UNIT 3 CONSTRUCTION OF EVALUATION TOOLS

Structure

3.0 OBJECTIVES

- After going through this unit, you should be able to:
- identify the principles of test construction;
- explain the processes involved in test construction;
- describe and differentiate item formats; and
- list the quality of a test and identify the different approaches to use a test.

3.1 INTRODUCTION

Educational testing involves four stages of activity: (i) planning the test structure; (ii) constructing the test; (iii) administering the test; and (iv) assessing and interpreting the learners' performance. At the planning stage decisions have to be taken with regard to the objectives, choice of content area, choice of skills/abilities, the length/duration of the test, etc. At the construction stage the choice of item formats for the proposed test-points, the nature of sampling, the sequencing and grouping of items, drafting of instructions, etc. are to be decided upon. At the administration stage the main concern is to provide the appropriate condition, facilities, accessories, etc., uniformly to all learners who take the test.

At the stage of interpretation, conclusions with regard to the testees' performance, ability, their achievement in terms of the standards set, their relative standing in the group, etc., are to be arrived at.

These stages of activities are not independent of one another. But rather they are closely interdependent. Test construction is not only guided by a test plan but also governed by considerations of available conditions of administration. The test-plan conceives not only the areas of test-construction activities but also the norms by which the learner-performance is to be interpreted.

Beyond these four stages is the stage of test validation—a stage where we undertake the evaluation of a test. This activity serves two important purposes:

i) the purpose of immediate concern—ensuring whether the test procedure and the test are dependable in terms of the objectives and measurement;

ii) the purpose of long-term concern—envisaging and directing the reformative efforts that are required to improve the examination procedures (through consistent and systematic efforts over a period of years of feeding the findings of the evaluation of the work in one examination into the planning operations of the subsequent examinations).

There are different criteria to be taken into consideration while evaluating a test or determining the worth or quality of a test. Chief among these are provided by the concepts of 'validity' and 'reliability'. There are a host of other features to consider which pertain, in general, to the question of 'usability'. We shall discuss these in the following sections. But before we take them up for discussion, let us look into some of the general questions concerning evaluation of educational tests.

3.2 PRINCIPLES OF EVALUATION TOOL CONSTRUCTION

A test in order to be good must follow certain principles. Usually, test construction passes through four distinct phases i.e., planning, preparing, trial and evaluation. The test contains a carefully prepared and fixed set of items and procedures for administering and scoring. In other words, we can say that any test tool is first to be carefully planned, secondly items are to be prepared, then the test is to be tried out and lastly it must be evaluated from different angles before it is used.

3.2.1 Planning

It is the first step in test construction that whatever we do we plan it in advance, so that our act will be systematic. Obviously so many questions come to our mind before we prepare a test. What content area is to be covered by the test? What types of items are to be asked and what are the objectives that are going to be tested? Are the objectives that are going to be tested clear? etc. Suppose we want to prepare a good test in General Science for grade X. If most of the questions are asked from Physics and Chemistry and a few from Zoology, will it be a good test? Anybody would offer the criticism that the test is defective because it would fail to measure the achievement of pupils in Geology, Astronomy, Physiology, Hygiene and Botany. So it is desirable that all the content areas be duly represented in the test. Moreover, due Weightage should be given to different contents. Whenever we want to prepare a test, the weight to be given to different content areas must be decided at the beginning.

If we are preparing a test on Physics we must decide the weight to be given to different chapters taught in Physics. Thus, at the planning stage, it is the first task to decide the weight to be given to different content areas. Usually the weight is decided by taking expert advice. It must be noted that the Weightage must be in conformity with the amount of content taught under each content area. To draw up an objective list of such weight, the average of the available expert opinion may be taken.

Even though we give due weight to content areas, if all the questions (test items) aim at testing only the memory of the pupil, can it be called a good test? No. It must not only cover the area of knowledge (recall or memory comes under knowledge), but also other objectives like understanding, application, skill etc. A good test must aim at measuring all the behavioural areas. But can all the objectives be tested through one test? At this stage expert opinion may be sought as to which objectives are to be covered through our test. Usually in achievement tests four major objectives viz., knowledge, understanding, application and skill are tested. Now another question comes to our mind. What weights are to be given to different objectives chosen? Here again expert opinion can be taken.

