answer. It is, in this sense, that objective type test is more reliable, i.e. consistent over different raters.

In the above paragraphs, reliability is defined in terms of consistency of scores. And, scoring objectivity is now being defined in terms of consistency among different scorers in assigning scores to the answer on a test. Thus, scoring objectivity can be considered as **inter-scorer reliability** (or inter-rater reliability).

Self-assessment

4. Why is scoring objectivity known as inter-scorer reliability?

.....
.....

Factors affecting the reliability

Reliability is an important characteristic of a tool of evaluation, indicating the extent to which a test measures the actual performance of the students. Therefore, we should know the factors which affect reliability so that we could develop reliable test. Below are given some factors which affect reliability so that we could develop reliable test. Below are given some factors which affect reliability.

i) Errors of Measurement

These errors are related to the test items. Sometimes item is ambiguous. It does not clearly indicate what is required of students. Due to the misunderstanding of the question, a student may write the wrong answer although he knows the correct answer too. Another type of error comes in the measurement when items provide scope of bluffing or guessing to the students. Therefore, one should develop items which are unambiguous and restrict the possibility of bluffing and guessing. It may be noted that essay type questions have more scope of vagueness and bluffing. Guessing is maximum in case of items which have only two alternatives.

ii) Length of the Test

Tests which are short and do not cover the behavioural changes in students adequately are less reliable than longer duration test covering behavioural changes comprehensively. It may also be noted that essay type test with options having the same length as an objective and short answer type test would be less reliable. However, in general, longer the test, greater is its reliability. It should also be taken into consideration that length of the test should not exceed 3 hours writing in terms of time.

iii) The scoring procedure

If examiners use different yardsticks and expect different answers then marks will not represent actual performance of students.

iv) The time gap between two administration

When method of establishing reliability requires test to be administered twice (test-re-test and equivalent-form reliability), the time gap should not be too long. If time is long, a student might forget or learn more which would change his score in second administration.

Placement of items in the test also can affect the reliability of the test. As a teacher if you place most difficult item of the test in the beginning, it may disturb mental set of many students particularly sensitive ones. It would affect students' performance on other items too. Therefore, answers will not represent actual performance of students. It is always desirable to arrange items in the test from easy to difficult.

Is reliability a sufficient condition for validity?

It may be mentioned that the essential characteristics of tools of evaluation, viz., validity, reliability and objectivity are interrelated. However, it may be noted that if a test is valid, it has to be reliable to a great extent. But reliability is not a sufficient condition for the validity. For example, consistency of a test can be increased by having more questions which require definite and restricted answers, but if this test measures only knowledge, and not other instructional objectives, it cannot be considered valid. Thus, we can say that reliability is not a sufficient condition for validity.

Similarly, if you prepare a test which contains objective type items, then, the test will be quite reliable. But, think of a situation when one prepares these objective type items only on 50% of the specified content. This shows that test will be objective but it will not be that valid. Relevance and reliability both are essential components of validity.

Usability of the Test

In addition to reliability and objectivity, another characteristic of the test is also important. This is the usability of the test. You know that a variety of tools are required for evaluating students with reference to different instructional objectives.

Different tools are likely to differ with respect to the "procedures of administration", "time required" and "cost involved". All these aspects induce a teacher to use tests or other tools of evaluation. Before using a test, teacher would like to ensure, that, the test's procedure of administration and scoring should be easy and it should not be very much time consuming and costly. These practical aspects are referred to as the usability of a tool of evaluation. Therefore, a test which is easy to construct, administer, score and interpret is usable besides being cost effective.

It may thus be observed that a teacher should look for **validity, reliability, objectivity** and **usability** while choosing a particular tool of evaluation.

Self-assessment

4. List four factors which reduce the reliability of a test:

- i)
- ii)
- iii)
- iv)

Parameters concerning an evaluation item

After a question paper has been prepared, it has to be administered to the students. A good question paper, you will agree, should provide clear directions to the students. Such directions should include amongst other things, information regarding the following:

- Purpose of the test
- Time allowed to complete the test

- Total marks for the test
- How to record the answer
- Whether guesses are allowed when in doubt about the correct answer
- Weightage given to the different questions or parts of questions

The answer scripts are to be scored after the test is over. Scoring procedure for objective type items and essay type items will be different. Objective type items can be scored by any person or sometimes by machines (computers), if the scoring key is provided. Essay type questions have to be scored only by those teachers or scorers who are familiar with what has been taught. Sometimes, correction for guessing formulae are applied while scoring objective type items in which answers can be guessed. One such formulae is as follows:

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

In this formulae “n” equals the number of alternatives in each question.

Suppose, a student scores 40 items right and 9 items wrong on a 50 item multiple choice test, using four alternatives in each item, his corrected score will then be as follows:

$$40 - 9/3 = 40 - 3 = 37$$

It is, of course, not necessary to use any such formulae if the test is made by the teacher for classroom use and the discussion/guidance to students after his scores are communicated to him. Even in a selection situation, if the number of items in that test is large (above thirty), it is not necessary to use this formulae as it does not yield better discrimination.

Item analysis

After a test paper has been scored, it is desirable to appraise the effectiveness of the different questions or items. This can be done by considering the responses of students to each item. Such a procedure is called **item analysis**. When systematically done, item analysis would provide information regarding the following aspects:

- The difficulty level of the item.
- The discriminating power of the item.
- The effectiveness of each alternative.

We can find out through item analysis whether an item was too easy or too difficult, to what extent it was able to discriminate between high and low achievers and the effectiveness with which the different alternatives functioned. Such an information may be useful in improving the items for later use. The teacher also gets useful information from item analysis for providing feedback to students regarding their performance in the test.

For example, answer to difficult questions can be explained fully, common errors and misconceptions of students pointed out and the remedial learning activities suggested.

In the section that follows certain simplified item analysis procedures will be discussed.

Difficulty level

The difficulty level of a test item is indicated by the percentage of students who answer the item correctly. The following formula can be used to estimate the difficult value of an item.

$$\text{Difficulty level} = \frac{R}{T} \times 100$$

Where R = The number of students who get the right and

T = the total number of students who have tried the item.

A simple but effective procedure for calculating the difficulty level of an item is given below. Let us take 32 papers for item analysis. The following steps illustrate the procedure:

- i) Arrange all the 32 papers in order from the highest score to the lowest score.
- ii) Select approximately one-third of the papers (say 10 papers) with the highest scores and call this the upper group.

Similarly, select the lower group of the same number of one-third papers (10 papers) from those which have the lowest scores. The middle group of papers (12 papers) need not be used.

- iii) For each item, count the number of students in the upper group who have selected the correct answer. Make the same count for the lower group. If the item has a number of alternatives, the response of students in the upper and lower group can also be noted.

- iv) Record the count as follows:

Item 1:	Alternatives	A	B*	C	D
	Upper Group (10)	0	6	3	1
	Lower Group (10)	3	2	2	3

* Correct answer.

- v) Estimate the item difficulty level by the percentages of students who get the item right in the upper and lower group. The difficulty level of the item in the above example can be computed as follows:

$$\text{Difficulty level} = \frac{(6+2)}{20} \times 100$$

Although calculation of the item difficulty is based on the performance of students in the upper and lower groups only, it provides a close approximation to the estimates obtained by considering the performance of the entire group. It may be noted that the difficulty level of an item refers to the percentage of students getting the item right. The smaller the percentage figure, the more difficult will be the item. And because of this nature, it is sometimes called facility index (rather than difficulty index), as it indicates the number of students who could answer the item with facility, i.e., correctly. If you are constructing a Criterion Referenced Test (CRT) or any other test for classroom (instructional) use, you may use this information about difficulty as a feedback to students, i.e., discuss those topics once again with students. But if you are constructing a Norm Referenced Test (NRT) in the case of final examinations or in selection situation, you may not like to include too many difficult items in your question paper, neither would you like to have too many easy items in it. Such test makers usually like to have items at the 50 per cent level in a test. Too many easy or difficult items in a test paper, will reduce its effectiveness in terms of its capability of discriminating among students. This is what one should remember when designing a test for final examination which is to be used for ranking purposes.

The procedure described above is suitable for computing the difficulty level of objective type items. But in the case of essay type question, students are not categorized as successful and unsuccessful, but would score between zero and the maximum marks assigned to each question. The answers to essay type questions can be partially correct. If such partially correct answers are treated as correct answers for computing the difficulty level of the question, the figure so obtained may be erroneous. It would be better to assign a partial weightage to a partially correct answer. For example, if a

student is given four marks for essay question for which the maximum marks assigned were ten, the student's marks for that question can be considered to be 0.4. Similarly, scores of other students can be converted into such proportions. Then the scores of all students can be averaged out to estimate the difficulty level of the essay type question.

Index of discrimination

The discriminative power of a test item refers to its ability to discriminate between the high and the low achievers. One would expect that more students from the high scorers, i.e. the upper group will get an item right in comparison to the number of students in the lower group, if an item is effective. The degree to which a test item discriminates between students with high and low achievement is known as its **index of discrimination**. For computing the index of discrimination, you can use a simple procedure. Just subtract the number of students in **lower** group who have got the item right from the number of students in the **upper** group who have got it right, and divide this by the total number, in each group. Thus, for item 1 in the above example, the index of discrimination would be as follows:

$$\text{Index of discrimination} = \frac{6-2}{10} = .4$$

It may be observed that the index of discrimination is indicated by a decimal fraction. The discrimination power of an item may range from -1.00 to 1.00. You may like to include items in the test, with higher values of the index of discrimination. Items with lower values and those with negative indices should be dropped from the test or modified before revising.

Effectiveness of alternatives

When alternative answers are provided in objective type test items, one would expect that for an effective item, more students from the upper group will choose the correct alternative in comparison to the students from the lower group. Let us consider the responses of the students to an item in the following illustration:

Alternatives	A	B*	C	D
Upper group (10)	5	5	0	2
Lower group (10)	4	3	0	5

*Correct answer.

It can be seen from the above illustration that the discriminating index of the item, as indicated by the choice of the correct alternative, is not high ($D = .2$), which may be due to the distracters being poor. It may be noted that distracter A is poor, because it attracts more students from the upper group than the lower group. Distracter C is completely ineffective as it has attracted none of the students. Only distracter D is functioning effectively, as it has attracted a larger portion of students from the lower group in comparison to students from the upper group. Thus, the effectiveness of the distracters in objective type items can be determined by an inspection of the responses of the students in the upper group. Thus, the effectiveness of the distracters in objective type items can be determined by an inspection of the responses of the students in the upper and lower group. Poor distracters suggest that the item is ineffective and needs to be revised. It should be noted that a good distracter will attract more students from the lower group than from the upper group.

We have so far discussed only the simplified procedures for analyzing test items. These procedures can be modified according to the demands of the testing situations. Sometimes, merely inspecting the test scores may be all that is necessary. It may be

desirable to select the upper and lower groups from the top and bottom 25 per cent of the students, if the group is large. If the group is small, the upper and lower halves may be considered. Selecting the top and bottom 27 per cent of the group is recommended for more refined item analysis and is rarely required for analyzing teacher made classroom tests.

Interpreting item analysis data

Item analysis information should be used with caution, particularly when it relates to the performance of a small number of students on classroom tests. Difficulty level and index of discrimination of a test item are expected to vary from group to group. It is not proper to set minimum level of the discriminating power of a test item for inclusion to set a question paper. Similarly, limits cannot be set to the levels of difficulty, within which items will be acceptable. As stated earlier, other things being equal, in a NRT situation, you should favour items around the 50 per cent level of difficulty and those with highest levels of discriminating indices. However, if an item does not show very high or low level of difficulty, it provides a positive index of discrimination and if all the distracters are functioning effectively, it should be retained. All items retained should be placed in a file for future use. It would be a good practice if each item is written on a separate card, which is then filed. Each time an item is used, data regarding the item analysis may be recorded on the card. Such data will show the variability of an item's difficulty level and discriminative power when used with different groups of students and make the information more interpretable.

Analysis of criterion-referenced tests

In unit 13, you have studied about criterion-referenced evaluation and norm-referenced evaluation. Our discussion about item analysis in the preceding sections was related to norm-referenced testing procedures. The item analysis procedures used for norm-referenced test items are not applicable to criterion-referenced mastery tests. The purpose of criterion-referenced tests is not to rank students in terms of their performance on tests, they are designed to describe the learning tasks which can be performed by students.

As the purpose of criterion-referenced tests is not to discriminate between high and low achievers, the discriminating index of an item as defined above is not very meaningful for preparing such test. The difficulty level of items in criterion-referenced tests should be determined by the specific learning outcome each item is designed to measure. So, if a learning task is of moderate level of difficulty, the test item designed to measure it, should possess a moderate level of difficulty. The standard formula for calculating difficulty level, can be applied to criterion-referenced test items, but the results are not used to select items or to manipulate item difficulty. Most items on a criterion-referenced mastery test, if the teaching is effective, should have a high difficulty index, i.e. high facility level indicating high percentage of students passing the test.

As regards the discrimination index in a CRT also, one would expect that more students in the upper group would solve the item correctly, than those in the lower group. In other words, if the discrimination is negative or low, i.e., if more students in the lower group are getting the item correct, the test constructor should look into the item as it indicates that the more intelligent students—the high achievers are getting punished. Thus here, the logic of revising an item is the same as in NRT, the only difference being that a constructor of a CRT will not make special efforts to get items of 50 per cent difficulty level or of high discrimination, because his purpose is not to rank students, or for that purpose, to discriminate between them. His purpose is only to see how far his students have achieved the stipulated instructional objectives or expected learning outcomes—for which the teaching/instruction was devised.

How do we analyse the criterion-referenced test item? A crucial question in judging the effectiveness of a criterion-referenced mastery test is, "to what extent did the test item measure the effectiveness of the instruction?" This question may be answered by testing the students before and after the instruction and comparing the results. For this purpose, the students need to be administered a pre- and post-test, which measures the same learning outcomes. A simple item by item chart for comparing the performance of students on the pre- and post-test can be prepared. The following is an example of such a chart.

Table 1
Item response chart showing correct (+) and incorrect (-) responses of students before and after instruction

Students	Items										No. of eqvt. items
	1		2		3		4		5		
	A	B	A	B	A	B	A	B	A	B	
1.	-	+	+	+	-	-	+	-	-	+	
2.	-	+	+	+	-	-	+	-	+	+	
3.	-	+	+	+	-	-	+	-	-	+	
4.	-	+	+	+	-	-	+	-	-	+	
5.	-	+	+	+	-	-	+	-	+	+	

Note: A represents Pre-test
B represents Post-test

It can be seen that Item 1 is an ideal item for a criterion-referenced test, as all students have answered it incorrectly before instruction and correctly after instruction. Item 2 was too easy and Item 3 was too difficult, as indicated by the students' performance on the pre-test and post-test. Item 4 is extremely defective as the students appear to have unlearned what they knew earlier or have got confused after the teaching. Item 5 can be considered as effective item showing a fairly typical response pattern, where some students have got the item right before instruction but a larger proportion did so, after instruction.

The following formula can be used to estimate the effectiveness of a criterion-referenced test. It measures the sensitivity of the item to instructional effects.

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

Where

- RB = number of students who have got the item right after instruction,
RA = number of students who have got the item right before instruction, and
T = total number of students who tried the item both times.

The sensitivity index will range from +1 to -1. The ideal item will naturally have the sensitivity index of +1. The higher the value, the more sensitive the item is, to the effects of instruction. The sensitivity index measures the degree to which an item reflects the intended effects of instruction occurring between the pre-test and post-test. Thus, criterion-referenced test item should be analysed by computing the index of sensitivity, rather than through the index of difficulty level and discriminative power of item.

We have discussed in this section how analyzing the test items, after question papers have been scored, help in further improvement of items and in diagnosing students' difficulties and/or to provide them the specific feedback. You will agree that on the basis of such analysis, items can be stored for future use. This is the beginning of the idea of a question bank.

Item analysis for short answer essay type items

In unit 15, you have studied about answer and essay type items. In case of Higher Education these items are used to a great extent. Therefore, we must know the procedures of analysing these items and selecting quality items. Usually we do not find procedures for finding out difficulty level and index of discrimination for essay type questions in the literature available on educational evaluation world-wide. The Centre of Advanced Study in Education (CASE), The M.S. University of Baroda, has developed the procedure for calculating the difficulty value and index of discrimination for short answer and essay type items. This is given hereinafter.

Facility index for short-answer and essay items

The difficulty value of an item indicates how easy or difficult an item is for students. In case of objective type items it is indicated by the percentage of students who answer the item correctly. In case of essay and short answer items difficulty value of an item can be indicated by the ratio of the average performance of both the groups (upper and lower) on an item and total marks of the item multiplied by 2. The following formula can be used to estimate the difficulty value of an item (Edwin Harper, 1990).

$$\text{Facility Index} = \frac{\text{MU} + \text{ML}}{2(\text{FM})}$$

Where MU = Mean achievement of students in the upper group (1/5th) students

ML = Mean achievement of students in the lower group (1/5th) students

FM = Full marks of the item.

The procedure of calculating difficulty value is given below. Let us take about 30 answer sheets for the item analysis. The following steps illustrate the procedure. First two steps are the same as in objective type question.

- Arrange all the 30 answer sheets in order from highest score to the lowest score.
- Select approximately one third of the answer sheets (10) from the top with the highest score and call this the upper group. Similarly, select the lower group of the same number from those who have the lowest scores and call this the lower group.
- For each item calculate the mean for upper group as well as for lower group.

This is shown for an item of 5 marks.

Sr. No.	Marks obtained	
	Upper group	Lower group
1	4	1
2	4.5	2
3	5	1
4	4	2
5	3	1.5
6	4	1
7	4.5	0
8	3	1
9	5	1.5
10	4	1
Total	41	12
Mean	$\frac{41}{10} = 4.1$	$\frac{12}{10} = 1.2$

Thus, for this item, we have required values as follows:

$$FM = 5, MU = 4.1 \text{ and } ML = 1.2$$

$$\text{Facility Index (FI)} = \frac{MU + ML}{2 (FM)}$$

$$= \frac{4.1 + 1.2}{2 (5)}$$

$$= \frac{5.3}{10}$$

$$\text{Facility Index} = .53$$

Following the same procedure, difficulty value can be calculated for all the items. It would range from 0 to 1. More the difficulty value, easier would be the item. The item having facility value around .50 would be of average difficulty. But what is a good question as far as facility index is concerned? According to Harper, in 4 choice items the range for FI is 60–75, and in 3 choice items the preferable range for FI is 65–80.

Index of discrimination for short-answer and essay items

The index of discrimination which was discussed earlier is an important parameter for selecting quality items in educational evaluation. For computing index of discrimination for short answer and essay type question, the following formula can be used.

$$\text{Index of Discrimination} = \frac{MU - ML}{FM}$$

Thus for the example taken for computing difficulty value, the index of discrimination would be as follows:

$$\text{Index of discrimination} = \frac{4.1 - 1.2}{5}$$

$$= \frac{2.9}{5}$$

$$=.58$$

Index of discrimination is .58

It may be noted that the index of discrimination is indicated by a decimal fraction. The discrimination value of an item may range from –1 to +1. You may include items in the test, with higher discrimination values. Thus, items with lower values and those with negative indices should be dropped or modified.

Self-assessment

6. Why is the 50 per cent level of difficulty recommended for effective norm-referenced test items? What level of difficulty is best for criterion-referenced test items?
7. An item was answered correctly by 9 out of 10 students in the upper group, and 5 out of 10 students in the lower group. What is the difficulty level of the item? Should the item be considered difficult or easy? What would be its index of discrimination?

Question banks

In the preceding section, we learnt about the parameters of test items, namely the

difficulty level and the index of discrimination. It was observed, however, such parameters obtained through the item analysis process, help a teacher to estimate the level of effectiveness and the efficiency of the item or question, to evaluate the variety of specific learning outcomes of students under different subject areas. Obviously, such effective items of evaluation are not developed for a one time use only, but are expected to be used again and again with the necessary refinement and modifications. A pool of such items, developed for a particular group of students, studying a specific subject area, will be helpful to the concerned teacher to frame a good question paper. The teacher will be able to select items or questions from the pool, to set the question papers for the test (to be administered to his students). This is basically the purpose of developing a question bank.

Examination reform and question banks

You have studied in status of evaluation at the higher education stage in India. Examination has been identified as the major stumbling block in qualitative improvement of instruction at various levels of education in India. The Education Commission (1964–66) has said: “examination reform has become crucial to all progress and has to go hand in hand with the improvements in teaching”. You have also studied the examination reforms initiated in the field of higher education from time to time in Unit 12. In that context, you came across the recommendations of the University Grants Commission about the examination reforms in universities and colleges in India. Mainly, these recommendations are related to four major aspects of examination reform, viz., internal evaluation, grading system, question banks and national examinations. The U.G.C. had published a monograph entitled “Examination Reform—A Plan of Action” in 1973, setting out the guidelines for the suggested reforms in examinations at the higher education stage. The Association of Indian Universities has sponsored a number of research projects concerning different aspects of examination reforms, including question banks. A.I.U. has published a monograph on question banking for universities in 1977, which has suggested a plan of action for developing question banks. You may go through the A.I.U. monograph and similar publication to learn more about the systematic procedures for developing question banks, as also the work done by A.I.U. and other agencies, in this connection.

Purpose of question banks

The main purpose of a question bank is to eliminate or reduce to a considerable extent, the defects in setting a question paper. At present, examiners and paper setters usually adopt an adhoc procedure. The selection of questions is not done systematically and it is not uncommon for paper setters to just repeat questions from the last few years' question papers. There is no attempt to work out a table of specifications and to select/write questions to represent all the cell in such a table. A systematically developed pool of questions or items of evaluation, namely a question bank, offers a chance to set up valid and reliable question paper, to compare the performance of a group of students with that of the wider population to match the examination to the prescribed curriculum, instead of the other way round.

Developing a pool of items

Question banks consisting of items of evaluation may be developed at different levels: an individual teacher preparing a pool of questions for teaching his subject, a group of teachers teaching the same subject in a college developing the bank for their use, or the preparation of a pool of items at the university level for use in examinations for all the students in different colleges. The A.I.U. has developed question banks for students at the undergraduate level in Indian universities in a number of subjects.

Developing such a pool of items can be undertaken in a number of ways. Questions asked in past examinations can be scrutinized and effective questions or items can be selected from there for inclusion in the question bank. Another approach, which has been successfully tried out by the A.I.U., is to organise workshops for teachers to develop effective items. Still another approach, to develop the bank, is to get the items written by certain experts. Writing good items or questions for question bank, whichever way it is done, involves the steps we have already discussed in the previous unit as well as in the preceding sections of this Unit. These steps require the course outline in a particular subject areas to be analysed in terms of the content units and specific instructional objectives, preparing a table of specifications or blueprint for the question paper, writing appropriate items or questions based on the content units and specific learning outcomes, testing the items for characteristics of a good evaluation tool, namely, validity, reliability, objectivity, etc., and computing the parameters of the evaluation items such as difficulty level and discriminative index after a tryout on a group of students studying the course. Acceptable items retained after screening, according to the procedure stated above, should be stored in the question bank for use on a card with details regarding the specific learning outcome and content it measures, as well as its characteristics and psychometric properties. The items may be stored in other forms also, such as brochures or monographs.

Storing such information in a computer system will facilitate its use to a much greater extent. Universities and other examining bodies, in fact, need well-developed computer systems to store and retrieve the large item pools in various subjects. Developing the question bank is a continuous process. Items in the question bank should be appraised from time to time, new items to be continuously added to the existing pool and outdated items to be discarded. A bank, to be effective, should consist of a large number of items of known technical value. Obviously, the bank should be composed of all types of items, such as the essay type, short answer and the objective type, in different formats. The bank may not be limited to question for use only in written test, but, may also include questions for use in oral practical examination.

Using a question bank

A teacher should use a question bank for preparing a question paper to test the learning outcomes of students with reference to the content units he has taught. This requires the use of a two-way table of specification or blueprint as we have discussed in the previous Unit. The questions or items, which constitute the question paper, can be chosen from the bank so that, specific learning outcomes under different content unit are measures. While choosing the appropriate items, the teacher will have to consider the specific instructional objectives to be evaluated by the items, the difficulty value and the index of discrimination of the items, etc. He/she may also have to consider the purpose of the test. For example, if the test is a norm-referenced test, the teacher would like to select those items which possess the highest discriminative indices and which have difficulty level around 50 per cent. On the other hand, if his purpose is to judge the mastery level of the student's performance through a criterion-referenced test, he would look for items which have much higher levels of difficulty.

Similar procedures can be used to develop university level examinations, where a group of teachers or paper setters go through the steps described above, to set a question paper.

If items are stored properly in a computer, question papers related to specific content units and objectives may be provided by the computer. It may also be possible to develop two, three or more alternate forms of the question paper through a computer.

Another use of the question bank is for devising diagnostic tests to identify the learning difficulties of students related to specific contents and objectives. A suitable set of

questions, related to those areas of content where students' performance is low, can be used to diagnose their specific learning disabilities by developing diagnostic tests on those specific areas and providing needed corrections.

Thus, framing question papers for a variety of tests is possible, provided a systematically developed pool of items, that is a question bank, is available.

Systematising instruction through question banks

You may like to think about the consequences of a question bank on teaching-learning activities. If students have access to the question bank items, there is a likelihood that they will focus their attention on the questions in the bank, instead of trying to achieve the specific instructional objectives. Similarly, is teaching going to become examination centred, if teachers organise their instruction on the basis of a question bank? There is, however, a rare possibility of the above taking place. When questions are made specific and expected answers are standardised, students will develop more faith in examinations. As the questions are based on specific instructional objectives in particular subjects, students will prepare to achieve the specific learning outcomes, which are intended to be measured through these questions. Thus, students' learning efforts will be more focussed on the learning tasks and they will develop greater faith in the examination system. Teaching will also become more pointed and its effectiveness will improve considerably. As the teachers will be required to focus their teaching towards the items in the bank, their accountability to students will also increase. Thus, the need for institutional question banks for use by students and by teachers for formative tests cannot be over emphasized.

It may also be emphasised here that a question bank should have a very large pool of questions—some hundreds of questions in a particular subject at a particular level so that only a student who has thoroughly mastered the subject at a particular level can answer all these questions. The availability of such a large pool from which a sample—a particular question paper is drawn—will reduce the fear of the bank being known to some students, as it happens in the usual leakage of question papers in the present day examinations.

Development of confidential question banks with psychometric data in different subjects by A.I.U. and by different universities is the pressing need to improve quality of question papers. Developing big pools of questions in question banks is an ongoing process which an examination agency requires in the form of confidential bank. Besides improving the validity and reliability of question paper, it does provide basis (indirectly) for teachers and paper setters to understand the expected learning outcomes especially the higher order instructional objectives requiring more emphasis.

Thus, it can be observed that systematising evaluation through reforms such as question banks will result in improvements in the instruction. Evaluation and instruction, thus, are closely interlinked with each other and reforms in one should go hand in hand with the other.

Self-assessment

8. *Questions in a question bank should be in different formats such as essay type item, short answer and objective type item. List three reasons in support of this statement.*

Summary

We have discussed in brief the essential characteristics of an effective evaluation tool. We mentioned about four essential characteristics of an evaluation tool. These are

validity, reliability, objectivity and usability. There are four types of validity, which have been identified in the unit, namely content validity, predictive validity, concurrent validity and construct validity. We have also touched upon the three approaches of checking the reliability, i.e. test-retest, equivalent-forms method and split-half.

Objectivity is also known as inter-scorer reliability. Under usability, we stressed on the ease of administration, ease of scoring, time required and the cost involved. For judging the worth of the items, we focussed mainly on two parameters: the difficulty level and the index of discrimination for both types of items—objective and essay-type. We also did some simple computations of these parameters. Towards the end, we discussed the question banks which are the systematically developed pools of evaluation items of known technical value, for framing effective question for a variety of tests for use in formative and summative testing.

Unit-end activities

1. Construct a composite test on any topic in your subject including at least ten objective type items, testing different abilities/competencies.
2. Give this test to a student group of at least thirty students. Compute the item statistics—difficulty and discrimination indices as if the test is a criterion-referenced test. Evaluate test in terms of its capacity to give feedback to students and to you about the students' progress and/or their difficulties.

Points for discussions

Discuss with your colleagues in the college a plan for setting up a question bank in your own subject. What will be the steps involved in implementing such a plan?

Suggested readings

- Association of Indian Universities (1977) *Monograph on Question Banking for Universities*. New Delhi.
- Anastasi, Anne (1976) *Psychological Testing*. (4th ed.), New York: Macmillan Publishing Co, Inc.
- Cronback, L.J. (1970) *Essentials of Psychological Testing*. (3rd ed.), New York: Harper & Row.
- Freeman, Richard and Lewis, Roger (1998) *Planning and Implementing Assessment*. London: Kogan Page.
- Gronlund, Norman E. (1990) *Measurement and Evaluation in Teaching*. (6th ed.), New York: Macmillan.
- Thorndike, R.L. and Hagen, E.P. (1977) *Measurement and Evaluation in Psychology and Education*. (4th ed.), New York: John Wiley & Sons.
- Singh, Pritam (1996) *Towards Better Questions*. New Delhi: Navraj Publication.

Answers to self-assessment

1. Predictive validity.
2. A valid test will have a certain degree of reliability, but a reliable test may not have validity.
3. a.

4. Because it refers to the consistency in scores assigned by different scorers.
5.
 - a) Time between test and retest too large
 - b) Sample of items too small
 - c) The scoring procedure not objective
 - d) Vague items
6. Because at this difficulty level, maximum discrimination between the performance of students is possible, which is the purpose of an item in a norm-referenced test. The best level of difficulty for criterion-referenced test items is 100 per cent, if the teaching is effective.
7. Difficulty level of the item

$$\text{Difficulty level} = \frac{9+5}{20} \times 100 = \frac{14}{20} \times 100 = 70\%$$

The item should be considered as easy.

$$\text{Index of discrimination} = \frac{9-5}{10} = \frac{4}{10} = 0.4$$

This is a satisfactory discrimination

8. Because different types of items are required for measuring the variety of specific learning outcomes under different content units in a subject area.

Introduction

In Units 15 and 16, you have learnt about writing different types of test items, development of question papers and characteristics of a good question paper. When question papers are administered to students and their responses are evaluated by examiners, we get evaluation data generally in terms of marks or grades. These marks and grades are used for purposes of certification, selection, promotion, placement and improvement of teaching and learning processes by different institutions and universities. In the present unit, you will study about processing and interpretation of evaluation data and relative merits of grading and marking systems. You will also learn how test norms such as central tendency, variance and standard scores can be computed from grouped and ungrouped data.

Learning Outcomes

After going through this Unit, you will be able to:

- describe different methods of grading;
- interpret grades and marks;
- compare the merits and limitations of marking and grading systems;
- discuss the importance of test norms;
- compute test norms such as central tendency, variance and standard scores from grouped and ungrouped data.

Marking and grading systems

Marks and grades are the most tangible and perhaps the important components of an evaluation system. Marks/grades determine which course a student can take up for higher studies; how he/she is progressing and what is his/her achievement at the end of the course. Traditionally our universities, boards of examinations and other educational institutes certify the achievement of students in terms of marks. Teachers, students and parents are all more familiar with marks and consider marks as indicator of absolute achievement of a student. Comparison among students are made on the basis of a single score or sometimes on fraction of a mark. This kind of comparisons raises questions about the reliability of marking system. A number of studies have been conducted in India and in other countries on the subjectivity of marking system which contribute to the unreliability of marks. The research studies indicate that marks

awarded to a student are not the true marks. These contain errors of measurement. If the standard error of measurement for a test is 5, then in 68 cases out of 100, the true score of a student will be between ± 5 of the obtained score. (The standard error of measurement discussed in the preceding section). That means we are 68 percent sure that the true score of a student with obtained score of 60 lies between 55 and 65.

Therefore, it was considered that it is more appropriate to certify the achievement of students in terms of range or band of marks. These ranges or bands can be expressed by a letter grade such as A, B, C and so on. More systematic methods of grading have been evolved which can be used for various purposes. It is important for you to understand the mechanics of marking and grading, so that you can draw your own conclusions and make your own judgment.

What do marks and grades reflect?

You must have appeared in a number of examinations—written, oral or practical—for purposes of certification, selection or promotion. Your achievement/performance in these examinations is generally indicated in terms of marks, grade or rank position. What do these marks and grades indicate? Marks generally indicate the estimates of *absolute* achievement of a student on a test or an examination comprising of tests whereas grades reflect the *relative* achievement of students in the group. Marks are assigned on a continuous *ratio scale* while grades are assigned on *interval/ordinal scale*. You may recall that in a ratio scale absolute zero exists and all operations such as +, −, $\times$, $\div$ are possible. In the interval scale absolute zero does not exist but the interval between two consecutive classes is the same. Since marks are on a more precise ratio scale, you tend to compare these in absolute terms such as the achievement of a student securing 60 marks is better than the one securing 59 marks and is poorer than the one securing 61 marks. However, this interpretation may not be as precise and accurate because marks contain errors of measurement. The sources of these errors of measurement can be subjectivity of the examiner or question paper itself. Marks must, therefore, be interpreted with reference to the test norms or a priori defined criterion. Grades can also be interpreted in a number of ways depending on the method of grading. Grading can also be done in norm referenced or criterion referenced fashion.

Many a times decisions are taken on the basis of aggregate marks or grade point average. You might have experienced that sometimes your aggregate of all the subjects becomes low because of low score in one of the papers only. For example your aggregate becomes less because of low score in say, Mathematics. This, however, does not reflect your overall poor performance or less ability. Marks assigned on each paper indicate a student's achievement in that particular subject. Marks on this paper also provide basis for comparison of a student's achievement in that particular subject. Marks of a student in a particular subject should not be taken as indicator of his/her personality, attitude or conduct. Also an average score of say, English, Mathematics, Music, Science, Social Science, Arts cannot be taken as a single measure of all hodgepodge of factors. A student's marks in English is a measure of different abilities, may be reading, comprehension, grammar and so on, whereas his/her marks in Mathematics may indicate his/her numerical ability or ability to solve problems of space, money, etc. A Mathematics teacher may use the full range of marks from 0 to 100 whereas an English teacher may use only a limited range of 30 to 70. Therefore, comparison of students on an average score may not give the true picture.

Comparison on the basis of grades has another dimension, that is, whether grading has been done on a norm-referenced scale or a criterion-referenced scale. Comparison of grades obtained by two different methods is fallacious. Also, grades obtained on norm-referenced scale should not be interpreted in absolute terms. Grade 'B' of one university may be as good as grade 'A' of another university in terms of absolute achievement.

Similarly, grade 'C' of one year may be as good as grade 'B' of previous year in terms of absolute achievement. Norm-referenced grades must, therefore, be interpreted in relative terms. Grades for co-scholastic abilities such as attitude, emotional and social competence, conduct, regularity and punctuality, etc., should be shown separately.

The Marking system

Marks are generally assigned on the basis of number of correct responses or the extent to which a student's answer matches with the expected answer. Marks are assigned on each item/question and then mathematically added to arrive at the total marks in a paper. The maximum marks in a paper can be 50, 100, 200 and so on. For purposes of comparison and interpretation marks are expressed in terms of percentage and there we use 101 point scale, from 0 to 100. Marks are generally interpreted in terms of absolute achievement but these are only estimates of absolute achievement. Also, the marks obtained on a paper are not the true marks, these contain the errors of measurement. The errors of measurement can be expressed as standard error of measurement.

$$\text{Obtained score} = \text{True score} + \text{Error score}$$

The sources of errors of measurement can be many such as the test itself, examiners, examinees, test administration, scoring of test, etc.

(a) Error within the test

You might have heard students discussing that a particular paper was too difficult/easy, all the questions were from a few units of the syllabus; questions were ambiguous; scope of answers was not indicated; items were poorly constructed; question paper was lengthy; readability of question paper was poor, etc. If the question paper is poorly designed then the obtained score can be lower or higher than the true score. For example, if a question paper contains too many clues, then the obtained score would be higher than the true score and if the questions are ambiguous or the reading level of questions is high then the obtained score may be lower than the true score. There can be inter test as well as intra test errors.

(b) Errors due to examiner's subjectivity

Errors in obtained scores can be due to examiner's subjectivity. If you ask five teachers to score the same set of answer sheets, you may find that five examiners have different standards of scoring. Not only that their range, means and standard deviations are different, their score on the same script varies greatly. Research studies indicate that subjectivity in essay type of question papers is so high that sometimes a student failed by one examiner is assigned first division marks by another examiner. Examiner subjectivity has been found in all the subjects. Also, if you ask the same examiner to score the same answer script on two different occasions, his/her own scoring on the same answer script may be different on two occasions.

(c) Errors within the examinees

You might have experienced that sometimes you do not fare well in the examinations due to illness, emotional problem, fatigue or sleeplessness. You get relatively low score in the examination, which is lower than your true score. A student may accidentally see other students' answer and may get higher score than his true score. All these situations are temporary but these contribute to intra individual errors. Therefore, the factors within the examinees also contribute to errors of measurement.

(d) Errors in test administrations

Sometimes you might have experienced a power cut during the examination. On another occasion there was too much noise outside the examination hall or the supervisors were talking to each other and you could not as well in the examination as you could have done otherwise. You might have experienced that different supervisors or test administrators provide differing amounts of information to test takers. Some keep you reminding about the time, or advise you not to guess, whereas others remain fairly distant. The test administrator can be emotionally supportive to the students or can also be harsh in his behaviour. These test administration factors also contribute to error of measurement.

(e) Errors in scoring

There can also be errors in scoring. In the manual scoring of tests the likelihood of errors is more. Though computerized test scoring may have less errors, these also are not completely free from errors.

The errors in marks due to different sources also affect the reliability of examination results. Measures can be taken to minimize the errors of measurement but we can never eliminate these errors completely.

Activity 1

Suppose you have 10 students in your tutorial group. Compare the marks of each of these 10 students on four assignments/tests that you gave them during this session. What is the range of marks of each student? Try to provide adequate explanation for variation in each student's marks. Also ask students to explain why they got less marks on one occasion and more marks on another occasion

Activity 2

Here is a question picked up from an actual question paper. Administer this question to 5 students of your class fellows. Ask another 5 (examiners) persons to score the 5 scripts. Find out the cause of variation in nature of responses and the variation in award of marks by the five examiners.

Q.: What is parasitism?

(4 marks)

Inadequacies of marking system

The marking system has inherent subjectivity. Firstly, the obtained marks are not true marks and these contain errors of measurement. As discussed in the previous section, the errors of measurement can be due to examiner subjectivity, within test, within examinee, test administration or test scoring. Even the best constructed tests have standard error of measurement of about 3 to 5. Therefore, all test scores are fallible. Secondly, marks are only the estimates of achievement. These do not reflect the absolute achievement of a student. For this estimated value also, different examiners use different scales of measurement. Even, when the maximum marks are 100, one examiner may use range of 20-80, the second examiner may use range of 40-70, the third examiner may use the entire scale of 9-100. Practically, all examiners do not use the 101 point scale. And you may find that the mean and variance for these examiners is different. If you draw the distribution of marks assigned to students in different papers, you will find that some distributions are positively skewed whereas others are negatively skewed. Still, we consider that 60 marks in each paper means the same level of achievement.

Treating marks as absolute achievement of students creates problems. Although marks are on ratio scale and these can be added, subtracted, multiplied and divided but the abilities, achievement of which these marks indicate may not be additive. Therefore, another problem comes when marks of all the subjects are added and total marks are taken as measure of overall ability. Sometimes these marks are also considered as measure of students' conduct, attitude and other personality factors. In fact marks indicate a student's level of achievement in that particular subject. In order to understand the accuracy and errors of marking system, it is important for you to understand the concept of standard error of measurement.