Suppose that we are going to construct a test of Mathematics for students of grade IX. After due expert opinion we decide that the weight to be given to the different content areas viz. Arithmetic, Algebra, Mensuration and Geometry are 20%, 30%, 20% and 30% respectively and the weight to be given to the objectives of knowledge, understanding, application and skill are 40%, 30%, 20% and 10% respectively. After the weight to be given to the different content areas and different objectives is decided upon, a blue-print can be prepared. A blue-print for the proposed test on Mathematics is presented below.

Table 3.1: Blue-print (two dimensional chart)

Objectives	Knowledge	Understanding	Application	Skill	Total
Arithmetic	8	6	4	2	20
Algebra	12	9	6	3	30
Mensuration	8	6	4	2	20
Geometry	12	9	6	3	30
Total	40	30	20	10	100

The chart shown above is a content-behaviour chart. It is called the blue-print. Here, we have shown the weight to be given in two dimensions, viz., content and objectives. Thus, it can be called a ‘Two-dimensional chart’. The blue print is just a design or a plan of the test to be prepared.

Another thing comes to mind—whether we would ask ‘essay type’ or short-Answer type or ‘objective type questions’? If we decide to include only objective type questions the two dimensional chart will serve the purpose. If we want to include all the three types of questions, here again we have to seek expert opinion to decide the weight to be given to different forms of questions. Now the type of question (or form of questions) will be another dimension and as such a three-dimensional chart may be prepared.

Now another question arises. Should all the questions be easy or difficult or average standard? The usual practice is to include all the three categories. So a decision also has to be taken concerning the distribution of questions of different difficulty level. However, for a normal group the percentage of difficult, average and easy items to be included are 15%, 70% and 15% respectively.

We should also decide whether there would be provision for options or not. Whether we should have overall options (as, ‘answer any ten’) or internal options (each question has an alternative)? Provisions for options tend to lower the validity of questions. However, we may use internal options if the two questions are comparable in most respects (i.e., they test the same objective based on the same content, are equally difficult and would require same time to complete). So, whether or not options be introduced is to be decided at the planning stage.

Further we must decide the time within which an average student can answer the test. The test will have to be accordingly planned. Moreover, the total marks of the test is be decided and according to the weight fixed the marks are to be divided. The conditions under which the testing will be done should also be thought of in advance.

If a test is to be successful, a careful planning must precede its construction. From the foregoing discussions we feel that the planning of a test is not so easy. To sum up, planning of a test involves the following:

i) A detailed study of the text books, reference books, journals, test manuals, old questions, other reports, etc. is to be made.

ii) Weight to be given to different content areas is to be decided.

iii) Weight to be given to different objectives is to be decided.

iv) Weight to be given to different forms of questions is to be decided.

v) Whether or not provision for options is to be made.

vi) Weightage given to different categories of difficult level of questions is to be fixed.

vii) Total marks of the test along with the time required for its administration, the conditions of administration, etc. are to be planned in advance.

After all these considerations, a blue print of the test is to be prepared. It would not only give us a picture of the question of the test, but also serve as a guide for the preparation of the test.

3.2.2 Preparation

The second step in test construction is the preparation of the test itself. At this stage we have to prepare:

i) the test items

ii) the directions to test items

iii) the directions for administration

iv) the directions for scoring

v) a question wise analysis chart.

i) **Preparation of the test items**

Items must be prepared in conformity with the blue print. We have to choose appropriate items (test situations) which would test the specified objectives in the specific content area. Construction of test items is not so easy. It is the task of test-specialists and experts. An experienced teacher who is sufficiently trained in test-construction can prepare appropriate test items. There are certain rules and guidelines for construction of test items. Separate guidelines are there for construction of ‘essay type’, ‘short-answer type’ and ‘objective type’ tests. Even for construction of different types of objective-type tests, specific guidelines are prescribed. One must have access to all these guidelines and also access into the taxonomy of objectives before constructing test items. In general, the test items must be clear, comprehensive and free from ambiguity. They must be aimed at measuring the desired pupil-behaviour. They must fulfill their functions to ensure validity.