The standard error of measurement

Error scores are assumed to be random. As such, they cancel each other out, that is, the mean of error scores for a test is zero. The distribution of error scores approximate a normal distribution. The standard deviation of this distribution of error scores is called standard error of measurement (S_m) as illustrated in Fig. 1.

Mathematically,

$$S_m = SD \sqrt{1-r}$$

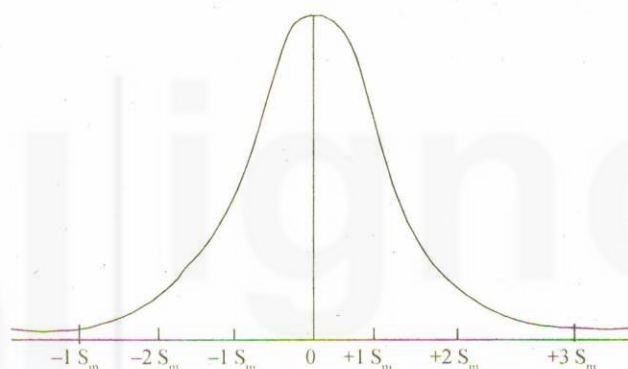


Fig. 1 : The Error Score Distribution

where SD is the standard deviation and r is the reliability of test. To understand this concept, let us work out an example. Suppose a test has standard deviation of 10 points and reliability of .75, then the standard error of measurement will be

$$S_m = 10 \sqrt{1-0.75} = 10 \sqrt{0.25} = 5.0$$

If a student obtains a score of 70 (X) on a test, his true score would lie between

65 and 75 (68% of the time)	$X \pm 1S_m$
60 and 80 (95% of the time)	$X \pm 2S_m$
55 and 85 (99% of the time)	$X \pm 3S_m$

If the test is perfectly reliable ($r=1.00$), then a student's true score will be equal to obtained score and S_m will be zero. But no test is perfectly reliable and the best made tests have errors of measurement.

Activity 3

Suppose your test has standard deviation of 15 and reliability of 0.60, if a student has obtained 50 marks, in what range his true scores would lie? How would you compare the obtained scores of 60 and 50?

The Grading system

In order to overcome some of the inadequacies of marking system (discussed in earlier

section). University Grants Commission (U.G.C.), National Curriculum Framework for School Education (NCFSE, 2000) developed by NCERT and National Policy on Education (NPE, 1986) recommended use of grades in place of marks to certify the achievement of students. Most of the Universities and Boards of Examinations have accepted to use the grading system for certification purpose but still there is no unanimity as to which method of grading should be used. This debate is reflected in newspapers and other media, where you must have been participating. In order to draw your own conclusions and take your own decisions, you must understand different methods of grading.

(a) Methods of grading

Many a times, teachers award/assign grades on your assignments/class work/tutorial work. How do teachers assign these grades? Sometimes for a small sample of 20–25 students, teachers use a simplified, rudimentary method where they classify the assignments in five different categories—very good (A), good (B), average (C), below average (D) and poor (E). They make this classification according to their own standards of performance without specifying the number of scripts to be placed in each category. However, for large samples and for purposes of decision making as in the University/Board examination, standard statistical procedures should be used. The statistical theory of probability says that in a large population, any trait, ability, or characteristic is normally distributed. A normal distribution is a bell shaped curve in which 68% of the cases lie between $M \pm 1$ S.D. (where M is the mean and S.D. is the standard deviation of the distribution) 95% of the cases lie between $M \pm 2$ S.D. Grades can also be awarded on the basis of normal distribution, presuming that in a large sample achievement of students in different subjects is normally distributed. This type of grading is called 'grading on the curve' or 'norm-referenced' grading.

(i) Norm-referenced grading

Norm-referenced grading is based on relative achievement, that is how your achievement compare with the achievement of other students in the group. In this method certain proportion of students are assigned A's and B's regardless of their absolute achievement on a test. For example, a student with 60% correct items can be assigned grade 'A', if a large proportion of students have answered less than 60% items correctly and a student with 60 percent correct items can be assigned grade 'B', if more students have answered more than 60 percent items correctly. Similarly, student with 50 per cent correct item might get a B or even D grade depending on how his achievement of 50 per cent compares with other students achievement. If this score of 50 per cent is higher than 95 per cent of the students, then the student is likely to get an 'A' grade. Sometimes universities or board of examinations specify the percentage of students who will be assigned various grades. The grade can be assigned in a five, seven or nine point grading system. The proportion of students in each grade according to normal distribution (Fig. 2) can be as follows:

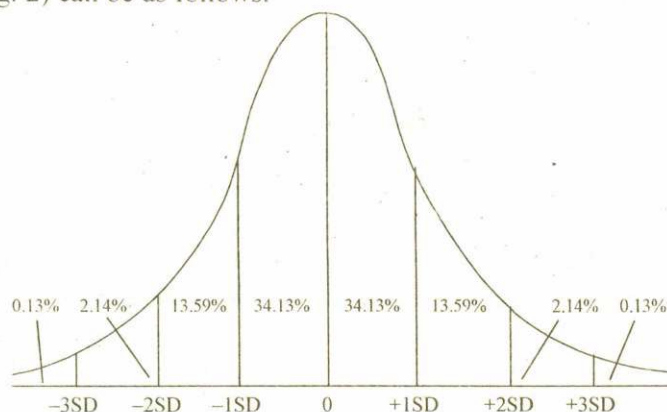


Fig. 2 : Percent of Cases Under Portions of the Normal Curve

Table 1: Norm-referenced grading on a five point scale

<i>Letter Grade</i>	<i>Interval</i>	<i>No. of Cases</i>	<i>Grade Value</i>
A	1.5 to ∞	7%	5
B	0.5 to 1.5	24%	4
C	-0.5 to 0.5	38%	3
D	-1.5 to -0.5	24%	2
E	$-\infty$ to -1.5	7%	1

Source: Grading in Schools, NCERT 2000.

If the University feels that they have less students towards lower achievement tail and comparatively more students towards higher achievement tail, that distribution is slightly positively skewed, then distribution can also be like this.

<i>Grade</i>	<i>Percentage of students</i>
A	10
B	25
C	40
D	20
E	05

The distribution of students over nine points scale can be follows

Table 2: Norm-referenced grading on a nine point scale

<i>Letter Grade</i>	<i>Interval</i>	<i>No. of Cases</i>	<i>Grade Value</i>
A	1.75 to ∞	4%	9
B	1.25 to 1.75	7%	8
C	0.75 to 1.25	12%	7
D	0.25 to 0.75	17%	6
E	-0.25 to 0.25	20%	5
F	-0.75 to -0.25	17%	4
G	-1.25 to -0.75	12%	3
H	-1.75 to -1.25	7%	2
I	$-\infty$ to -1.75	4%	1

Source: Grading in Schools, NCERT, 2000.

(ii) Criterion referenced grading

In this method of grading, it is possible that all students get grade 'A' or grade 'E' or any other grade in between. The number of students getting each of these grades varies from year to year, university to university and subject to subject. For example, in Mathematics a large proportion of students may get 'A' and 'E' whereas in History no student may get 'A' or 'E'. Similarly one year, only 5 percent may get 'A' in Mathematics and next year 30 percent may get 'A'. The grade is assigned according to a student's level of achievement or performance, regardless of other students' achievement or performance. The assignment of grades can be based on range of marks as shown below:

Table 3: Criterion-grading on a five point scale

<i>Letter Grade</i>	<i>Range of marks (%)</i>	<i>Grade Value</i>
A	75 – 100	5
B	60 – 74	4
C	45 – 59	3
D	33 – 44	2
E	0 – 32	1

Table 4: Criterion-referenced grading on a nine point scale

<i>Letter Grade</i>	<i>Range of marks (%)</i>	<i>Grade Value</i>
A	90 – 100	9
B	80 – 89	8
C	70 – 79	7
D	60 – 69	6
E	50 – 59	5
F	40 – 49	4
G	30 – 39	3
H	20 – 29	2
I	0 – 19	1

For using this system of grading, the examiners and teachers should have some prior knowledge of what level of achievement or performance is reasonable to expect. Teachers should also have knowledge about the difficulty level of test and each of the items in the test. If you compare the distribution of students in five point and nine point scales as shown in Tables 1 and 2, you will find that 7% students are placed in grade 'A' in five point scale whereas in nine point scale, only 4% students get grade 'A'. The percentage of students getting grade 'B' is 24 in five point scale and only 7 in nine point scale. That means nine point grading makes discrimination among 24% placed in grade 'B' in five point scale.

Similarly, if you compare the range of marks for each grade on a five point and nine point scale, as shown in Tables 3 and 4, the 'B' grade in a five point scale has a range of 60–74 and 'A' grade has a range of 75–100, whereas on nine point scale, 'A' grade is given to only those who get 90% or more marks and 'B' grade is given to those getting 80–90% marks. That means nine point scale discriminates between students getting 90%, 80% or 75% marks, whereas five point scale does not make such finer discriminations. The choice of 5 point, 7 point or 9 point scale depends on the purpose of evaluation. If this is a class evaluation, classifying students in five categories may serve your purpose. If these grades are required for placement or selection, finer discrimination on a point scale may be required. For certification, UGC considers 7 point scale appropriate.

Merits and limitations of grading system

Though there are advantages in using grading systems, each method of grading has its own merits and limitations. Which method of grading should be used? It depends on the purpose for which grading is required. For improvement of teaching

and learning process, criterion referenced grading is more useful. Now, you understand both the methods of grading, try to list the merits and limitations of each of these methods.

(a) Merits and limitations of norm referenced grading

The advantage of a norm-referenced grading is that it simplifies decision-making processes. It tells you where a student stands in the group, whether he/she is among top ten or twenty percent. Students also do not go for unhealthy competition for getting particular grade. They do not agonize over what cut-off scores are.

There are several disadvantages of this methods.

- (i) In this system it is previously decided that some students will get the 'A' grade and some students will get the 'E' grade or the fail grade. If such a distribution is used for a group of bright or gifted students, where all achieve good marks, suddenly some of these bright or high achievers are assigned C, D, or E grade. Students might feel that this is injustice to them. Similarly, if such a distribution is imposed on a class of low achievers some students get 'A' regardless of their level of achievement. In this system, comparison of students from two universities/ colleges and students of different years is sometimes difficult.
- (ii) Sometimes the distribution of students in different grades also poses problem, that is why 5 percent 'A's or 10 percent 'A's?
- (iii) Interpretation of grades by students and teachers is difficult. For example, a student with 'A' grade may be presumed to have mastered the content. But if the performance of group as a whole was poor, then even 'A' grade student may not have achieved mastery of the content.

(b) Merits and limitations of criterion-referenced grading

This system of grading is particularly useful for improvement of teaching and learning and provides feedback to both teachers and students. The second advantage of this method is that it is possible for all students to get high grades if they put forth sufficient efforts. These grades indicate the achievement. This type of grading tells whether a student has answered 60–70 percent items correctly or 30–40 percent items correctly. The students can know about their level of performance. Comparison of grades assigned by different universities in different years can be made if question papers of these universities are comparable in terms of validity, reliability and difficulty level.

Criterion referenced grading also has some limitations:

- (i) Establishing a standard is not easy, what is reasonably expected for an 'A' in one university or one subject may vary from another university/subject and from time to time.
- (ii) Decision for selection or placement becomes difficult if many students have 'A' or if all the students get A or B.

University Grants Commission (UGC) recommends use of grading system as one of the reform system.

University Grants Commission recommendations on the grading system

We have discussed the examination reforms suggested by the UGC through its monograph entitled 'Examination reform – A plan of action', in the Unit 11 of Block 3. One of the reforms in examination suggested by the UGC is related to the adoption

of the grading system in Universities and colleges. The following section lists the important recommendations of the UGC regarding the grading system.

(a) Regarding Scaling

- (i) A 7-point grading scale is recommended. It would keep evaluation within reasonable limits of accuracy.
- (ii) There should be no objection if, within a university, some faculties like engineering, medicine, etc., use a scale other than the 7-point scale for grading.
- (iii) The **rationale** for the 7-point scale is that it conforms, in **qualitative terms**, to the distinctions normally made by teachers in classifying students' performance.
- (iv) The distinction normally made by teachers are: Outstanding, Very Good, Good, Average/Fair, Below Average/satisfactory, Poor and Very Poor. These may be indicated by the following letter grades O, A, B, C, D, E and F respectively.

(b) Regarding mechanics of grading

- (i) There was broad consensus in favour of **direct grading** by teachers.
- (ii) Each answer in the answer script should be separately evaluated. The overall grade for the answer script can be arrived at on the basis of points assigned to each grade. Points suggested are: 'O' – 6, 'A' – 5, 'B' – 4, 'C' – 3, 'D' – 2, 'E' – 1 and 'F' – 0. The grade points for different questions can be added and the total divided by the number of questions to arrive at the grade point average. The overall grade can be arrived at on the basis of the grade point average. The following table illustrates this procedure.

Table 1: Determining overall grade

Question No.	1	2	3	4	5	Total
Grades	O	O	A	B	O	B
Grades points	6	5	4	3	6	26

$$\text{Grade point average} = \frac{26}{5} = 5.2$$

Overall grade A.

The fractional value .5 and above can be rounded up to the next higher grade e.g., 5.6 may be assigned O grade, whereas fractional value below .5 may be rounded up to lower grade e.g., 5.2 may be assigned A grade.

If all the questions are not of equal value, then points will have to be assigned to each grade, according to the value of the question.

The same procedure may be adopted in combining grade points in different papers in a subject.

- (iii) Marks given by teachers can be converted into grades with the help of **distribution tables**, as explained below. Such a conversion from marks to grades may be necessary until the universities have switched over to the grading system. For converting the marks to grades, the following procedure is suggested:

Take the top marks usually awarded in the subject and university under consideration and award it an O. Award E and F to the fail mark and distribute the remaining marks on the basis of equal intervals. For example, suppose the pass-mark in History is 35 and the highest mark awarded during the last three years is 74, the distribution of grades will be as under:

67 and above	O (Outstanding)
59 – 66	A
51 – 58	B
43 – 50	C
35 – 42	D
25 – 34	E (Failed)
24 and below	F (Failed)

Now suppose the highest mark awarded in Mathematics during the last three years is 99 and the fail/pass cut off score is 35, that is, minimum 35 marks are required to pass in the paper, the distribution will be as under:

87 and above	O (Outstanding)
74 – 86	A
61 – 73	B
48 – 60	C
35 – 47	D
25 – 34	E (Failed)
24 and below	E (Failed)

Similar tables may be constructed for each subject or paper, and this will provide a more rational basis for combining grades in different subjects.

- (iv) The grades in internal assessment and external examinations should be shown separately.
- (v) Along with the grade of the student in each subject, the **grade point average** can also be indicated. This will provide a large number of discriminations and enable us to distinguish amongst students belonging to a certain category.
- (vi) The minimum grade point average in each subject or the course as a whole will have to be indicated by each university according to its needs. It is recommended that the minimum grade point average for a course in general, should not be less than 2.00. A higher average may be prescribed for post-graduate and professional examinations.

Only the main recommendations of the UGC regarding the adoption of the grading system are stated above. For a more detailed description of these recommendations, you may refer to the UGC monograph, “Examination reform – A plan of action” 1973.

Self-assessment

1.
 - (i) Explain the term ‘error’ score.
 - (ii) List inadequacies of the marking system.
 - (iii) What are the sources of error scores?
 - (iv) What is the difference between norm-referenced and criterion-referenced grading?
 - (v) What are the limitations of norm-referenced grading?

Upgraded data

Scores on a test are raw scores, which have to be processed before they can be used by teachers to understand the performance of students. Therefore, such unprocessed scores are called raw scores. A jumbled mass of raw score is confusing and difficult for the mind to grasp. It has to be organized into the most meaningful form by reducing and summarizing. This is done by converting the ungrouped set of raw scores into a frequency distribution table or grouping them into certain meaningful categories. Suppose the performance of a group of students on a test is indicated by the scores given below:

Table 2: Scores of 33 students on a test

15	13	14	13	15	17	16
22	20	24	24	19	24	21
8	22	14	8	23	12	15
23	18	24	15	6	25	
9	25	23	25	7	14	

It may be observed that the scores of the above group of students range from 6 to 25, which are the lowest and highest scores respectively. These scores can be arranged to form several groups, each group representing the scores of a particular range such as 8 to 10, 11 to 13, etc. The following frequency distribution table is constructed from the 33 scores given in Table 2 above.

Table 3: Frequency distribution of scores of 33 students

Class Interval (C.I.)	Tallies	Frequency (f)
23–25		10
20–22		4
17–19		3
14–16		8
11–13		3
8–10		3
5–7		2
	Total	33

In the second column above, each tally is written by a mark 'I' and 5 tallies are, generally, noted by crossing the four tallies by another line, such as '||||'. Particular range of scores, such as 8–10, 11–13, etc., are called class intervals (C.I.), and the number of scores within a class interval is known as its frequency and denoted by the symbol 'f'.

Each class interval represents a group of scores. In Table 2, each class interval represents three scores. The lowest score included in the interval is taken as its lower limit and the highest score in it is known as the upper limit. In the class interval 14–16, the lower and upper limits are denoted by the scores 14 and 16 respectively. But, a score is generally considered as the midpoint of a distance extending from half a unit below to half a unit above this number. Thus, score 14 is actually the midpoint of the distance 13.5 to 14.5. Therefore, the actual limits of the class interval 14–16 will be 13.5 ($14 - 0.5$) to 16.5 ($16 + 0.5$).

One can easily understand the nature of performance of the 33 students from the above table. For example, it is observed that about one-third of the students have got scores in the highest class interval 23–25, suggesting the high level of their performance. Again, very few of the students, only two, manifested the lowest level of performance. Similar observations can be made regarding the levels of performance of the other students, which is suggested by the frequencies of scores in different class intervals.

The frequency distribution can be represented graphically. The class intervals can be represented on the X-axis (horizontal axis) and the frequencies can be shown on the Y-axis (vertical axis). The two axes can be divided into different units of equal width and each unit on the X and Y axes be taken to represent class intervals and frequencies respectively. The frequency distribution in Table 3 can be graphically represented as follows:

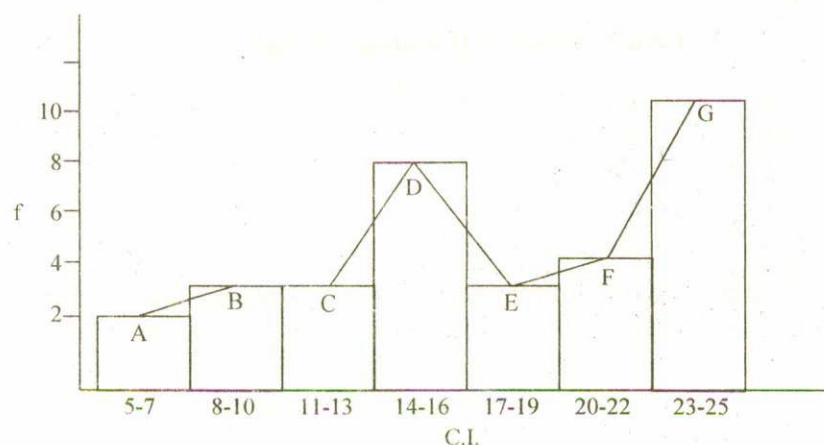


Fig. 3: Graphical representation of frequency distribution

In the above figure, frequency of scores in different class intervals is represented by heights of rectangles. Each rectangle has for its width the size of the class interval. Such a graphical representation is called a histogram. In Figure 1 above, there are lines joining the points A, B, C, D, etc. These points are actually the mid-points of different class intervals, and each point is at a height which represents the frequency of the particular class interval. Thus, lines joining the points A, B, C, D, etc., from a polygon, which represents the frequency distribution. This is known as a frequency polygon. Both, the histogram and the frequency polygon, help the teacher to make comparisons on the performance of different groups of students in a class. He can also know the position of each student in relation to others in the class.