After the test items are framed they must be arranged properly and assembled into a test. If different forms of test items are being used, they should preferably be grouped form-wise. Moreover, easy items are to be given a place in the beginning, the difficult items at the end. The test items may be arranged in the order of difficulty. Of course, there are various ways of assembling the questions and we may assemble the questions according to our purpose and convenience of interpretation.

ii) **Preparation of directions to test items**

Appropriate directions to test items should be prepared. The directions must be clear and concise so that the students will understand them easily. The students should know as to whether he/she has to write the response or put a tick against the right response or to mark his/her response in some squares provided on the right side of the question or to mark his/her response on a separate answer sheet etc. Sometimes the directions to test items are so ambiguous that the students cannot follow them and as such he/she responds to the items in a manner which he/she thinks fit at that instant or simply passes on to the next item leaving it unanswered. Due to lack of clarity or directions students will respond differently at different times which would lower the reliability of the test. It is essential that the directions to the test items must be carefully prepared and they must be as clear and simple as possible. If necessary, full guidelines (even demonstration) for responding on item may be given.

iii) **Preparation of directions for administration**

A clear and detailed direction as to how the test is to be administered is to be provided. The conditions under which the test is to be administered, when the test is to be administered (whether in the middle of the session or at the end of the session etc.), within what time limit it is to be administered etc. are to be stated clearly. If the test has separate sections, time limits to cover each section must be mentioned. The materials required (if any) for the test such as graph papers, logarithm tables etc. must be mentioned. The directions must state clearly what precautions the administrator should take at the time of administration. So it is important that appropriate and clear directions for test-administration be prepared.

Preparation of direction for scoring

To facilitate objectivity in scoring, 'scoring keys' are to be provided. Scoring key is a prepared list of answers to a given set of objective-type questions. Suppose there are 10 multiple-choice objective type questions (each having four options, A, B, C, D) in a section of the test, the scoring-key will be as follows:

Section-I Scoring key

Q.N.	1	2	3	4	5	6	7	8	9	10
Key	D	C	C	A	B	D	B	A	C	B

A scoring key is prepared by listing serially the key (or right answer) to each question against each item.

For short answer type questions and essay type questions, marking schemes are to be prepared (i.e., marks allotted to different parts of the answer or to different important points etc. are to be mentioned). Such scoring keys and marking schemes must be carefully prepared. They serve as guides at the time of scoring the test and they ensure objectivity in scoring.

Moreover, it must be clearly stated as to how scoring is to be done. For example, if a strip-key or a window stencil is used, appropriate directions for using them are to be provided.

In certain cases corrections for guessing is necessary. Under the directions, it must be clearly stated for which type of items such corrections are to bee made. The formula used for 'correction for guessing' is given below:

$$S = R - \frac{W}{N-1}$$

where

S	=	the corrected score
R	=	No. of right responses
W	=	No. of wrong responses
N	=	Total no. of options

Thus, such specific directions for scoring as are likely to be necessary must be prepared. Of course, these may vary from test to test.

v) **Preparation of a question-wise analysis chart**

A question-wise analysis chart is given here. In this chart every question is analysed. This chart shows the content area (topic) the question covers, the objectives (with specifications) that it intends to measure, its type, the marks allotted to it, expected difficulty level and time taken to answer it. This chart not only analyses the items, but also gives us a picture of coverage of contents, objectives, type of questions and coverage of different difficulty levels etc. Moreover this gives us some idea about the total time to be taken for taking the test. This chart further helps us check whether the test has been prepared as per the blue print or not.