Self-assessment

- Given below are the scores of forty eight students on a college test. Prepare a frequency distribution table. Using the same table, also prepare a histogram and a frequency polygon. Let the size of the class interval be 5.

Scores of forty eight students on a college test

57	49	37	42	63	44	47	69	65	33	24	30	40
55	58	52	36	41	45	27	35	48	36	44	60	
17	29	45	48	53	31	59	50	43	51	35	22	
32	49	39	54	42	39	41	38	54	46	50		

Test norms

As you are aware that the tests are used to measure the learning achievement of students. The achievement of students is indicated in terms of marks or grades. These marks or grades are used for purposes of certification, selection and placement. The interpretation of marks and grades is important because these are used for decision making. Test norms are important for interpretation of marks and grades. These tell us what is the average score on a test and what is the position of a student with reference to this mean. These also tell how the scores of a group/class of students are spread around the mean and what is the shape of distribution. Comparison of scores also becomes more meaningful if these are compared with reference to a norm. For example, comparison of scores from two distributions with common mean and variance is not as meaningful as scores from distributions with common mean and variance. Let us understand what these test norms are and how these are useful in interpreting the marks or grades.

Measures of central tendency

You must have read in your earlier classes about mean, median, and mode. These are three measures of central tendency. Each of these measures is used in different situations.

(a) The Mean

Many a times, you try to justify that your achievement is good because it is above average. How do you find out average score of a group? Generally, by adding scores of all the students in a class and dividing by the total number of students. This average score is nothing but mean. It is the most commonly used measure of central tendency because it is more stable. In the computation of mean, all the scores are considered and therefore, it is stable overtime as compared to other measures of central tendency. Another characteristic of mean is that sum of difference of each score from the mean is equal to zero. Mean is affected by extreme scores, that is, a few very high scores in a distribution can pull up the mean value whereas a few very low scores in the distribution can pull down the mean value. Mean, therefore, is not very useful in skewed distributions, where few have very high or very low scores. In such cases median may be used as a measure of central tendency.

(b) The Median

Median is the score below and above which lie 50 percent of the cases. The median is also known as the fiftieth percentile. Since median is not affected by extreme scores, it is a better measure of central tendency in skewed distributions. For example, in Figure 3a illustrating positively skewed distribution, the mean is unrealistically pulled up by scores of 100 and above. Similarly, in negatively skewed distribution, shown in figure 3b, the mean is pulled down by low scores of 40 and below. These means may not give the true picture above the central tendency of the distribution. You may conclude from these examples that median is better measures of central tendency in skewed distributions.

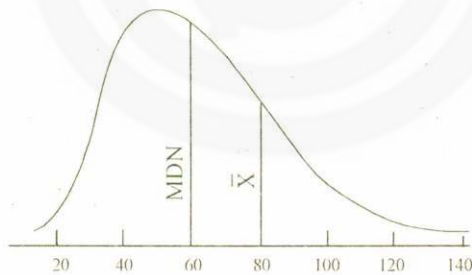


Fig. 3(a): Positively skewed distribution

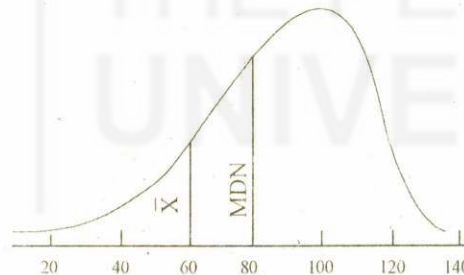


Fig. 3(b): Negatively skewed distribution

(c) The Mode

Mode is another measure of central tendency. The mode is that score in distribution which occurs most frequently. The mode is easy to determine but it is least stable. A few scores can influence the mode considerably. If a distribution has more than one score that occurs with equal frequency, then it has no mode as exemplified below:

Score	Frequency
100	1
95	2
90	6

80	10
75	10
60	9
50	10
45	4
40	1

Here, 80, 75 and 50 occur 10 times each. The distribution then becomes rectilinear.

Measures of variability

Mode is used as measure of central tendency in normal distribution only. Consider the scores of three classes A, B, & C with mean 50. Is the distribution same for three classes? No. Therefore, to describe a distribution you also need information about the spread of scores. Variability is the estimate of how “compressed” or “expanded” the distributions are. The measures of variability are the range, the Semi-Interquartile range and the standard deviation.

(a) The range

The easiest estimate of variability is the range (R). The range is the distance between the lowest score and the highest score and can be determined by subtracting the lowest score from the highest score. The range is unstable if the distributions have extreme scores. A single very high score or a single very low score can increase R, giving the impression that there is large spread of scores.

(b) The Semi-Inter Quartile Range (SIQR)

The instability of range due to extreme scores can be overcome by Semi-Inter Quartile Range (SIQR). In computation of SIQR, the lowest 25 percent and upper 25 percent are not considered.

$$SIQR = \frac{Q_3 - Q_1}{2}$$

Where Q_1 is the point below which 25 percent of the scores lie and Q_3 is the point below which 75 percent of the scores lie. Although SIQR is not influenced by extreme scores, it is not very stable because it considers only half the scores in the distribution.

(c) The Standard Deviation

The standard deviation is considered as more stable and useful measure of variability and its computation is based on all the scores. The standard deviation is determined on the basis of deviation of each score from the mean. Since sum of these deviations would be zero (property of mean), the calculations of standard deviation are based on squared deviations.

$$S.D = \sqrt{\frac{\sum x^2}{N}}$$

Where x is the deviation of each score from the mean, x^2 is the sum of squared deviations, and N is the number of students.

One standard deviation (+ and -) includes the middle 68% of the scores.

The Standard scores

If a student gets 85 marks in Physics and 70 marks in Biology he/she believes that his/her performance is better in Physics. But if 70 is the highest score in Biology and 95 is the highest score in Physics, he/she may conclude that he/she is better in Biology. Here, we are talking in terms of percentages. Comparison of raw scores becomes much more difficult if these marks are out of 200 in one case and 400 in second case. Comparison of raw **scores** from two distributions with different means and S.D.'s not useful. For example, consider these scores.

History	Geography
$X = 85$	$X = 80$
$X = 60$	$X = 120$
S.D = 10	S.D = 20

Maximum marks for History are 100 and Geography 200.

The conversion of these raw scores to a common scale may enable you to compare these scores more meaningfully. One way of doing this is to convert raw scores into z-scores. The conversion of raw scores into z-scores enables us to determine a score's exact position in the normal distribution.

$$z = \frac{X - M}{S.D}$$

Where

Z = z score (standard score)

X = obtained raw score.

M = Mean score.

S.D. = Standard Deviation.

In the above example,

History

$$z = \frac{85 - 60}{10} = \frac{25}{10} = 2.5$$

Geography

$$z = \frac{80 - 120}{20} = \frac{-40}{20} = -2.0$$

You know in normal distribution, mean is zero and S.D is one. Z score can be positive and also negative. Some of you might experience difficulty in dealing with negative numbers. To avoid this difficulty, you can convert z-scores to T-scores. The T-scores distribution has a mean of 50 and a standard deviation of 10.

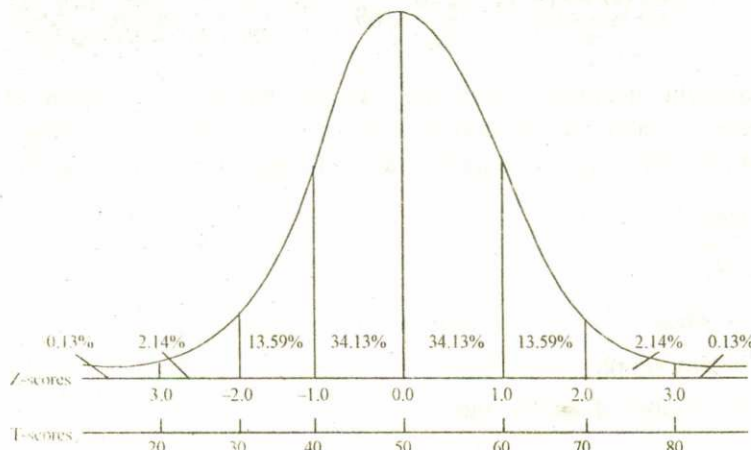


Fig. 4: Conversion of Z-scores to T-scores

Activity 4

Suppose a student gets 75 marks out of 150 in English from university A and another student gets 90 marks out of 200 in English from university B. Compare the marks of these two students on Z and T scale. The mean of university A is 60 and S.D. is 10 whereas mean of university B is 100 and S.D. is 15.

$$T = 10z + 50$$

The z score of 2.5 and -2.0 can be converted to T score as follows:

$$\begin{aligned} T &= 10(2.5) + 50 \text{ and } T = 10(-2.0) + 50 \\ &= 25 + 50 &= (-20) + 50 \\ &= 75 &= 30 \end{aligned}$$

T-distribution gives negative numbers only when raw scores more than five standard deviations below the mean are obtained which is rare.

Self-assessment

3. What are three measures of central tendency?
Which of these is used the most and why?
4. What are the limitations of using range (R) and Semi-InterQuartile Range (SIQR) as measures of variability?
5. What are the advantages of converting raw scores to z scores and T-scores?

Computation of test norms

The test norms discussed in the earlier section can be computed from both ungrouped and grouped data using simple statistical formulae.

Measures of central tendencies

It may be necessary to represent the level of performance of a group of students by a single score on certain occasions, particularly when the group is large—say, with more than 25 students. One of the ways to do this is to calculate the average of the set of scores. For convenience, let us calculate the average performance of four students with the scores 64, 69, 71 and 72.

$$= \frac{64 + 69 + 71 + 72}{4} = \frac{276}{4} = 69$$

This is known as the arithmetic mean and it represents the typical value of a set of scores. The mean is calculated by adding a set of scores and then dividing this sum by the number of scores. For ungrouped data, the formula for computing the mean is.

$$M = \frac{\sum X}{N}$$

Where M = Mean
X = any score,
N = number of scores, and
 $\sum$ = sum of

Apart from the mean, two other measures of central tendency are the median and the mode. The median is the mid point of a set of scores, that is, the point on either side of which, half the scores occur. If the number of scores is even, the median is half way between the two middle-most scores. When the number of the scores is odd, the median is the middle score. The mode is the score which occurs most frequently.

Consider the following set of scores obtained by 13 students. These are arranged in a descending order.

42
40
36
33
32
29 – Median (Mid Point i.e. 7th number out of 13 students)
28
28
28
27
25
20

Score 29 is the median because it is the mid point – the 7th number out of 13 students. In this distribution, score 28 is the mode because that is the score obtained by 3 students – the highest number of students getting a particular score.

It may be noted that the mean takes into account the value of each score. Even one extremely high or low score could have an appreciable effect on the mean. The median is only a point which divides a set of scores into two equal halves. It is, therefore, not affected by an extremely high or low score. The mode is simply the most frequent or popular score in the set. It is the least reliable measure of central tendency and is used merely as a preliminary and crude estimate.

Measures of variability

In addition to measures of central tendency, which indicate the performance of a group of students by a single score, one may like to know about the variability, spread or dispersion of the scores. One measure of variability is the range, that is, the difference between the highest and lowest scores. Another measure of variability is called the quartile deviation. It is based on the range on the middle 50 percent of the scores. It is the distance between the third and the first quartiles divided by two. Quartiles are merely points which divide a set of scores into quarter, and are indicated by Q_3 , Q_2 and Q_1 . The third and first quartiles (Q_3 and Q_1) respectively, are points below which 75% and 25% of the scores lie. The middle (second) quartile, or Q_2 , is the median. Quartile deviation can be computed by applying the following formula.

$$Q = \frac{Q_3 - Q_1}{2}$$

You may use the same counting procedure to locate Q_3 and Q_1 that was used to find the median.

The most useful measure of variability of scores is the standard deviation. It is an average of the degree to which a set of scores deviate from the mean. It is a more stable measure of the variability than the range or quartile deviation, because it takes into

account the amount by which each score deviates from the mean. The computation of the standard deviation can be expressed by the following formula.

$$S.D. = \sqrt{\frac{\sum x^2}{N}}$$

Where X = deviation of each score from the mean and

$\sum X$ = summations of all such derivations squared.

One standard deviation (plus and minus) includes the middle 68 per cent of the scores approximately.

Measures of relative positions

A teacher may like to know about the performance of a student in relation to the performance of other students in the class. You have seen that the median is a point below which 50% of the scores lie. If a student's performance is indicated by the median i.e., say he is above the median, then the teacher infers that his level of performance is more than half the students in his class. Similarly, a student who is above the Q_3 level, indicates that his superior performance is better than 75% of his class-mates. Now, one can easily compute other points in a set of scores below which a certain percentage of scores are distributed. Suppose a student gets a score of 39, has surpassed the performance of 90% of the students in his group, then it may be said that the percentile and (or percentile score) of the student is 90. Other scores which indicate the student's relative level of performance in terms of the percentage of students scoring below him, can be similarly found out. Thus, a percentile rank describes an individual's performance in terms of the percentages of students he surpasses in some clearly defined group. Tables presenting percentile norms can be prepared to indicate percentile ranks corresponding to various raw scores. It may be noted that the procedure used for calculating the median Q_3 and Q_1 may be applied for computing percentile ranks.

Computing statistics from grouped data

In the preceding sections, we discussed how to compute statistics from ungrouped data. Such procedures are helpful where the number of scores is small. When the scores to be analysed are more than twenty five, it is desirable to prepare a frequency distribution table and then proceed to compute the statistics. Procedures for computing the median, mean and standard deviation from the grouped data are given below.

Computing the median

Consider the frequency distribution table.

Table 4: Frequency distribution of a set of scores

Class Interval	Frequencies
(C.I.)	(f)
23-25	10
20-22	5
lower limit = 16.5 17-19	3
14-16	7
11-13	3
5-7	2
N = 33	

You know that the median is the 50 per cent point in a set of scores. We count off one half of the scores to locate the median in ungrouped data. The same counting procedure is used when computing the median from grouped data, but some adjustments are necessary because the scores are grouped into class intervals. The following steps illustrate the procedure for computing the median from the frequency distribution table given above.

- (i) Divide the total number (N) of scores by 2.
 $= 33/2 = 16.5$
- (ii) Starting at the low end of the frequency column, add the frequencies until the C.I. containing the median is reached ($2+3+3+7=15$). The median will be at some point in the C.I. 17-19.
- (iii) Subtract the sum obtained in step 2 from the number required to reach the median ($16.5-15=1.5$).
- (iv) Find the proportion of the median interval that is to be added to its lower limit (16.5) by dividing the number obtained in step 3 (1.5) by the number of frequencies in the median class interval (3) and multiplying this by the size of the C.I. i.e., $(3) = 1.5/3 \times 3 = 1.5$.
- (v) Add the amount obtained in step 4 to the actual lower limit of the median C.I.

This sum is the median $= \frac{1.5}{3} \times 3 = 1.5$.

The following formula can be used to compute the median

$$L + \frac{\left(\frac{N}{2} - F\right)}{f} \times i$$

Where

L = actual lower limit of the C.I. in which the Mdn. lies, (16.5)

N = total number of scores, (33)

F = sum of all frequencies below the C.I. in which the Mdn. lies, (15)

f = frequency of the C.I. in which the Mdn. lies, and (3)

i = the size of the C.I. (3)

Applying this formula, the Mdn. can be computed as follows:

$$\begin{aligned} & L + \frac{\left(\frac{N}{2} - F\right)}{f} \times i \\ &= 16.5 + \frac{16.5 - 15}{3} \times 3 \\ &= 16.5 + \frac{1.5}{3} \times 3 \\ &= 16.5 + 1.5 = 18 \end{aligned}$$

The same procedure, with necessary modifications, can be used to compute Q1

Q3, etc., from grouped data. Formula for Q1 and Q3 will be as follows:

$$L + \frac{\left(\frac{N}{4} - F\right)}{f} \times i \text{ and } L + \frac{\left(\frac{3N}{4} - F\right)}{f} \times i$$

It may be noted that different percentile points can also be computed by suitably modifying these formulae.

Self-assessment

6. Compute the median from the frequency distribution table given below:

Frequency distribution table

(C.I.)	(f)
36-38	2
33-34	4
30-32	6
27-29	7
24-26	5
21-23	4
18-20	3
15-17	2
12-14	2
$N = 35$	

Computing the mean and standard deviation

You know that mean is determined by dividing the sum of scores by the number of scores, and the standard deviation is a measure of the variability of scores around the mean. The formulae for computing these statistics from ungrouped data have been discussed earlier. For computing the mean and standard deviation when data are grouped, procedures and formulae are modified slightly. The following steps illustrate these. Let us consider the frequency distribution table used earlier to compute the mean and S.D.

Table 5: Frequency distribution of a set of scores

C.I.	Md.pt.	f	d(Deviation)l*	fd	fd ²
23-25	24	10	3	30	90
20-22	21	5	2	10	20
17-17	18	3	1	3	3
14-16	15	7	0	0	0
11-13	12	3	-1	-3	3
8-10	9	2	-2	-6	12
5-7	6	2	-3	-6	18

$$N = 33$$

$$\sum fd = 28 \quad \sum fd^2 = 146$$

(*Deviation is also expressed by 'X').

- Select any class interval near the middle of the distribution and call this interval zero in the deviation (d) column. The mid-point of this C.I. is the assumed mean, or A.M. (in table 4, the zero interval is 14-16, and A.M. = 15).

- (ii) determine the deviation (d) of each C.I. from the zero interval and enter in the (d) column.
- (iii) In each row, multiply the entry in the fd column by the entry in the d column, and enter the result in the fd2 column.
- (iv) In each row, again multiply the entry in the fd column by the entry in the d column, and enter the result in the fd² column.
- (v) Add the entries in the fd column to get $\sum fd$ (in table 5, $\sum fd = 28$)
- (vi) Add the entries in the fd² column to obtain $\sum fd^2$ (in table 5, $\sum fd^2 = 146$)

Now the mean can be computed by using the formula:

$$M = AM + Ci$$

Where AM is the assumed mean, C is correction and i the size of the C.I.

$$C = \frac{\sum fd}{N}$$

$$M = AM + Ci$$

$$= 15 + \frac{28}{33} \times \frac{3}{1}$$

$$= 17.54$$

Standard deviation can be calculated by using the following formula:

$$S.D. = i \sqrt{\frac{\sum fd^2}{N} - C^2}$$

Applying this formula to the data in table 5, S.D. can be computed as follows:

$$S.D. = \sqrt{\frac{\sum fd^2}{N} - C^2}$$

$$= 3 \sqrt{\frac{146}{33} - \left(\frac{28}{33}\right)^2}$$

$$= 3 \sqrt{4.424 - 0.719}$$

$$= 3 \sqrt{3.705}$$

$$= 3 \times 1.925$$

$$= 5.77$$

The computations above indicate that the distribution of scores present in table 4 can be described by a mean of 17.54 and a standard deviation of 5.77. Thus we can expect approximately two-thirds of the scores to fall between 11.77 and 23.31.

Computing standard scores

You know that the performance of students in different subjects cannot be directly compared with the help of raw scores. The raw scores have to be converted into standard scores, which are based on the means and standard deviations of the distribution of scores in different subjects. Standard scores can be computed by using the formula given below:

$$Z = \frac{X - M}{S.D.}$$

Where Z is the standard score,

X is the raw score,

M is the mean, and

$S.D.$ is the standard deviation

The standard scores are also called Z scores.

The derived score distribution will always have $M = 0$ and $S.D. = 1$. Thus, they can be directly compared. With the help of these converted scores, the teacher can directly compare the scores of students on tests of different subjects or characteristics.

Self-assessment

7. The raw scores of 42 students on a test are given below. Prepare a frequency distribution table, taking 10 as the size of the C.I. Let the highest C.I. be 70-79.

17	16	14	20	22	50
79	12	18	16	15	30
33	34	56	57	62	71
23	24	21	19	16	25
40	40	45	36	37	27
32	32	35	35	53	66
42	45	46	22	25	36

8. Compare the mean and $S.D.$ from the frequency distribution table prepared by you.
9. What is the use of converting raw scores into standard scores?

Summary

In this unit you have learnt how evaluation data obtained from tests can be made more meaningful for decision making purposes such as selection, promotion or placement. Pupil evaluation data are generally in the form of marks or grades. We have discussed advantages and disadvantages of both marking and grading systems. The grading can be norm-referenced or criterion-referenced. For improvement of teaching-learning process in the classroom, criterion-referenced grading is useful whereas for classification, selection and certification, norm-referenced grading may be used. Marks and grades must be interpreted with reference to test norms. The test norms are measures of central tendency, measures of variance and standard scores. These can be computed from ungrouped and grouped data using simple statistical formulae.

Unit-end activities

Take the marks obtained by a class of students in the last year examination conducted in your college/university and compute the mean, median, mode and standard deviation. If the sample is too large, draw a sub-sample of 100 students for the computations. Compute these statistics for at least three subjects in that examination taken by the same class of students. Are the statistics same for these subjects?

Points for discussion

(To be taken up in Counselling Sessions and Peer Groups)

What will be the consequences, if grading system is adopted by your university? Would you support this reform? Would student and general public accept it? What can be done to get the students to support this reform?

Suggested readings

- Association of Indian Universities (1973), *Examination Reforms—A Plan of Action*. AIU: New Delhi.
- Association of Indian Universities (1977), *Monograph on Grading for Universities*. AIU: New Delhi.
- Freeman, Richard and Lewis, Roger (1998) *Planning and Implementing Assessment*. London: Kogan Page.
- Garrett, Henry E. and Woodworth, R.S. (1961), *Statistics in Psychology and Education*. Vaklis: Feffer and Sinns Ltd. Bombay.
- Government of India (1986), *National Policy on Education*. MHRD Department of Education: New Delhi.
- Gronlund, Norman E. (1981), *Measurement and Evaluation in Teaching*. (4th ed.), Macmillan: New York.
- Pole, Christopher J. (1993), *Assessing and Recording Achievement*. Open University Press: Buckingham, Philadelphia.
- Prakash, V. and others (2000), *Grading in Schools*. National Council of Educational Research & Training: New Delhi.
- Torrance, H. (1995), *Evaluating Authentic Assessment*. Open University Press: Buckingham, Philadelphia.
- Ward Annie, Stoker Howard and Murry-Ward Mildred (1996), *Educational Measurement*. University Press of America Inc.

Answers to self-assessment

- I. (i) No test is perfectly reliable. All tests have error of measurement. Because of this error of measurement, the obtained scores are not true cores. These contain error scores.

$$\text{Obtained Score} = \text{True Score} + \text{Error Score}$$
 Error score thus represents the deviation of obtained scores from true score.
- (ii) Marks are estimates of students' achievement. Taking marks as absolute measure of achievement creates problems.
 Obtained marks contain errors of measurement.
 All examiners do not use the same scale of measurement. Some examiners may use the entire scale of 0 to 100 whereas others may use 30 to 70 scale. Comparison of marks assigned by these examiners creates problems. Subjectivity in marks due to examiners' different standards, test, examinees, etc. makes marking system unreliable.
- (iii) The sources of error scores are:
 - (a) Examiner subjectivity, i.e., different examiners have different standard of marking.
 - (b) Subjectivity in test itself: The tests have subjectivity in selection of sample of content area objectives and also in scope of answers.
 - (c) Conditions of test administration and attitude of supervisors during examination may also contribute to error scores.
 - (d) The examinee also is a source of error in scores. The examiners physical and emotional well being, fatigue, sleeplessness and other factors also continue to score subjectivity.
 - (e) Errors in scoring also contribute to error score.

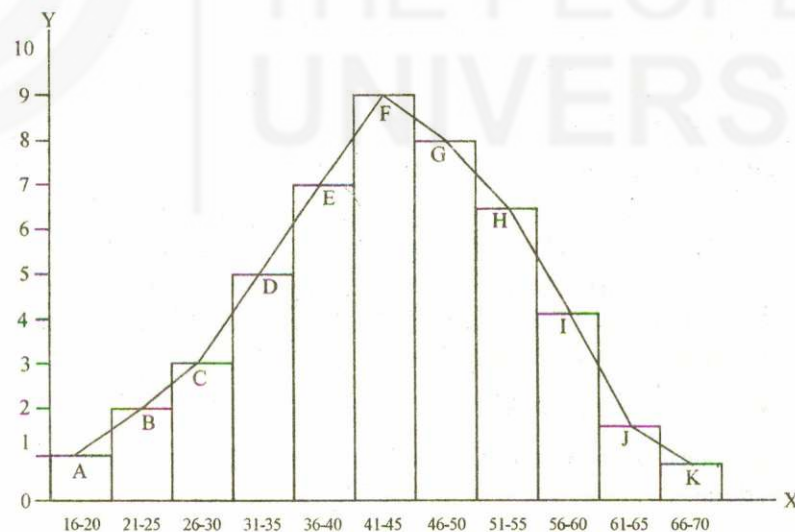
(iv) Norm referenced grading is based on relative achievement, that is, how a student's achievement compares with achievement of other students in the class or group. Criterion referenced grading compares the achievement of a student with reference to a prior defined criterion or standard, that is, whether a student has answered more than 60% items correctly or not, irrespective of other students' performance.

(v) Limitations of norm-referenced grading

- The number of students to be placed in grades A, B, C or D are prior defined.
- The same number of students get 'A' or 'B' grade each year and in all the universities.
- Therefore, there is no indication or even feedback for teachers teaching poorly. This does not help in improving learning process.
- Interpretation of grades by students and teachers is difficult.

2. Frequency distribution table:

C.I.	f
66-70	1
61-65	2
56-60	4
51-55	6
46-50	8
41-45	9
36-40	7
31-35	5
26-30	3
21-25	2
16-20	1
N = 48	



Fig

$$\begin{aligned}
 10. \text{ Mdn} &= 1 + \frac{N}{(2-F)} \times i \\
 &= 26.5 \frac{17.5-16}{3} \times \frac{3}{1} \\
 &= 26.5 + 1.5 = 28
 \end{aligned}$$

3. The three measures of central tendency are: Mean, Median and Mode. Mean is the most frequently used measure of central tendency because it is more stable.
4. Range is unstable if the distribution has extreme scores. SIQR does not take into account the entire range of scores. It considers only half the scores in the distribution.
5. Raw scores from two different distributions with different means and SDs are not comparable. Conversion of raw scores to two scores or T scores makes this comparison possible.

7.

C.I.		f
70-79	//	2
60-69	//	2
50-59	////	4
40-49	//// /	6
30-39	/// /// /	11
20-19	/// ////	9

N = 42

8.

C.I.	Mid Pt	f	d	fd	fd ²
70-79	74.5	2	4	8	12
60-69	64.5	2	3	6	18
50-59	54.5	4	2	8	18
40-49	44.5	6	1	6	6
30-39	34.5	11	0	0	0
20-19	24.5	8	-1	-8	8
10-19	14.5	9	-2	-18	36

N = 43 $\sum fd = 2$ $\sum fd^2 = 116$

$$M = AM + C_i$$

$$= 34.5 + \frac{2}{42} \times \frac{10}{1}$$

$$= 34.5 + .47$$

$$= 34.97$$

$$S.D. = i \sqrt{\frac{\sum fd^2}{N} - c^2}$$

$$= 10 \sqrt{\frac{116 - (.47)^2}{42}}$$

$$= 10 \sqrt{2.76 - .2209}$$

$$= 10 \sqrt{2.5391}$$

$$= 10 \times 1.6$$

$$= 10.6$$

9. Standard Scores can be directly compared.

Alternative Evaluation Procedures

Introduction

Evaluation is an integral part of teaching-learning process and not apart from it. One of the important functions of evaluation is to provide feedback to students and teachers, which may help in the improvement of teaching-learning process. Written tests whether objective or essay type provide information about the product of a task but do not provide adequate information about the process of learning in terms of students' performance on the learning tasks or about the performance itself. These tests also fail to evaluate the other equally important non-cognitive objectives such as habits, interests, attitudes, values, ability to work in collaborative situations, etc. For evaluation of students on these learning outcomes, we need evaluation tools other than tests, such as anecdotal records, rating scales, checklists, inventories, rating scales, socio-metric techniques, self-reporting techniques, etc. The objects of evaluation can be students' performance on learning tasks, portfolios, responses on inventories, teachers' observations, peer appraisal, etc. These performances or observations or self reports can be evaluated qualitatively. The qualitative data generated by these procedures can be meaningfully interpreted, and reported to different users.

This unit discusses these alternative procedures of evaluation, which are important for evaluation of process objectives, product of task, performance itself, attitudes and values.

Learning outcomes

After going through this Unit, you will be able to:

- explain the need for alternate evaluation procedures;
- illustrate the use of performance assessment and portfolio assessment ;
- develop rating scales, checklists, anecdotal records, etc.;
- use alternate evaluation assessment procedures for continuous and comprehensive evaluation of students;
- process the qualitative data; and
- interpret the results and provide feedback to the students.

Alternative techniques of evaluation

As discussed in Unit 15, tests are generally used to evaluate the outcomes/products of learning in the form of assessment objective like knowledge, comprehension, analysis,

synthesis or evaluation. For evaluation of process objectives we require other techniques of evaluation such as observation, self reporting and peer appraisal. These techniques also help in the assessment of co-scholastic areas such as arts, music, physical education and personal and social qualities such as social and emotional competence, interest, attitudes, values, etc. A variety of evaluation tools can be constructed using each of these techniques. Anecdotal records, rating scales, observation schedules, checklists and performance tasks are based on observational techniques. The self-reporting technique uses tools such as portfolio, inventories and interviews. Peer appraisal technique involves use of tools like socio-metric tools. These techniques generate qualitative evaluation data. In some cases, evaluation can be in terms of specific comments or it can be descriptive.

The following section discusses techniques and tools of evaluation, which generate qualitative data.

Observational technique

Observation is one of the most prevalent methods of inquiring characterizing behavioural research. A good observation is purposive, carefully planned systematically organized and expertly recorded. Observational techniques can be used to evaluate process objectives. Teacher's observation of students' behaviour in collaborative work situations, participation in group discussions, interaction during teaching-learning process form the basis of observational techniques. While students perform learning tasks, they may provide useful feedback to both teachers and learners. The observational tools found particularly useful by teachers are observation schedule, rating scales, check lists, anecdotal records and score cards.

(a) Observation Schedule

It is one of the most common tools of observation used to record observational data in different situations. To exemplify, let us take a case where the observer/s observe/s a teacher's class to see whether his/her teaching is effective. For careful observation and making it purposeful, the observer must be cognizant of criteria of improved or effective teaching.

1. List of significant criteria may be worked out.

For example

- (a) Stating intended learning outcomes (ILOS)
 - (b) Pre-requisite testing of entry behaviours
 - (c) Relevance of instructional Strategy to ILOS
 - (d) Index of student's involvement in teaching
 - (e) Use of quality questions as a teaching desire
 - (f) Use of thought provoking assignments
2. Observing teachers' lesson and record evidences corresponding to each criterion applied during teaching. For example:
S/he can observe
 - (a) (i) announcement of ILOs. at introductory stage
 - (ii) distribution of list of ILOs etc
 - (b) (i) use of questions to test relevant basic concepts
 - (ii) clarify, elaborate or re-teach those concepts, if needed
 - (c) (i) use of problem solving approach

- (ii) use of relevant teaching learning aids
 - (iii) undertaking formative evaluation
 - (d) (i) teacher domination in lower level ILOs
 - (ii) pupil domination in higher level ILOs
 - (iii) emphasis on independent learning
 - (e) (i) use of different levels of questions catering to all
 - (ii) encouraging students to pose questions at higher level
 - (f) (i) giving assignments for reinforcement and enrichment of learning
 - (ii) requiring students to undertake projects
3. Observer is to rate against each evidence gathered corresponding to each criterion, on a three (A-B-C) or 5 point (A-B – C-D-E) Scale using Good, Satisfactory, Poor for 3 points or Excellent, Good, Satisfactory, Unsatisfactory, Poor in case of 5-point scale.

Each evidence collected should be properly recorded, analysed and interpreted cautiously. If more than one observers (2 or 3) are there, rating can be ensured more objectively.

(b) Anecdotal records

Some aspects of behaviour cannot be evaluated by observing the student under a controlled situation at a particular fixed time decided by the evaluator. They have to be evaluated through certain incidents or anecdotes as they take place in the usual course of the day. Such **incidents or anecdotes** may be recorded as and when they occur. From these recordings the teacher will be able to evaluate specific aspects of behaviour of the students. The technique, where **information is provided in terms of record of anecdotes**, is called anecdotal records. An anecdotal record is a verbal snapshot of a single but significant event recorded in its setting to describe observations made by the observer. These descriptions are factual record and not interpretations of what happened. Such records are to be interpreted very cautiously separately keeping in view the context or setting in which the event or behaviour has occurred.

Suppose you find that a student always remains aloof and avoids taking part in group activities, it shows that the particular student is not socially well adjusted. As a teacher you may arrange certain programmes through which the student may try to change his behaviour. And you would like to evaluate such changes in behaviour through observation. Needless to say, you would like to record several anecdotes, which describe such changes in the social behaviour of the student. This record of social incidents may be considered as an anecdotal record. The anecdotal record is just a description of certain specific but significant incidents. In using this technique, you will have to record several incidents with regard to the behaviours to be evaluated over a fairly long period of time. And then, you as a teacher, will have to note down any consistency or otherwise and interpret them. Such interpretations would refer to any pattern that is emerging—the learning outcomes desired as the effect of the programmes arranged by you. The anecdotal records are also useful for **noting** and then **evaluating** certain **unusual aspects of social and emotional behaviour**, which otherwise are missed or ignored as irrelevant. However, each anecdotal record refers to a single but significant event or behaviour observed in a particular context or setting and therefore, interpreted cautiously in terms of its setting or context.

(c) Rating Scales

Rating scales can be used to assess the performance of students on learning tasks and are useful tools for performance assessment involving qualitative description of a limited number of traits of a person or aspects of a thing.

For instance, you want to assess your students on oral communication performance task and written communication performance task.

Task – 1

Your students want to bring some reforms in college library by changing rules and procedures. Ask them to persuade you as principal of the college to change the rules and procedures.

The performance of students can be evaluated using a rating scale. For that, you first work out the elements on which you want to evaluate your students and then assign a value to each of these elements.

For example, you select the following elements for evaluation:

- (vi) Ideas/points in the message
- (vii) Language to convey the message
- (viii) Delivering the message

You may evaluate your students on each of these points as very good, good, fair, satisfactory and unsatisfactory.

Task – 2

Your students want to send the message to the principal of college through written communication. For this task, you may select the elements as:

- (ii) Content of the message
- (iii) Organisation of the message
- (iv) Mechanics, that is, grammar, spelling, punctuation
- (v) Language—use of effective words and phrases.

You may rate your student on each of these elements as very good (5), good (4), fair (3), satisfactory (2), and unsatisfactory (1).

The performance can be rated 'holistically' or 'analytically'

Such as a technique, specifying the different aspects being observed and providing for rating each aspect on a separate scale is called **rating scale**. A rating scale thus can be descriptive—referring to the specific aspect of the task, or it can be quantitative, or both. Sometimes the descriptions are mainly classificatory; e.g. (A) sports persons (including in the University team), (B) those interested in extra-curricular activities, (C) those interested only in academic studies.

But in many cases the rating suggests **some order**—a **higher or lower rank** on a particular dimension.

For example:

- (ii) Very good in sports
- (iii) Participating in sports
- (iv) Not participating in sports but likes to observe others playing
- (v) Not interested in sports at all

In such cases where there is some order (ordinal scale) or certain quantitative ranking—the rating scale can be indicated by a quantitative numerical score such as 4, 3, 2, 1 or 3, 2, 1, 1 for the four categories in the above illustration. The numerical order indicates the increase in quality. These numbers can be assigned to each aspect of the performance. The total of these numbers is sometimes taken as specifying the amount of learning. Rating scales are mainly used to evaluate different skills (handwriting,

signing, etc.) as also several aspects of personal-social behaviour. Rating scale can be of 3, 5 or 7 point depending upon the refinement and practicability. In classroom situation 3 point or 5 point rating scale is more usable. Such scales can be

- (a) descriptive like, Excellent, Good, Average, Fair, Poor
- (b) numerical 4, 3, 2, 1, 0
- (c) graphical when 0-100 continuum on a line is to be checked by the student at appropriate place.



Difficulty in defining the traits, halo effect of carrying effect of a judgment from one trait to another, tendency of raters to be generous in rating are some of the glaring limitations of rating scales.

(d) Checklists

The checklist is similar in appearance and use to the rating scale. The basic difference between the two is in the **type of judgement** called for. A rating scale provides an opportunity to indicate the degree to which a characteristic is present. The checklist on the other hand, basically records whether a characteristic is present or absent, or whether an action has taken place or not. **Checklists are useful in evaluating those performances**, which can be divided into a series of clearly defined specific actions. For example, it may be possible, through a checklist, to record the actions of a student as he attempts to locate an object under the microscope. In the area of personal-social development, the **checklist** can serve as a convenient method of recording evidence of growth toward specific learning outcomes. (compare the checklist and rating scale given in Unit 14, Block 3).

Checklist is just like a laundry list which can be recorded or indicated in the form of 'Yes' or 'No'. Half way responses are not admissible in such a list. To prepare a checklist we have to analyse the job, the task, skills, etc. into various components which could be included in the list. Such a checklist facilitates the recording of observations. Work habits, health habits and study habits or skills are relevant areas for assessment through checklists. Example of a checklist on Study Skill.

- (a) Can I read silently?
- (b) Can I consult dictionary?
- (c) Can I read at reasonable speed?
- (d) Can I relate ideas in the Para?
- (e) Can I grasp the central idea of the Para, etc.?

Students are asked to tick mark the statement or cross mark depending upon Yes or No response. Analysis of responses indicate adequacies or the inadequacies in Study Skills for self evaluation.

(e) Score Card

It provides appraisal of a large number of traits or aspects and carries the advantage of both the checklist and the rating scale. Like a checklist different aspects of assessments are identified and listed. Like the rating scale each aspect is weighted by giving predetermined values. Numerical values say 4, 3, 2, 1, 0 (5 point) is given in various rating points. When any aspect is to be rated (say review exercises in a textbook) the weighted score for each aspect can be calculated as also the total weighted score as shown in the example given in Table 1.

Table 1: Score Card

Aspect	Wieghtage	Rating (0-4)	Aspect-wise weighted score
(a) Coverage of unit objectives	25	✓	25x2=50
(b) Coverage of content of unit	25	✓	25x3=75
(c) Precise, unambiguous language	15	✓	15x2=30
(d) Use of appropriate directional words	10	✓	10=10
(e) Variety of level of questions	10	✓	10x4=40
(f) Clarity about intended response	15	✓	15x2=30
	100	Total Weighted Score = 235	

Score cards are normally used for evaluating colleges, communities, textbooks, socio-economic status etc. Limitation is the difficulty of choosing relevant aspects to be evaluated, identifying and quantifying them.,

The checklist is not very useful in summarizing a teacher's general impressions regarding the personality and adjustment of students. In evaluating characteristics such as initiative, social maturity and emotional stability, for example, it is seldom sufficient to note merely whether the trait is present or absent. Here, we are more concerned with the degree to which a trait is present. In such a situation, the use of rating scale will be more appropriate.