Table 3.2: Question-wise analysis chart

Topic	**Question**	**Objectives with specifications**	**type**	**Mark**
Solitary Reaper	1. Give a tick(✓) mark against the right answer			
(A)	Q. When the poet saw the girl. She was: a) sitting alone b) passing gently c) listening to the songs of the nightingale d) reaping grains and singing	Instructional objective (*Knowledge*) Specific/ behavioural objective (*recall*)	Multiple choice	01
(B)	Q. "The music in my heart I bore, long after it was heard no more." What was the effect of music on the poet later? (Answer in three sentences only.)	Understanding or *Comprehension* Specific or behavioural objective (*interprets)*	Short answer	03

Note: Preparation of such a chart is a necessity for teacher made tests; but for standardised tests we may or may not do it at the preparation stage. However, such analysis may be necessary at the time of editing the final form of the tests.

3.2.3 Try-out

The questions may be carefully constructed, but there is no guarantee that they will operate in the same manner as planned. So before the final form of the test is prepared it is necessary to have a try out.

For the trial of the items, a preliminary form of the test is generally prepared. This contains more number of items than are actually required for the final form. Usually the number of items included in the trial form should nearly be double of the number of items required for the final form. The lesson is that at the item-analysis stage many items will be discarded. A detailed scoring key of the trial form should therefore be prepared.

i) **Preliminary try-out:** After the test items, directions for response, administration and scoring are prepared it is tried out on a few 'sample' students just to ascertain how it works. At this stage 10 to 15 students of different abilities are selected and the test is administered. The aim of doing is to detect the omissions or mistakes if any, to examine whether the directions to items are actually being followed by students, to examine whether the time allowed is sufficient etc. Although a test is constructed with caution it may have some errors or ambiguity of directions here and there. The preliminary tryout will be to bring these to light. This helps us to modify or revise the items or directions whenever necessary. After due corrections the test is edited.

ii) **Final try-out:** At this stage the test is administered to a representative sample. The sample may not be too large. As it is just a pilot study a sample of 200 to 300 will do. But it must be borne in mind that this sample must be a representative sample of poor, average and brilliant students. The aim of such a tryout is to identify the defects and deficiencies of the test and to provide data for evaluating the test.

The purpose of the tryout can be summed up as follows:

a) to identify the defective or ambiguous items.

b) to discover the weaknesses in the mechanism of test administration.

c) to identify the non-functioning or implausible distracters in case of multiple choice tests.

d) to provide data for determining the discriminating value of items.

e) to determine the number of items to be included in the final form of the test.

f) to determine the time limit for the final form.

At the tryout stage the directions must be strictly followed. Conditions for test administration must be normal. The atmosphere should be calm and quiet. There should be proper seating arrangements, light, ventilation and water arrangements. Proper investigation and supervision must be ensured. A wrong administration of the test will give us wrong data for its evaluation.

iii) **Scoring:** After the try-out form of the test is administered the answer sheets are scored as per the scoring key and scoring directions. 'Corrections for guessing' are also done if required under scoring directions. Now the scores are ready for item analysis and evaluation of the test.

3.2.4 Evaluation

After scoring is complete, the test must be evaluated to examine whether the test items are good and whether the test is reliable and valid. For this purpose we:

i) analyse the items to examine their worth of inclusion in the test (Item-analysis);

ii) determine the validity of the test;

iii) determine the reliability of the test;

iv) assess the usability of the test.

i) **Item analysis**

Item analysis is a procedure by which we analyse the items to judge their suitability or unsuitability for inclusion in the test. As we know, the quality or merit of a test depends upon the individual items which constitute it. So only those items which suit our purpose are to be retained. Item analysis is an integral part of the reliability and validity of a test. The worth of an item is judged from three main angles viz.

a) Difficulty index of the item

b) Discriminating power of the item

c) Its internal consistency with the whole test.

a) ***Item difficulty***

When an item is too easy, all the students would answer it. If it is too hard, nobody would answer it. What is the use of having such items in a test? If all the students get equal scores, the very purpose of the test (i.e. to assess the ability of students) is defeated. So it is clear that too easy and too difficult items are to be totally discarded. It is desirable that items of a medium difficulty level must be included in a test. Item difficulty is calculated by different methods.