Self-assessment

1. What are the specific uses of the following techniques of evaluation?

- (a) Observation schedules
- (b) Anecdotal records
- (c) Rating scales
- (d) Checklistss

Self reporting techniques

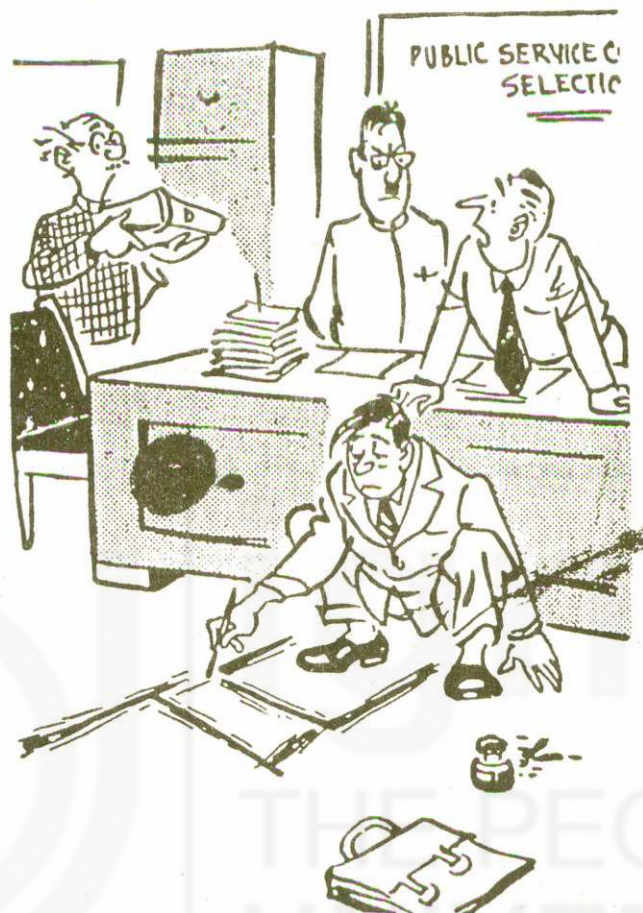
You have been told earlier that under the testing and observational techniques, information about the student is obtained **indirectly** from the product of a performance, or the performance itself. In some areas of learning it is desirable to supplement the teacher's observations with the information obtained directly from the student. For example, information on the student's **past experiences** and his own perceptions of his inner life, which are inaccessible by other means, can be collected directly from him. The commonly used self-reporting techniques are the interview and the questionnaire.

(a) Interview

Interview is a procedure under self-reporting technique in which students are made to answer several questions put to them related to specific aspect, in a face-to-face situation. In conducting the interview, one may use a set of questions specified before hand. This is known as the **structured interview**. On the other hand, if the interviewer does not have a list of pre-specified questions while conducting the interview, it is called an **unstructured interview**.

The interview as a self-reporting technique provides considerable flexibility to the interviewer. Questions can be clarified, if necessary, the interviewee can be given an opportunity to qualify or modify his answers and the interviewer can carefully observe

the interviewee during the session, noting down the feeling attached to his answers—the topics on which he seems to be evasive, and the areas in which he is most vocal. Despite these advantages, the interview suffers from two serious limitations. One, it is extremely time consuming and two, information obtained is not standard from one person to another. Often the interviewer's own biases influence what questions are asked (different questions to different candidates) and what information is recorded in the minds of the interviewer for later evaluation. Some of the interviewers become too hasty



I just asked the candidate to draw a chair and sit down!

Courtesy: R.K. Laxman

in taking judgements, while some others get influenced by others' judgements. For best results an interview has to be scheduled well planned meticulously, conducted with proper report and interpreted objectively. Therefore, in the interest of both feasibility and greater comparability of results, in large scale studies, the questionnaire is commonly used in place of personal interview, as a self-reporting.

(b) Portfolio

A portfolio is a systematic collection of a student's work. Traditionally, it was used by photographers, journalists, architects, artists, etc. Recently portfolio assessment attracted attention of educationists and has been considered useful particularly for assessment of process objectives and learning progress.

The advocates of portfolio assessment believe that this mode of evaluation is more useful for implementation of continuous comprehensive evaluation. Each student's portfolio must be updated continuously as his/her competence/skill grows. This provides students an opportunity for self-assessment. Students can be evaluated against their own earlier achievement. The portfolio assessment thus has many advantages and its use can overcome the deficiencies of test-based assessment.

- Portfolio assessment links teaching to learning and can be used for improvement of teaching and learning.
- Measures each student's achievement against his/her earlier achievement.
- Assessment is more objective and less competitive allowing for individual differences.
- Students have a goal and progress and their progress can be monitored.
- Represents wider range of content area and objectives.
- This is collaborative approach to assessment where students, teachers and parents participate.
- You must have been convinced by now that portfolio assessment serves many functions, which tests cannot. The next question that you may ask is, 'how can portfolio assessment be implemented in the classroom'? For this, first of all teacher and students must understand what is a portfolio, what kind of work should be stored in it and how it can be evaluated. You may follow the following sequencing for implementation of portfolio assessment.
 - (i) Portfolio is collection of students' work.
 - (ii) Make it **working portfolio** by allowing students to self evaluate their work, present and discuss self-evaluation among peers and teachers, compare their earlier and latter work. Continuous update the portfolio by new works.
 - (iii) What kind of **work samples be collected**? Teacher must decide in collaboration with the students, what kind of work samples can be included in a portfolio? The nature of work to be included in portfolio will vary from subject to subject. In general work samples should be collected from wide range of content requiring different kinds of abilities and measuring different objectives. That is, it should sample larger content areas and objectives.
 - (iv) **Storing work samples:** Students may submit their sample of works in folders or notebooks, which can be stored in a file cabinet or storage almirah. The piece of art works, teaching aid, etc. may be stored separately. Students may be allowed to revisit their portfolio whenever they want.
 - (v) **Selecting criteria for evaluation of work samples:** Setting up a criterion by which students' work and its quality can be judged is very important. The criteria can be decided in collaboration with students or by a group of experts. Since portfolio assessment allows diversity of products in each student's portfolio, the identification of evaluation criteria is not an easy task. The criteria statement should be clear and self-explanatory so that students do not have difficulty in evaluating their own efforts and thereafter to strive for their improvement.
 - (vi) **Continuous evaluation of portfolio products by students and teachers:** Using the agreed upon criteria students may evaluate their own work. Evaluation can be holistic or analytic or both. Self-evaluation sheet of the students must be stapled with the work sample when teacher evaluates the work, specific comments or suggestions for improvement must be indicated on the students' work and these must be conveyed to the students. Teachers' evaluation sheet must also be attached to the students' work. During second evaluation of the portfolio products, progress made from first to second evaluation must be indicated. Even if the progress is small, this will encourage students to move further towards their goal.
 - (vii) **Conduct portfolio conferences for peer interaction:** The interaction between students and teachers would give students and teachers insight into the potential use of portfolio and for better use of these assessments made. Students will be

able to judge their own evaluation against peer evaluation and teacher evaluation. Students must be well prepared for the conference so that it becomes time efficient strategy of self-evaluation. Considering the various suggestions from peers and teachers, the presenter student can improve his work sample. For portfolio conference, topics must be carefully chosen.

- (viii) **Make portfolio assessment a participatory and collaborative activity:** Parents must also be informed and oriented about portfolio assessment. Parents should be allowed to see the work samples of their wards. This will serve two purposes. Firstly, the parents will be able to know what kind of activities are done in the institutue. Secondly, they will be able to monitor the progress of their wards. When parents become actively involved in the work of their wards and their progress, students get the message that portfolio is the mirror of their progress through which parents, teachers, peers and students themselves can see their progress. Collaborative evaluation makes portfolio assessment really worthwhile.
- (ix) Finally, the show case portfolio may contain the best work samples of your students. Display of show case portfolio would enhance the image of institution, students and teachers. These may be considered for final evaluation of students.

The portfolio may thus contain the information such as:

- guidelines for portfolio reviewers
- the evaluation criteria
- skills or abilities being tested
- entries/work samples submitted by the student with dates (representative of student's best work)
- student's own evaluation on all entries
- teacher's evaluation on all entries

(c) Questionnaires

Instead of asking the questions personally the teacher may present the questions in a written form and obtain the answers from the students. This tool consists of questions to which the student answers in hiw own words. Such questionnaires are referred to as **open ended questionnaires**. On the other hand, some may have closed questions where the possible answers are already provided, such as 'yes', 'no', 'undecided', etc. The respondent has only to indicate his choice in terms of one of the options given. Such questionnaires are called **closed questionnaires**.

Questionnaires are also known as **self-report inventories**. The inventory, which is a sort of standardized written interview, may be used to obtain information regarding the student's past experiences, present activities, personality characteristics or personal-social adjustment. Questionnaires and inventories may also be used to measure the interest, attitudes and opinions of the students.

(d) Inventories

Self-reporting inventories are useful for assessment of affective outcomes such as interests, attitudes and values. Likert scales are most widely used for evaluation of affective outcomes. For example the statement, 'The stories in Panchtantra represent Indian culture and values'. The options for self-reporter can be strongly agree, agree, uncertain, disagree and strongly disagree.

Developing an inventory

For developing an inventory, the following steps may be followed:

- (i) Identify the attitude or value you want to assess. For example, you want to assess the attitude of today's youth (college students) towards their teachers.

- (ii) Write down a series of favourable and unfavourable statements which may reflect students' attitude towards their teachers. For example, the positive statement can be: 'Teachers continuously update their knowledge of the subject'. The negative statement can be: 'Teachers are busy with private tuitions ignoring their duties in college'. This way, you can generate about 15-20 positive and negative statements. Try to construct an equal number of positive and negative statements.
- (iii) Get several teachers, students or others concerned to classify these statements as positive, negative or neutral.
- (iv) Decide the response options for each statement. The response options can be Strongly Disagree (SD), Disagree (D), Not Sure (NS), Agree (A), and Strongly Agree (SA).
- (v) Prepare inventory giving directions for self-reporting, that is, how a student has to answer the inventory.
- (vi) Administer inventory to the students.
- (vii) Score the inventories. You may assign a value to each of the response options. For example, Strongly Agree (5) in positive statements and Strongly Disagree (5) as negative statements.
- (viii) Eliminate statements that fail to function. One way of finding the functionality of statement is to compute the correlation co-efficient of scores on each statement with the total score. Statements with correlations that are not statistically significant may be eliminated.

Peer appraisal

In certain situations, it is possible to have students rate their peers (fellow students). When such a procedure is adopted, it is known as peer appraisal. The two most widely used techniques in this area are the 'guess who' technique and the sociometric technique.

(a) 'Guess who' technique

In this procedure, each student is presented with a **series of brief behaviour descriptions** and asked to name those students who best fit each description. For example, the students may be presented the following descriptions.

- Here is some one who is willing to share ideas and materials with others.
- Here is someone who does not care to share ideas and materials with others.
- Here is some one who is always worried.
- Here is some one who always helps you.

It may be noted that both positive and negative descriptions are used here. This technique is based on obtaining peer ratings and is scored by simply counting the number of mentions each student receives on each description. If both positive and negative descriptions are used, as in the illustration above, the number of negative mentions on each characteristic is subtracted from the number of positive mentions. The pattern of scores for each student indicates the reputation he holds among his peers. This may or may not agree with the teacher's ratings, but it is nonetheless significant information concerning personal-social development. It makes the teacher aware of the feelings and attitudes among students that he is unable to detect through direct observation. This technique yields peer judgements which are interesting and significant and are useful in the study of individual roles.

(b) Sociometric technique

You may have observed that certain students always stay together, certain students are more liked by all students, some of them are not liked by anyone, and so on. These social relationships amongst different students, naturally influence all aspects of their

development. It is, therefore, necessary for the teacher to evaluate these **social relationships** that prevail among the students. The sociometric technique is a method for evaluating the social acceptance of individual students. It is based on students' choices of companions for some group situation or activity. For example, on the play ground the teacher may ask the 'students' whom do you want to play with? Or in another situation where a group project is to be carried out, he may ask 'whom do you want to be your leader'? The choices of the students indicate what type of relationship they have with others. By studying the choices, the teacher can determine the social relationship among the students.

The sociometric technique involves presentation of specific social situations in which the students indicate their choices or preferences for certain particular students. This technique is especially useful in identifying those who are isolated, that is not preferred by any other student. It is also useful for identifying those who are liked by many others and who can be good leaders. Depending upon the choices made in terms of performance, a sociogram can be developed indicating 1st, 2nd, 3rd choices with respect to the type of social relationships one is interested in explaining some symbol like O for boys, A for girls. Arrow, mutual arrows between two etc. can be used to develop sociogram for a class. Most of arrow going towards a boy is STAR. Arrow gives away from a boy is an ISOLATE. Three boys A, B, and C voting for each other forming a triangle shows a clique. Such a Sociogram is quite relevant for determining social relationship. (More about these techniques based on group dynamics are dealt with in Course 3).

Processing qualitative evaluation data

We have discussed in the preceding sections how certain techniques of evaluation generate qualitative data. In such cases, the learning outcomes **cannot be measured in exact quantities** that can be indicated by numbers. Instead, the learning outcomes have to be determined through the **description of changes**, which have taken place. Such descriptions help the teacher to determine the quality of the learning outcome rather than its quantity.

Sometimes the varying levels of the qualitative changes can be denoted by numbers. For example, when a rating scale is used, the levels of performance such as Very Good, Good, Fair, Poor and Very Poor can be represented by the numbers 5, 4, 3, 2, and 1. In certain other situations, when a checklist is used the frequency of occurrence of certain behaviours or presence of certain characteristics can be determined. When qualitative changes can be expressed in terms of numbers, it may be possible to compare the performance of different students in terms of the frequency of occurrence or presence of desirable events or traits. Suppose, while performing an experiment, a student is required to go through ten operations, the teacher may then compare the performance of students through a checklist. A student who performs eight operations correctly can be considered to be on a higher level of achievement than one who has correctly performed only six operations. By converting the frequencies into percentages (in this case 80 per cent and 60 per cent) the comparison of the performance of two students can be made more meaningful. This procedure is known as **frequency and percentage analysis**. It can be used in those situations where qualitative changes can be expressed through numbers.

So far we have discussed, in the previous Unit, how the performance of a group of students can be graphically represented through a histogram and a frequency polygon. Such graphical representations can be utilized to denote the varying levels of performance of a group of students. When evaluation has been carried out through such techniques as the checklists or rating scales, one may also be able to use other types of graphical representations such as line graphs, bar diagrams, pie-graphs, etc. in certain other situations.

You will agree that techniques like the open ended questionnaires, unstructured interviews, anecdotal records, etc. do not generate data, which can be represented by numbers. Verbal responses and descriptions of behaviour do not lend themselves to immediate analysis in the way that number or 'yes-no' questions do. Such data are usually analysed by the process called **content analysis**.

Content analysis

Content analysis is defined as a procedure for the categorization of verbal or behavioural data for purpose of **classification, summarization and tabulation**. Content analysis was first used as a technique for the analysis and quantification of communication material such as contents of newspapers, magazines and radio broadcast. In recent years, the technique has been used for a wide variety of purposes. In general, content analysis can be applied to data generated through open ended questioning and observation. Content analysis involves the development of a set of categories intended to represent the dimensions of the actual content of the responses. The content of responses can be examined and coded in terms of the set categories developed. It may be noted that such categorized data will be influenced to a great extent by the teacher's expectations as reflected in the categories he develops, the nature of the content selected and the nature of the analysis attempted. In a sense, content analysis is an "order" imposed by researcher and in that sense a personal statement by the teacher, on his perceptions of the data. Such analyses may, therefore, differ from person to person.

However, content analysis to be useful, needs to reduce the subjective nature of its procedure as much as possible. For this, the categories of analysis should be defined so precisely that different analysts can apply them to the same body of content and secure the same results. Essentially the procedure involves the **codification** of verbal (or written) materials which can take many diverse forms and still 'mean' the same thing, in terms of the categories under which they are subsumed. Coding is based upon the presence or absence of the categorized material, not upon its frequency; upon whether, or not how much. For example, in a study of ethnic prejudices of a group of American veterans, the respondents were classified as the intense (anti-Semitic or anti-Negro), the outspoken, the stereotyped and the tolerant, according to whether they made spontaneous restrictive remarks, elicited restrictive remarks, stereotyped responses or explicit references of a tolerant nature. Group behaviour can be analysed by categorizing statements made by group members in terms of such categories as "shows solidarity", "shows tension release", "agrees", "gives suggestion", "gives opinion", "gives orientation", etc.

The above is only a very brief outline of content analysis as a technique for processing qualitative data. You may like to read more about the details regarding the procedure of content analysis.

Self-assessment

2. *What are the advantages and disadvantages of the following techniques for evaluating the students?*
 - (a) *Interview*
 - (b) *Questionnaire*
 - (c) *Portfolio*
3. *Match the techniques of List-A with the situations of List-B. Some techniques can be useful in more than one situation.*

List-A

- A. *Anecdotal records of students' behaviour*
- B. *Rating scale*
- C. *Sociometric technique*
- D. *Interview*
- E. *Questionnaire*
- F. *Checklist*

List-B

- 1) *To rank students on the quality of leadership.*
- 2) *To find out who the most popular leader is among the students.*
- 3) *To find out problems of students regarding transport.*
- 4) *To find out problems of students against the administration of the college.*
- 5) *To rank students on their skill of performing a laboratory experiment.*
- 6) *To find out why certain students had become drug addicts.*
- 7) *To find out specific difficulties of students regarding a particular subject taught in the college.*
- 8) *To find out presence or absence of certain study skills.*

Reporting the results of evaluation

You will agree that the results of evaluation are useful to students, teachers, administrators and employers in various ways. Therefore, such results should be reported to users **in the form of a report**, which is meaningful and usable. In this section, we shall discuss the reporting of the results of evaluation to different users.

Consideration for reporting the performance

You are aware that the purpose of evaluation is to find out the extent to which students have achieved various instructional objectives. You also know that evaluation is carried out with regard to a variety of instructional objectives, which the students have to achieve. This suggests that the results of evaluation should contain information with respect to different areas of learning and all aspects of the development of students. In other words, the report of the results of evaluation should be **comprehensive**, covering both scholastic and co-scholastic aspects.

Results of evaluation contain grades or marks indicating student's performance level in different areas of learning. The user is not likely to know much about the student's performance from the letter grades or from numerical marks alone. Grades or marks, which are simply signs and symbols to indicate the student's learning outcomes need to be made understandable to all users. One way to give meaning to marks (or grades) of a student is to compare his performance with that of other students in his group. This enables the user to get some idea of an individual student in comparison to his peers. We have already discussed how in norm-referenced testing, the student's performance is compared with others in his group. The performance of a student can be compared by ranking him in the class, based on his performance. As we have discussed earlier, you may also compute the percentile rank of a student to indicate the relative level of his performance with reference to other students. It may also be possible to convert the raw scores into standard scores, so that the relative performance of students is indicated more systematically. So, evaluation results should be reported in a form which is understandable by the users.

Another, and sometimes a more useful, way is to compare the student's performance with the standard/level of achievement expected from that student in the context of his entry behaviour. Criterion-referenced tests, which we have discussed in Unit 13 of Block 3 are designed from this perspective. The feedback a student gets from the teacher from this perspective, guides and helps him to improve in his efforts to attain the mastery expected of him. This perspective also helps the teacher to provide remedial material or advice/counselling as to what a student should do in future.

Evaluation is carried out at different stages of instruction. The student is able to judge his/her progress if evaluation is carried out from time to time. Users, other than the student, may also be interested in knowing how the student has progressed. Therefore, results of **past** as well as the **present** evaluation should be reported to the users. In other words, the report of evaluation results should be cumulative in nature. Thus, a good evaluation report should be **comprehensive, understandable** and **cumulative**.

Scheme of reporting

Evaluation results are reported as grade or marks. Marks obtained in different subjects are shown in the mark sheet. It also contains the total marks in different subjects and the classification of level of performance, based on the percentage of marks obtained, such as first class, second class, pass class, etc. We have discussed in the previous Unit, the superiority of the grading system over the marking system. When performance levels are indicated by grades instead of marks, many defects of the marking system are removed. So, levels of performance of students in different areas of learning should be indicated by grades. For this purpose, the grading system has to be adopted for evaluating students' performance as we have discussed earlier in Unit 17.

Although a raw score or grade provides the summary of a student's test performance, it is not very meaningful without further information. In general, we can provide meaning to a raw score either by converting it into a description of the specific tasks the student can perform (criterion-referenced interpretation), or by converting it into some type of derived score that indicates the student's relative position in a clearly defined reference group (norm-referenced interpretation).

(a) Criterion-referenced interpretation

Criterion-referenced interpretation permits us to describe an individual's performance without referring to the performance of others. We may describe a student's performance in terms of the **speed** with which a task is performed (e.g., types 45 words per minute), the **precision** with which a task is performed (e.g., measures the diameter of a sphere within 0.1 cms), or the percentage of items correct on certain clearly defined tasks (e.g., solves 80 per cent of the questions on quadratic equations). The percentage of correct score is widely used in criterion-referenced test interpretation. You will agree that criterion-referenced interpretation of test results can be meaningful only when the test has been specifically designed for this purpose.

(b) Norm-referenced interpretation

Norm-referenced test interpretation tells us how an individual compares with other students who have **taken** the same test. As we have discussed earlier, the simplest type of comparison is to rank the scores from the **highest** to the **lowest** and to note where a particular student's score falls in the set of ranked scores. While such a simple ranking of raw scores may be useful for reporting the results of a classroom test, it is of limited value beyond the immediate situation, because, the meaning of a rank depends on the number of group members. To obtain a more general framework for norm-referenced

interpretation, raw scores are typically converted into type of **derived scores**. We have already discussed one such derived scores, the standard score in Unit 16. Other types of derived scores, which are also commonly used, are grade and age equivalents, which describe test performance in terms of the grade or age group in which an individual's raw score is just average. For converting raw scores into derived scores, one has to simply consult the table of norms. Tables of norms present scores earned by students in clearly defined reference groups, such as first year undergraduate students or eighteen year old students. These scores are obtained by administering the tests to a large number of students belonging to the reference group and represent the 'normal' or 'typical' performance of the students. Such norms enable us to answer questions such as the following:

- (i) How does a student's performance compare with that of other students?
- (ii) How does a student's performance in one test compare with his performance in another test?
- (iii) How does a student's performance in a test compare with his performance in a test administered at an earlier date?

One of the most widely used and easily understood methods of describing test performances is that of percentile rank. A **percentile rank** (or percentile score) indicates a student's relative position in a group in terms of the percentage of students scoring below him. Thus, if a student's raw score of 49 equals a percentile rank of 75, we know that his performance surpasses that of 75 per cent of the group. Percentile norms enable us to find out the percentile ranks corresponding to raw scores.

Another method indicating a student's relative position in a group is by showing how far his raw score is above or below average. This is an approach used with standard scores. As we have discussed earlier, standard scores express test performance in terms of standard deviation units from the mean.

Thus, raw scores should be converted into derived scores in reports of evaluation. One of the major advantages of converting raw scores to derived scores is that a student's performance on different tests can be directly compared. The derived scores can be computed from the raw scores of a large number of students appearing at university examinations in different subjects. They can also be worked out for college level examinations and tests by teachers. What is important is to give meaning to the raw scores (or grades) by converting them into derived score.

Combining marks or grades over different subjects

You are aware of the practice of combining marks or grades in different subjects. We have discussed how this is not a correct practice when marks on the 101 point scale are used. If marks in different subjects are combined, the absolute total is incorrect, because the ranking of a student, based on total marks in different subjects, is affected by the average marks in a subject, as well as the spread of marks in it. This is why the UGC has recommended the adoption of the grading system with fewer scale units instead of the 101 point marking system. But, you will agree that marks or grades in different subjects represent the performance of students with regard to different educational objectives. Therefore, the levels of performance in different subjects as indicated by marks or grades, should be separately shown in the reports of evaluation. By the same consideration, performance of students in internal and external examinations, as well as the theory papers and the practical should be mentioned separately in reports. This has also been suggested by the UGC while recommending the adoption of the grading system.

An indication of the total performance of a student is required in certain situations. It may be necessary while classifying the students' performance levels into distinction,

first class, second class, etc., to award medals and prizes and may be needed by employing agencies or admission committees. For this purpose, the UGC has recommended that a grade point average, calculated upto two decimal points could also be indicated. You may read more about the details of computing the grade point average in the UGC-monograph or 'Examination Reform—A Plan of Action'.

We have discussed earlier that evaluation is carried out to determine the extent to which a variety of instructional objectives have been achieved. So while evaluating the students, we should not only evaluate their achievement in different subjects, but also evaluate the various aspects of their personal and social development, such as their other interests, attitudes, social relationships, etc. The techniques of evaluation discussed in the preceding section under 'qualitative evaluation data' are useful in determining the changes in personal and social development. You will agree that an indication of changes in personal and social development should be given while reporting the results of evaluation. Generally, such changes can be presented in the form of qualitative rather than quantitative data. Needless to say, different aspects of socio-personal development have to be systematically evaluated, before they can be reported. This will require organization of a variety of learning experiences, most of them of a co-curricular nature, to enable the students to develop various aspects of their personalities.

Self-assessment

4. *Why should a good evaluation report be comprehensive and cumulative in nature?*
5. *What are the advantages of converting raw scores into derived scores, while reporting the results of evaluation of students' performance?*

Profiles

Performance of individual students can be represented by means of graphs. Such a graphical presentation, showing the levels of performance of an individual is called a 'profile'. Performance of the student is generally indicated by derived scores on the profile, so that his performance in different subjects or tests can be directly compared. The illustration indicates the levels of performance of a student in different subjects:

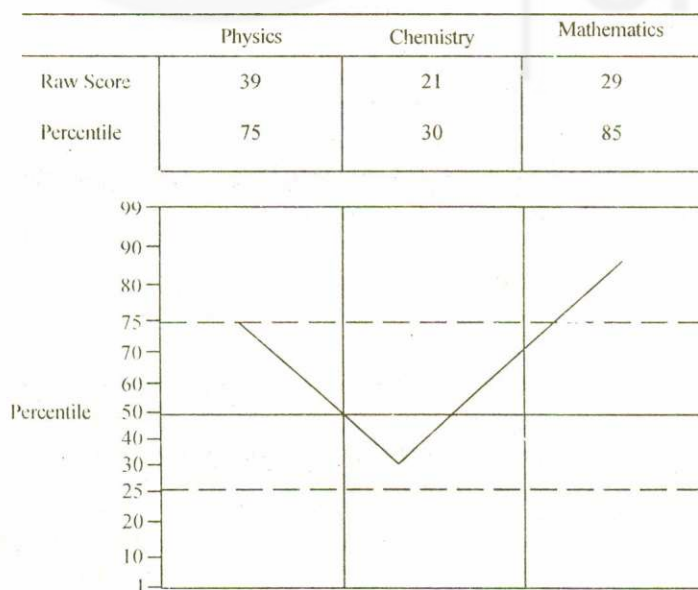


Fig. 1: Profile Chart

You can observe that the levels of performance of the student in three subjects are clearly indicated by the graphical representation above. It may be noted that although the raw scores in Physics and Mathematics are the same; the student's performance in Mathematics is superior to that of Physics, as indicated by the respective percentile ranks. Thus, by converting raw scores into derived scores and showing them graphically, a clear picture of the strong and weak points of the student's performance in different areas of learning can be indicated.

The performance of a group of students can also be shown through graphical representation. We have discussed how histogram and polygon can be drawn to represent the performance of a group of students. Performances of two or more groups of students can be indicated by showing the respective distribution of scores by different polygons in the same figure. Thus, individual and group profiles help the teachers to report the results of evaluations clearly and effectively.

Reporting results of evaluation to different users

Reports of evaluation are used by students, teachers, administrators, and others for various purposes. The main reason for reporting results of evaluation to students (and to parents), is to facilitate the student learning. For serving this specific function, the report should indicate the student's strengths and weaknesses in learning, promote greater understanding of his personal-social development, and motivate him to learn. From these considerations, most of the functions are best served by day to day evaluation and feedback during instruction. However, there also seems to be a need for a periodic summary of progress. The periodic progress report provides such a summary. In addition to giving students a general picture of how they are doing, such reports also provide them with a basis for checking the adequacy of their own learning progress. Thus, properly developed reports are likely to motivate students to learn. To serve these purposes adequately, the report will need to contain as much information and details as we can reasonably expect students (and parents) to comprehend and use.

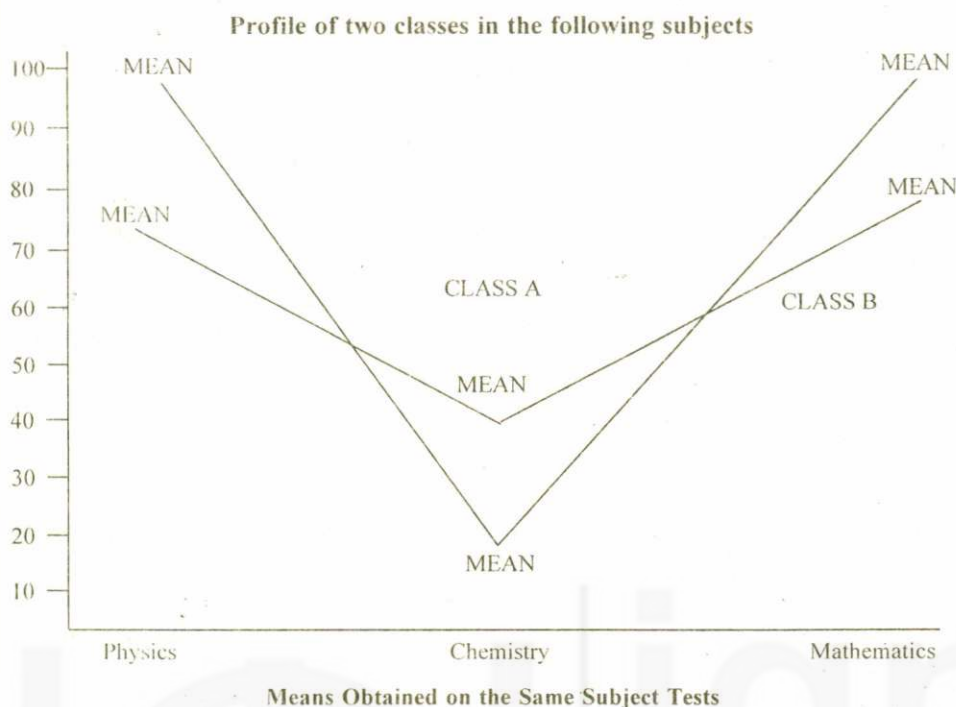
Reports of evaluation are useful to teachers in planning instruction, diagnosing learning difficulties of students and coping with social problems related to their personal-social development. Counsellors use these reports to help students develop increased self-understanding and make more realistic educational and vocational plans to carry out these functions, the reporting system needs to be both comprehensive and diagnostic.

Progress reports serve a number of administrative functions. They are used for determining promotion and graduation, awarding honours, determining eligibility levels, and for reporting student performances to other educational institutions and prospective employers. For most administrative purposes, a single letter grade (or set of scores) tends to be preferable, largely because such indicators are compact, easily recorded and averaged. But, administrative inconvenience should not be allowed to retard the development of more comprehensive and useful reports for the benefit of other users like the teachers, or learners themselves. It is possible to retain the use of their grades (or marks) for administrative purposes and to supplement them with the type of information needed by students, teachers, counsellors and employers. Employers will appreciate detailed information regarding, not only the achievement of the student, but many other aspects of his personality, so as to take decisions about the suitability of the individual, for induction into their organization.

Thus, you can see that reports of evaluation should contain information regarding the various aspects of socio-personal development of the students, so as to be useful to variety of users.

Self-assessment

Given below is the profile of two classes in the following subjects. Observe it carefully, answer the questions (6, 7) that follow the profile.



Note: Both the classes have about 60 students. But the teachers who teach the subjects are different. And the same teacher is teaching both the classes.

6. Which of the following statements are true?
- The mean score of Class A in Chemistry is higher than that of Class B.
 - The mean score of Class B in Chemistry is higher than that of Class A.
 - The composite mean score of Class B (over all subjects) is lower than that of Class A.
7. Which of the following hypotheses will you formulate for further probing?
- Students in both the classes are not interested in studies.
 - Class B students are more intelligent than Class A students.
 - The subject Chemistry needs to be looked into.
 - The teaching of Chemistry needs to be looked into.
 - The Mathematics teacher seems to be more effective than the Chemistry teacher.
 - The Mathematics test was easier than what it should have been.
 - Chemistry is more difficult than Mathematics.

Summary

Our major concern in this Unit was generation of qualitative evaluation data through alternate evaluation procedures and reporting of evaluation results. Qualitative evaluation data are generated when observational, self-reporting and peer-appraisal techniques are used. Under observational techniques, we discussed observation schedule, the anecdotal records, rating scale, checklists and score card. Under self-reporting techniques, the interview portfolio questionnaire techniques were discussed. And under peer-appraisal,

'the guess who' technique and socio-metric technique were discussed. We described how qualitative evaluation data could be processed through the content analysis method. Towards the end, we concluded that reports of evaluation should be comprehensive and cumulative. We also talked about the use of reports of evaluation to different consumers, like students, teachers, parents and administrators.

Unit-end activities

Are there any modes of evaluation other than the conventional written examinations used in your college/university? If yes, which of those are meant for assessment of non-cognitive outcomes? Collect the information about such tools/modes of evaluation from your colleagues, vocational guidance bureau or such other agencies in your institution. Also find out how the information collected through these tools is reported to the students themselves. Is it satisfactory? Would you like to make any changes in the reporting methods?

Points for discussions

We often say that education is for the development of total personality – both cognitive and non-cognitive aspects. How do we make this goal of education clearer/more explicit to ourselves and to our students? How can we get relevant information to decide whether our education system is or is not contributing to this goal of total personality development?

Suggested readings

Best, John W. and Kha, James V. (1988), *Research in Education*. Prentice Hall of India Pvt. Ltd.: New Delhi.

Fox, David J. (1969), *The Research Process in Education*. New York: Holt, Rinehart and Winston, Inc.

Freeman, Richard and Lewis, Roger (1998) *Planning and Implementing Assessment*. London: Kogan Page.

Gronlund, Norman E. (1981) *Measurement and Evaluation in Teaching*. (4th ed.), New York: Macmillan.

Gronlund, Norman E. (1974), *Improving Marking and Reporting in Classroom Instruction*. New York: Macmillan.

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

Yadav, M.S. and Govinda, R. (1977), *Educational Evaluation*. A package of auto-instructional material, Ahmedabad: Sahitya Mudranalaya.

McMilan, James H. (2001), *Classroom Assessment: Principles and Practice for Effective Instruction*. (2nd ed.).

Williams Jennifer (2000), "Implementing Portfolios and Student-led Conferences", ENC Focus.

Stiggins, Richard J. (2001), *Student-Involved Classroom Assessment*. (3rd ed.), Upper Saddle River, N: Prentice Hall.

Answers to self-assessment

1. (a) Observation schedule is a tool to record evidences about a phenomenon according to certain criteria.
(b) Anecdotal records are useful in evaluating certain unusual aspects of social and emotional behaviour.
(c) Rating scales are mainly used to evaluate different skills and aspects of personal-social behaviour.
(d) Checklists are useful in evaluating those performances which can be clearly divided into a series of clearly defined, specific actions.
2. (a) The interview provide considerable flexibility to the interviewer, but it is time consuming in nature and information collected through it is not standard from person to person.
(b) The questionnaire is useful for collecting data from a large number of individuals, but suffers from lack of face-to-face interaction with the respondent.
(c) Portfolios are tools to assess process objectives and learning progress.
3. A – 6 B – 1, 5 C – 2 D – 4, 7 E – 3
4. Report of the evaluation should be comprehensive because it represents the level of performance in different areas of learning and the aspects of personal-social development of the student. Indicators of performance, such as marks or grades, should be made understandable in terms of the performance of the group to which an individual belongs. To communicate the progress made by an individual, results of the past as well as the present evaluation should be reported. In other words, the report should be cumulative.
5. Derived score indicates the performance of an individual in terms of the group performance. Performance of a student in different subjects or tests can be compared by using the derived scores.
6. A
7. d, e, f.

Online/Web-Based Student Assessment

Introduction

If we talk about the testing in a conventional setup, student evaluation means that the student has to write an examination of two-three hours, which will test his level at a given point of time and in a locally similar type of environmental setting. The word locally is used here because, unlike Distance Education, Conventional Education is imparted in one campus only. These tests measure the knowledge and skills considered important in the learning. However, the goal of these evaluations is to just ascertain that at the time of taking the examination the student should be able to exhibit his performance on the concepts asked in the question paper. There is an immense importance of achieving intended learning outcomes. However, these kinds of evaluations are restricted to student achieving marks or grades that are as per the pre-defined pass level. These evaluations are, in their entirety, summative in nature. But what is happening today is, rather than using testing as a learning tool, students also become interested in getting taught for the test material itself. To be significant, the act of evaluation must enhance the learning process and thus comes the concept of continuous evaluation component, which is formative evaluation, intended to provide feedback to students on the errors, they have committed in understanding the concepts. In a conventional setup, it is easier to monitor the student's performance and provide feedback. However, in Distance Education setup, it is extremely difficult to do so. With the emergence of institutions like IGNOU, which has an enrollment of more than 10 lakhs of students, it is increasing becoming imminent that technology comes in a big way to the rescue. This is the main reason of our interest in online/web-based assessment which on the one hand, perform both summative and formative functions of evaluation and on the other hand remove the biases which are prevalent in human evaluation.

This unit is intended for the teachers who are using or are planning to use computers in their classrooms for enhancing the quality of the learning of their learners and also finding it difficult to manage the load of growing numbers. Computers have received their due importance in instructional activities, but there is a lot of scope for developing its role in assessment and feedback.

Learning Outcomes

After reading this unit, you would be able to:

- define the concept of online and web-based evaluation
- identify the issues involved in generating an electronic test
- discuss the theories on which the essay-evaluators work, and
- discuss other forms of evaluation which can be used in web-based evaluation.

Many of the recent advances made in educational measurement and evaluation could be attributed to computers. There is a continuous update in both hardware and software and hence there is an emergence of many newer applications. People have realized that application of these technologies can enhance the quality of learning experience. However, one must admit that the use of computers in testing is much less than its use in instructional activities. The use of computers in evaluation has probably risen due to the enhanced time, money and effort spent. Technology can be used for assessment purposes at multistages, from the management of the assessment information to a fully automated assessment system. In recent times, education world has seen emergence of several areas in computer-based assessment. There has been a growing interest and increasing practical experience in the use of computer to deliver objective tests. Objective testing is often taken to imply the use of multiple-choice questions. Many packages are available which are designed for the electronic delivery of objective tests, all of which support the delivery of a variety of question types.

In addition to using electronic packages to create unique tests, it is possible to use the computer to generate different tests automatically. Computerized delivery from a bank of questions or of randomly generated problems, supported by automated marking and feedback to the student is a flexible and efficient method of providing formative assessment, particularly when factual knowledge is an important component of the course.

In recent times the education world has seen the emergence of two efforts at using computers to evaluate the content of essays.

These are:

- Intelligent Essay Assessor (IEA), developed by Peter Follz of University of Colorado at Boulder
- Graduate Management Admission Test (GMAT), which employs a system called E-Rater

Reasons for their development were the difficulty in managing the load of grading the essays and the tendency of the evaluators to shift the focus of assessment to multiple-choice questions. In several studies IEA agreed with each human grader as often as they agreed with each other. These approaches not only save time but also provide consistency a computer can deliver. They are also helpful in training of evaluators and for giving feedback to the students. We have seen that these models are defined for Languages, Psychology, and Education, etc. But the interesting question is how do we use computerized models for Mathematics, Computer Science and other Sciences and Management subjects? Well, we have to go back to the Basics. If we go into details, there are 51 learning theories, 18 learning domains and 18 learning concepts. These theories provide an interesting opportunity for the teachers to assess their student's learning in a purposeful manner.