Method 1: Item difficulty (I.D.) is calculated by using the formula. ID = X 100 where R = no. of testees answering correctly, and N = Total no. of testees.

If in a test administered to 50 pupils an item is passed by (i.e. correctly marked by) 35 students the I.D = X 100 = 70

Here, we understand that the item is easy.

In essence, if the I.D. value is more, the item is easy and if the I.D. value is less than the item is considered to be difficult.

N.B. : Usually I.D. values in between 16 and 84 (or 15 to 85) are retained.

Method 2: Item variance and the difficulty level.

The proportion of passing an item is an index of difficulty. If 90% of a group pass an item, it is easy and when only 10% pass the item, it is difficult. If 'p' is the % of testees passing an item and 'q' is the % of testees failing in it.

S.D. = $\sqrt{Pq}$ or variance = pq

If p = .50, q is .50 and its variance .25

If p = .60, q is .40 and its variance .24

If p = .90, q is .10 and its variance .09

Items with more variance must be included in the test.

Method 3: Difficulty level of items can also be given in terms of standard deviation of the normal curve. For example, when 84% of pupils pass an item, it means that only 16% face the difficulty and its difficulty index in terms of S.D. of the normal curve will be - 1 . Other examples are given below.

Table 3.3: Item difficulty index

Item passed by	Difficulty index in terms of S.D. of the normal curve
16%	+ 1
84%	- 1
31%	+ .5
69%	-.5 etc.

Note: The table of areas under the *normal curve* may be referred to. Items with difficulty values in between ± 1 are usually retained

Method 5: The item analysis procedures used to obtain a reliable ranking of learners; indices of item difficulty and item discriminating power include the following:

i) Arrange the answer papers after scoring on the basis of merit (Highest mark at the top and lowest mark at the bottom).

ii) Select the 27% of the answer papers from the top and 27% of the answer papers from the bottom. The top 27% who have secured better marks constitute the higher group (H-group) and the bottom 27% who have secured poor marks constitute the lower group (L-group).

iii) Calculate W_H for each item i.e., for each item to determine the number of persons from the H-group who have wrongly answered an item or who have omitted it.

iv) Calculate W_L i.e., for each item calculates the number of persons in the L-group who have wrongly answered the item or omitted the item.

v) Calculate $W_H + W_L$

vi) I.D. = $\frac{W_H + W_1}{2n} \times 100$ where

n = number of persons in either lower group or higher group (n = 27% of N)

For multiple choice tests (where the options may be three or four) the following formula is used.

$$\text{I.D.} = \frac{W_H + W_1}{2n} \times \frac{100xoption}{option - 1}$$

Usually items in the range 16% to 84% of difficulty level are retained.

We can calculate the desired $W_H + W_L$ values from the following table.

Table 3.4: Calculation of item difficulty levels for multiple-choice questions

Difficulty level	$W_H + W_L$ values No. of options each item has			
	2	3	4	5
16%	.160n	.213n	.240n	.256n
84%	.840n	1.120n	1.260n	1.344

Suppose in a test consisting of all multiple-choice questions with 4 options and the number of persons in the H-group or the L-group is 120 (i.e. n = 120), what would be the $W_H + W_L$ values at 16%, 84% difficulty level?

Referring to the above table, $W_H + W_L$ value at 16% difficulty level = .240n = .24 x 120 = 28.8 = 29 (nearly).

$W_H + W_L$ value at 84% difficulty level = 1.260n = 1.260 × 120 = 151.2 or 151 (nearly).

Thus all items whose $W_H + W_L$ values are in between 29 and 151 are to be retained.

Any item whose $W_H + W_L$ value is less than 29 or more than 251 is rejected.

b) ***Discriminating index:*** To be considered good, an item must have discriminating power. For example, if an item is too easy or too difficult to all the testees, it can't discriminate between individuals. Logically, it is expected that a majority of students of a better standard and a few students of lower standard will answer an item correctly. Thus, an item must discriminate between persons of the high group and the low group. In other words,

W_L = Number of persons in the *lower group* (i.e. 27% of N) who have wrongly answer an item or omitted it.