There are immense benefits of using online and web-based evaluation. The first and foremost benefit is in the increased frequency of evaluation that performs both formative and summative functions. Putting quizzes and assignments on web and enabling the student to do a sort of self-assessment can perform the formative function. The teacher can monitor this self-assessment and feedback can be offered to a single or a group of students as the requirement is. It can also aid in summative function as it can considerably reduce time spent in marking, which in turn can be utilized for other meaningful activities. The teachers can declare the results with a click of a button, which will eventually aid the students in their planning of further studies.

Electronic delivery of objective tests

The prevalent evaluation system, in most of the institutions, has two components i.e. the continuous assessment component carrying a weight of 30% and a term-end component carrying a weight of 70%. The former component is further broken into two components based on the nature of the assignments i.e. Computer-marked assignment (CMA) and Tutor-marked assignment (TMA). Computer marked assignments are generally multiple-choice questions (MCQ). However, objective type tests can have a lot of other formats like fill in the blanks, word entry or number entry, identifying a segment etc.

Several packs are developed and are available for electronic delivery of objective tests. These packages offer a huge variety of questions formats. These packages are capable of giving feedback and doing assessment also. You will, in this section read about a few of them. To begin this discussion it is pertinent to mention CAPES which was developed by the National Informatics Centre, India in 1993. CAPES is called Computer Aided Paperless Examination System. This system was based on optical memory technology, especially CD-ROM where large amount of data pertaining to optimally designed questions banks and answers can be stored. The system then with the use of computer networks allows conduct of examination with minimum requirement of manpower for supervision and maintaining security of the examination. The system was also capable of delivering similar tests for all examinees in terms of difficulty levels, course coverage, and biases etc. by the use of stratified sampling techniques. A test of CAPES on 6800 examinees distributed over 20 days and 18 cities demonstrated that the system has not introduced any bias in the randomization process and the time spread also does not have any biases over the scoring ability of examinees taking the test towards the end duration as compared to those at the beginning.

IGNOU has also developed an online and on-demand model called IGOLEX, which is in the testing stage at the moment. IGOLEX is defined as the IGNOU Online Examination system. In this system, it is planned that every student who wishes to appear for an examination has to register for an examination. On the basis of his registration the student would be given a password for logging on to the site. The examination would produce a different paper of equivalent standard each time an examinee logs on. The question bank in this system is prepared with different characteristics like item difficulty, expected time taken, course unit, block etc. embedded in them. As soon as an examinee submits its work the system evaluates the work. The marks are reported to the student and simultaneously get recorded in the server at IGNOU for entry into the respective student grade card.

There are several other packages that are exclusively meant for delivering objective tests. We would discuss two of them, i.e. *Examine* and *Question Mark*. *Examine* is a multiple-choice delivery system for Windows developed by University of Nottingham. *Examine* can be used not only as an assessment tool but as a teaching tool also. It can help the learner to do a self-assessment and revision. The questions can be put into a text file that can be printed if one wants so. The *Examine* system consists of two components: *Examine* is delivery engine that is proposed for the student use and *Quiz Maker* is a tutors tool that enable teacher to generate quizzes from the available questions. *Examine*'s question bank is segregated into many question folders as per the chapters in the book. These questions are of many types like Multiple-choice questions, Multiple selections, True/False, Numeric and Comments. It offers the flexibility to offer random quizzes by selecting a random number of questions from each chapter.



Examine

Examine is a multiple-choice authoring and delivery system for use either as an adjunct to courseware or as a standalone means of on-line self-assessment. A number of question types are supported, as well as graphics and multimedia. There is also support for the Microcosm hypermedia system developed by Southampton University. Recently *Examine* has been licensed to McGraw Hill for distribution with a large collection of economics questions under the name of *Begg Examine* (this will form part of a teaching pack that supports the book *Economics* by David Begg et al).



Source: <http://ibis.nott.ac.uk/software/examine.html>

Questionmark Perception is another software that helps the teachers to write, administer and report on assessments, quizzes and surveys using computers. *Perception for Web* is a package for conducting assessments using the Internet and Entrants. *Questionmark* is an authoring system with many question types. It has a good feature that enable one to deliver the assessment in an Off-line mode but still capable of storing the answer. It has integrated Sound and Video with the questions and their answers so that it enhances the effect of assessment and overall learning. These clips can be configured to auto-run. Each question contains a text note and metadata. Apart from the question type discussed above it also contains question types like matching, pull-down list, rank in order and drag & drop kinds.

Q Questionmark

Questionmark software makes it easy for educators and trainers to write, administer and report on assessments, tests, exams and surveys using PCs, LANs, the Internet and intranets



Product
Information

FREE
Demo CD

FREE
Web Seminar

FREE
Newsletter

Source: <http://www.questionmark.com/us/home.htm>

So you can see that virtually all the tools are readily available for you to deliver online electronic tests. However, this area is yet to develop itself for the numeric and other disciplines that use special notations for the answer. However this can be overcome partially by the use of Drag & Drop feature. Thus online evaluation helps you by offering interesting possibilities that are not available with conventional paper-based systems. First, by creating a computerized question bank, you can generate as many different

questions as you want at the same time and with equivalent standard. It can also save your precious time that is otherwise used in making the question paper again and again. Second, by making an instant evaluation through computer you can give immediate feedback by storing the answer also in the question bank. Third, Online/Web-based evaluation through its ever readiness can enable you to offer on-demand examination to the students, i.e. giving them the flexibility to take the exam as and when they feel their learning is complete. Lastly, you can record the results immediately, as it is easier to analyze the group responses and do test and item analysis to provide feedback for future examinations.

Self-assessment

1. List the disadvantages of using objective tests only in the online assessment.
2. Mention any other evaluation tools that can be used in online assessment.

Possibilities in subjective tests

Technology has made it possible to think beyond objective test. Restrictions in including only right/wrong kind questions in online evaluation systems are no more in place. There do exist some packages that pop up a hint when a student is unable to answer a question. Some packages also offer a feedback on what kind of error is committed and what should be the correct answer. However, these packages will take answer input as a word or a numeral only. Can you think of a situation in which the computer grades an essay as you do? As we have mentioned it earlier, a beginning has been made in this direction and time is not very far when you will be equipped with customized software to evaluate students in your very specific course through the computer. The constraints are none but one: The student has to key his own answer in the computer. In 1998, a Professor of Psychology at University of Colorado at Boulder completed work on the software after a hard effort of 10 years. The software shifted its focus on the content from writing and found out that these two are highly correlated. As a teacher you have to feed information into the computer from the textbooks and other sources. The computer in turn forms relationships between the words. As a teacher you are again expected to give computer a set of model answers that are graded by the human evaluators, which would eventually provide the computer a basis for comparison. It is not as simple as it looks but then once you are through you would be able to evaluate thousands of essays with the click of a button. The software also provides feedback on what the student left out in the answer besides giving a grade to the answer. We will discuss two of such approaches in detail to give you a feel of how these softwares work.

Intelligent Essay Assessor (IEA) is based on a statistical approach for analyzing the essays and content information. This statistical approach is called Latent Semantic Analysis, which is used primarily to extract semantic content of the text. Many scientists have proved that latent semantic analysis captures the similarity of meanings expressed. This technique is successful in giving judgements very close to the human judgements. According to the developers of the IEA, the semantic analysis focuses on the conceptual content of the essay, i.e. on the correctness and completeness of the content, the soundness of the arguments and the fluency and comprehensibility of the writing. This was a deviation from the earlier approaches, which were more mechanical in their approach of giving importance to grammar, punctuation and spelling.

You will get to know about Latent Semantic Analysis and other techniques later. This intelligent Essay Assessor has been used in a course of Psycholinguistics at New Mexico State University as a web based application. One can visit the website <http://psch.nmsu.edu/essay> for a demonstration of this package. This package basically permits the students of that course to log on the web and submit their essay to get a feedback on the missing points in their essays. After going through the feedback, students



Knowledge Analysis Technologies

A World Leader in Automated Essay Assessment and Text Analysis



Discover Magazine
touts K-A-T's
Intelligent Essay
Assessor as an
"ingenious process."
Click [here](#) to view a
summary of the
article and link to the
full text

Source: <http://www.knowledge-technologies.com/>

can revise their essays and resubmit the work. The main benefit of this process is that on one hand the software is performing the function of summative evaluation and on the other hand it is also doing the function of formative evaluation.

E-rater is another software worth mentioning in the computerized essay evaluators segment. The Educational Testing Service uses it for the Graduate Management Admission Test for testing the analytical writing assessments. E-rater is used as a second rater replacing the second human rater and thus reducing large costs on evaluation. This system has been developed after 5 years of rigorous efforts in using the advanced computational linguistics techniques. Apart from GMAT test, e-rater was also applied on Test of English as a Foreign Language (TOEFL) test. E-rater follows the scoring guide used by expert human evaluators for scoring. This scoring guide has a six-point scoring scale. It checks the essay for the argument structure, syntactic structure and vocabulary structure. The software is based on three general classes of features: syntactic, rhetorical and topical content features. The features are extracted from the essay texts and qualified using computational linguistics techniques.



Research

The CriterionSM Online Essay Evaluation Service, including *e-rater*[®] and the writing analysis tools and *e-rater*TM, are systems that automatically evaluate free responses. These applications were developed by researchers at Educational Testing Service. *E-rater* and the writing analysis tools are complementary automated essay evaluation applications within Criterion. *E-rater* provides a holistic score for an essay, and the writing analysis tools provide real-time feedback about grammar, usage, mechanics and style, and organization and development. *C-rater* is an application for automated analysis of conceptual information in short-answer, free responses. Publications describing these capabilities are below. Some of the publications may be downloaded as Adobe PDF files.

Source: <http://www.ets.org/research/erater.html>

Methodologies of essay evaluators

In the earlier sections, you have learnt about the two efforts in evaluating the essays through computer and thus adding a new dimension to web-based student assessment. Since the theories behind these methodologies are a new in the educational and a little statistical in nature, it is worth giving an overview of them. This will enable you to appreciate these efforts better.

- a) **Latent Semantic Analysis** It is a statistical model of word usage that permits comparisons of the semantic similarity between pieces of textual information. LSA assumes that there is a latent or concealed structure in the pattern of word usage in the essays and statistical can be used to eliminate this masked thing. The method generates a depiction of words that are used in similar contexts and are more semantically associated. LSA generates a matrix of occurrences of each word in each essay and then decomposes these matrices into a set of hundreds of factors. Since the number of factors will be much smaller than the number of unique words, words will not be independent. If two terms are used in same context, they will have similar factors. Matching is then done between two pieces of textual information, even if they have no words in common.
- b) **Syntactic Structure Analysis** Syntactic variety is an important feature in evaluating essays. It defines the ratio and quantity of types of sentences, types of clauses and use of verbs. This analysis parses through each sentence in the essay and quantifies these features. This parsing is done on the basis of Microsoft's Natural Language Processing tool.
- c) **Rhetorical Structure Analysis** If an examinee is able to demonstrate that the essay he developed, organized the ideas logically and has connected them well, then he ought to get a good score. This analysis quantifies the evidence of organization of the essay. It quantifies the cue words and other structures of the essay. Examples of cue words that are used to define a relation are given by the developers are: In summary, In conclusion, etc.
- d) **Topical Content Analysis** A good essay resembles other good essays in its use of vocabulary use of patterns. This analysis evaluates the topical content of an essay by comparing it with the other model essays that are graded by the human evaluators. It uses two different measures of content similarity. First, one is based on the vocabulary use in the essay as a whole and other is based on specific vocabulary content of the argument found in the essay.

Implications for education

The computerized essay evaluators have the following implications for education:

- **Reliability Regulator** Many teachers grade an essay and then use a computerized essay evaluator to ascertain the correlation. Only computer can offer unbiasedness, consistency and robustness.
- **Mass Evaluation tool** Computerized essay evaluators can be used for a large number of examinees with a less cost and time under the identical evaluation yardsticks.
- **Friendly On-Demand Tutor** Computerized essay evaluators can work as a friend philosopher and guide to the students by offering comments and their work and thus improving the quality of their writing. This can be of great help in distance education where the student is fairly distant from the teacher.
- **Pre-Testing Tool** Computerized essay evaluators can work as a great tool for the students to know beforehand about their preparations for the final term-end exam. Students can write and review their essays at the end of each chapter and then assess the areas on which they have to work more.

- **Second Rater** Computerized essay evaluators can work as a second rater to cross check the major differences, if any, with the human evaluators. This is especially useful in the high stake assessments such as in GMAT and TOEFL.

In short, computerized essay evaluator presents a good tool for assessing the quality of the essay in question. These can be applied in the classrooms and in distance learning for deriving respective benefits. The need at the moment is to generate the artificial intelligence in varied subjects.

Self-assessment

3. *Mention an area in which computerized essay evaluators cannot perform well. Give reasons for your choice.*
4. *Take a test on some students in your class. Get it evaluated by 4-5 evaluators. Try to find the correlation between them. What did you observe?*

Other tests suitable for online/web-based assessment

Having dealt with objective and subjective tests, you might be wondering as to what other tools are left to complete the discussion on web-based student assessment. Well, there are many which you might be using in your classroom environment. Technology has made it possible to adapt them to the web-based evaluation model. You will learn about direct presentation tools, electronic seminars and conferencing, authoring tasks and simulations in this session.

Distance education can be imparted wonderfully by making it collaborative and networked. It is here that the *Direct Presentation tools* come into picture. Along with the growing use of computers in institutions, it is not very difficult for the students to generate their work on a word-processor or as a presentation in Power Point. After completing the work a student can put it on the web or the local area network so that peers see it and add a comment to it. The teacher as a moderator can see all this work and offer comments on it. *Common Space* is a good software for this activity. It can be used by a group of individuals who offer comments on an individual's work. These comments can be integrated to revise the work and hence improve the quality.

Electronic Seminars and Conferencing is another area where computer and web-based technologies offer a tool for the assessment. You must have seen various discussion groups where people share their view through emails. A teacher can work as the moderator of the discussion group and offer comments to the group members (essentially students of that course). Conferences are also held electronically now a days. In this conference one submits his work to a group of established experts through email for consideration and the review process.

Authoring Tasks Simple authoring tasks can be given to the students, which leads them to the development of small packages. For example, if you are teaching a course on educational multimedia, then the best way to assess the deeper understanding of the student is by asking him a small multimedia package. Similarly, creation of a lecture-note, creation of a web page in HTML, etc. can give a practical touch to your lectures. This is discussed here because computers help in a big way in execution of these works and web helps in submitting them for an assessment.

Advantages of online/web-based student assessment

McCormack and Jones (1998) have highlighted many advantages of the online/web-based student assessment. Assessment created for online/web-based assessment do require a lot of time when established for the first time, but in the long run proves beneficial by

reducing the time, cost and efforts. They enable generation of a true feedback to the student instantaneously, which is otherwise impossible especially in the context of Distance Education, where it takes more time to evaluate assignments and providing feedback on them. A study of the turn-around time for the assignments showed that sometime it takes months for the assignments to reach back to the student. Web-based assessment eliminates the with-in and between biases of the human evaluators and thus provides consistency and reliability which is impossible for a human evaluator. It can help in automatically recording the results as and when an assessment is completed. Remember, this generates an error-free data, which is otherwise scarce, when you deal with several thousands of students at a time. If the assignments and their assessment are done through web, a lot of efforts can be saved in dispatching and collecting them through conventional methods like postal service. It also reduces the chances of assignments missing and delay in results.

Offline use of computers in student assessment

When we talk about Online/Web-based student assessment, we are not necessarily talking about the delivery through web or a local area network (LAN). You will be surprised that most of these software work in off-line mode also. The only difference is that delivery of such a test in offline mode requires a little bit of adjustment, e.g. in place of web or LAN, one can use postal or personal delivery of the CD-ROM containing the computerized question banks from which any number of question papers can be made through random number generation. IGOLEX has such a feature that when a student would enter his enrolment number and password based on some formula, the computerized question bank will generate a unique question paper for him. The use of question bank feature can also be seen with QUIZMAKER tool of EXAMINE. Also you would recall that QUESTIONMARK has a feature that enables one to deliver the assessment in an off-line mode that is still capable of storing the answer.

Apart from generating the test, computers can do many other things in the off-line mode, e.g. analysis and recording of the awards received by the students, recording the most prevalent errors committed by the students and the type of feedback given when a particular error is committed. These are some of the tasks a computer can do while in off-line mode. E-rater and IEA can also work in off-line mode. If this is so, a question that arises is, "what is the difference between off-line and on-line mode then?" One can answer it by saying that on-line mode extends the range of possible receptors of the assessment, say, if in on-line mode thousands of students can be assessed, in off-line mode you have to restrict it to a classroom or a college. The second difference is that in on-line mode you can load the system on one computer and run the show on many networked computers at the same point of time, but in off-line mode you have to load the system on computer on which you would like to carry out an assessment. Another difference is on the count of recording the work of students. In online mode, system generates automatic delivery of the responses of examinees but in off-line mode, one has to record them separately and then deliver it by alternative means.

There are many other differences, can you think and try to list a few of them?

Summary

We are still in the infancy stage when we talk of online/web-based student assessment. The progress in imparting on-line learning is phenomenal, but on the online assessment front a lot is still to be done. But the good thing is that tools are available and are in abundance. The need is the change in the attitude. Although it is said again and again that a university's reputation is built by the element of sanctity in its evaluation procedures, there has been a slow progress in adopting web-based assessment techniques, which can assure this.

In this Unit, you have broken the myth that online/web-based evaluation is nothing but objective-type multiple-choice questions. You have seen that virtually all the questions type which exists in conventional testing are capable of being handled through web. These systems on one hand enhances the learning opportunities for the students and on the other hand provides tools for the teacher to do much more rather than wasting time on redundant assessments.

Suggested readings

Burstein Jill, Karen Kukich, Lisa Braden-Harder, Martin Chodorow, Shuyi Hua, Bruce Kaplan, Chi Lu, James Nolan, Don rock and Susane Wolf (1998a), Computer Analysis of Essay for automated Score Prediction: A Prototype automated scoring system for GMAT Analytical Score Prediction. (RR-98-15). Princeton, NJ: Educational Testing Service.

N. Seshagiri (1993), Capes-Computer aided paperless examination system Part-I: Theoretical Basis, INFOTEX'93, Bangalore, India.

M.V.V.S. Reddy, N. Seshagiri (1993), Capes-Computer aided paperless examination system Prt-II: Field test results, INFOTEX'93, Bangalore India.

Foltz, P.W. (1996), Latent Semantic Analysis for the Text-based research. *Behaviour Research Methods, Instruments and Computers*. 28(2), 197-202.

McCormack and Jones (1998), *Building a web-based eduction system*, Wiley Computer Publishing, New York, pp. 236-7.

Answers to self-assessment

1. Some important limitations are:
 - a) They are unable to measure higher-level skills. They only measure factual knowledge.
 - b) At the initial stage the implementation could be time consuming and tedious.
 - c) Setting-up of the infrastructure for the conduct and Hardware/software reliability is highly essential.
 - d) Students who are not familiar with computers will be in a disadvantageous situation.
2. Some other interesting possibilities could be checklists e.g. to undertake a project work, a student ticks all the steps he/she has to go through for completion of the project. Image mapping e.g. to identity hardware components of a computer the examinee select an image which resembles with the component typically.
3. Computerized essay evaluators are constrained to work with subjects, which are mathematical in nature. It is also not suitable for evaluating numerical. The subjects in whom an answer can be answered in many different ways can also make it difficult to generate intelligence for the computer.
4. You must have observed that there could be a lot of variability between the scores of different evaluators. In fact if you yourself evaluate these answers at different points of time. you will still find variability.

Notes