W_H = Number of persons in the *higher group* who have wrongly answered an item or omitted it.

It is expected that W_L will be always more than W_H i.e., $W_L - W_H$ will always be positive. If W_H is more than W_L the item is either ambiguous??? and it is to be totally rejected.

We need to calculate the $W_L - W_H$ value for each item. Representative minimum $W_L - W_H$ values for an item with different options for different 'n' (27% of N) have been provided in Table 5.

Table 3.5: Representative minimum Values

Total no. Tested	**No. in Low or High group (.27 N)**	**$W_L - W_H$ at or above which an item can be considered sufficiently discriminating**			
		True or false/ Two options)	3 options	4 options	5 options
350-353	95	13	14	14	14
443-446	120	14	15	16	16
1110-1112	300	22	24	24	25

By referring to the table we can find that for a'n' of 120, the minimum W_L - W_H, value for an item with 4 options should be 16. So, all the items whose W_L-W_H. value is 16 or above are considered to be sufficiently discriminating. If W_L-W_H value of an item is less than 16, it is to be rejected-

c) ***Internal consistency of items with the whole test***

Statistical methods are used to determine the internal consistency of items. Biserial correlation gives the correlation of an item with its sub-test scores and with total test-scores. This is the process of establishing internal validity. There are also other methods of assessing internal consistency of items and as they are beyond the scope of our present purpose, we have not discussed **them** here.

3.2.5 Finalisation

After **item** analysis, only good items with appropriate difficulty level and with satisfactory discriminating power are retained and these items form the final test. Time required for the test is determined by taking the average time taken by three students who represent three groups: bright, average and below average. Now the test is administered to a large representative sample and the test-papers are scored.

3.3 ITEM ANALYSIS

In Sub-section 3.2.4 of this unit we discussed that an **item** analysis is a procedure by which we analyse the items to judge their suitability or usability for inclusion in the test. In this section we shall talk about the characteristics of a good item (i.e. a question).

The concern about the quality of an item becomes immediate when we attempt to develop a test or an item bank or when we attempt to evaluate a test being put to use. To determine how sound or good an item is we ought to know the features that go into its constitution and the qualities which contribute to its soundness.

3.3.1 Mechanics of an Item

An item or a question is an 'instrument' that we use to measure learning-outcome. One's learning is measurable by another only when it is demonstrated in observable behavioural patterns. An item, intended to be used as an instrument to measure learning, should make a learner 'act' or 'behave' or 'respond' so as to demonstrate his/her mastery (or the extent of mastery) or otherwise with regard to the select 'bit' of learning. 'The stimulus may be in the form of a task. The task may require the learner to do a 'descriptive' and/or a 'practical' activity. Descriptive activities may be oral or graphic (involving linguistic, semiotic and other features of communication). Practical activities may involve the use of some tools and materials and they may be performed in realistic or stimulated conditions.

3.3.2 Required Functional Conditions of an Item

An item or a question is primarily the specification of a task, the response to which is expected to put a desired bit of learning to demonstration. This involves two distinct activities on the part of the item-writer:

i) devising a task to meet the specific objective of the test, and

ii) specifying the task precisely and adequately.

The devising of the task has to be done carefully so as to ascertain that the performance of the given task requires the learner to display the desired quantum of knowledge of a chosen content or the ability to use a skill in a desired way. The specification of task also has to be done with great care. The specification may defeat its purpose:

i) when it is not adequate, and

ii) When is not 'communicated' clearly.

Let us elaborate these conditions further.

The specification will not be adequate if it does not point out the conditions under which the task is to be performed (say, for instance, tools, materials, guidelines, facts and information, etc. to be provided and the stages at which they are to be provided to the learner). It will not be adequate also when the level of accomplishment to which the task is to be performed is not mentioned in clear terms.

And even when an adequate specification is conceived, its purpose may not be served if is not 'conveyed' or 'presented' to the learner properly. This means that simple language or some graphic signs or gestures or some other mode of communication (sometimes, more than one of these at a time) has to be employed to make the specification clear to the learner. If the mode of communication employed to present the specification is beyond the comprehension of the learner, then the item, however, adequately conceived, may not be of any use.

To summarise, an item, to be an effective instrument of measurement of learning, should meet the following requirements adequately:

i) The task that an item specifies should, in the process of learner-response, demand and reflect only those specific aspects of skills or bits of learning that are being tested.

ii) It should specify precisely:

a) what the learner is to do,

b) the conditions under which it is to be done, and

c) to what level/standard it is to be accomplished.

The medium (linguistic, graphic semiotic, etc.) used to present the task specification should be such that there may not be any gap in its communication to the prospective testee (i.e. the learner should be able to follow the medium without any misunderstanding.

3.3.3 Checking the Functional Conditions of an Item

These general requirements that make an item sound are all concerned with qualitative features. To determine whether these requirements are satisfied by a question or an item, there are no objective measures available. One will have to depend on one's own experience and subjective assessment and the conclusions will be essentially empirical in nature. Collective effort, in this regard, therefore, may result in more reliable conclusions. A group of content experts, in collaboration with an evaluation expert, may do better than an individual. The conclusions may be still more reliable if a comprehensive check list of criteria is prepared in advance for each type of item in every subject area and used while validating a question or an item.

This act of checking whether a question or an item meets the qualitative requirements to make it an effective instrument is known as 'pre-validation in the process of question/item bank development.

Check Your Progress 1

Notes: *a) Space is given below for your answers.*
b) Check your answers with the ones given at the end of this unit.

Two English teachers, say X and Y, working under similar conditions frames the two following questions (I and II respectively for use in their classroom after an hour of instruction. Their intention is to check whether the students are able to use correctly the different forms of given vocabulary items in contexts relevant to them. Study the two questions carefully and say which of the two is better. Give reasons for your judgement. You can consider the two questions from the points of view of, a) task-objective relationship and b) clarity, precision and adequacy of task-specification.

1) Write one sentence each of your own using the different forms of the following two vocabulary items:

   i) to relieve

   ....

   ....

   ....

   ii) to blast

   ....

   ....

   ....

2) Following are two words from the material you have just read. Change the form of these verbs to fit into the sentences given. In some cases you will have to change them into nouns and adjectives. Fill in the blank spaces with appropriate forms.

i) to relieve

   a) He was (________) to hear that his wife had not boarded the ill-fated airbus which was hijacked.

   b) Aspirin is supposed to provide (________) for a headache.

   c) Some people do not like to take pills as pain (________)

ii) to blast

   a) The (________) of the dynamite was very loud.

   b) The glass was (________) out of the window by the explosion of dynamite.

   c) The wind was (________) all night long because of the hurricane.

3.3.4 Behavioural Characteristics of an Item

However, good the collective effort and however well composed the expert group attempting to determine the qualitative attributes of a question or an item, it can never be predicted with any certainty as to how a question/item will actually 'behave' or perform with a likely group of testees. What an amount of difficulty an item will cause, what sources of difficulties it will develop (or with what amount of ease it will be taken up and what sources of facilitative clues it will develop), what kinds of responses it will generate, are general questions which can never be precisely and adequately answered by any test-constructor or test-evaluator.

Observations on task - objective relevance and adequacy of specifications, though good in themselves, cannot offer much guidance to determine such behaviour traits of an item with a prospective group of testees. Only the testees that take up the question/item can be the arbiters in deciding these traits. They alone can provide these bits of information which we need to determine the eligibility of a question or an item to build up a sound achievement test. An achievement test, as we have noted earlier, intends to provide a basis for selection and grading of learners- As such, questions/ items making up an achievement test are expected

i) to be neither too difficult not too easy for the prospective testees, and

ii) to discriminate effectively 'the more able' (among the testees) from 'the less able'.

We may also need information about the behaviour characteristics of an item when we construct tests of special specifications - like tests of same **difficulty** level ('parallel tests') and tests of progressive difficulty levels ('graded tests').

3.3.5 Measuring Behavioural Characteristics

The procedure used to get information about the behaviours/characteristics of a question or an item is known as item-analysis. The information is obtained in quantitative terms and they are given in the form of numeral indices, unlike the information about the qualitative features of an item which are given in the form of empirical statements. Two indices are given after analysing the responses to a trial test paper administered to a representative group of the testees:

i) the measure of difficulty of each question, and

ii) a measure of the extent to which each question discriminates between the high scores and the low scorers on the same test.

The first measure is known as the Facility Value (FV) and the second the Discrimination Index (DI).

Different ways of calculating the two measures are available and some of them are sophisticated statistical operations which can be done only with the help of computers. The procedures given below for calculating the two measures may not give very accurate indices, but they do give satisfactory ones. The operations involved are simple and can be carried out even by those who do not have any acquaintance with statistics as a discipline.

Tabulation of scores: To facilitate the computations of facility and discrimination indices of the item comprising a test, we need to tabulate the scores properly. An illustration of score tabulation is given below (Table 2.6). The illustration may help you follow the steps given as under.

Step 1: Arrange the response sheets in order of value of the total score (Col. IV) in the test. Put the highest score on top and the lowest score at the bottom. (Response sheets of the same total score may be put one below the other.) Number the response sheets serially (Col. If).

Step 2: Divide the response sheets into three ability-groups (Col. I).

Higher Ability Group (HAG),

Middle Ability Group (MAG), and

Lower Ability Group (LAG).

If the strength of the sample (number of testees) is less than 40, the top 50% of the response sheets can be taken up to form HAG and the rest to form LAG. If the sample is between 40 and 100, then 27% or 10% respectively of the top and of the bottom scripts may be taken to form HAG and LAG. (If the percentage works out to a fraction, the fractions may be ignored and either of the groups may be allowed to be larger by one or two numbers).

FV for item 1

$$= \frac{\text{Total Score of the sample}}{\text{Total No. of candidates}}$$

$$= \frac{9+63+2}{11+88+11}\times 100$$

$$= \frac{74}{110} x100 = 67.27\%$$

FV for item 4

$$= \frac{\text{Total Score of the sample}}{\text{Total No. of candidates}}$$

$$= \frac{9+72=5}{11+88+11}\times 100$$

$$= \frac{86}{110}\times 100 = 78.18\%$$

DI for item 1

= Facility in respect of HAG

- Facility in respect of LAG

$$= \frac{9}{11}-\frac{2}{11}$$

$$= \frac{9-2}{11}$$

$$= \frac{7}{11} = 0.636$$

DI for item 4

= Facility in respect of HAG

- Facility in respect of LAG

$$= \frac{9}{11}-\frac{5}{11}$$

$$= \frac{9-5}{11}$$

$$= \frac{4}{11} = 0.363$$

Table 3.6: Illustration: tabulation of scores to facilitate FV and DI computations

Ability Group	Sl. No.	Roll No. in order of ranking	Total individual score	Item-wise score V							
I	II	III	IV	1	2	3	4	5	6	7	8
H	1	891	25	√	√	√	√	√	√	√	
I	2	702	24	√	√	x	√	√	√	√	
G	3	801	23	x	√	√	√	√	√	√	
H	4	705	23	√	√	√	√	√	√	√	
E	5	712	23	√	√	√	√	x	√	x	
R	6	813	22	√	√	x	√	√	x	x	
	7	811	22	x	√	√	x	√	√	√	
	8	737	21	√	x	√	√	x	x	x	
	9	785	21	√	√	√	x	x	x	√	
	10	721	20	√	x	√	√	√	√	x	
	11	850	20	√	√	x	√	x	x	x	
Total		11 Candidates		9	9	8	9	7	7	6	
M	12										
I	.										
D	.										
D	.										
L	.										
E	99										
Total		88 Candidates		63	66	49	72	47	15	41	
L	100	786	8	x	x	x	√	x	x	√	
O	101	809	8	x	x	√	√	x	x	x	
W	102	722	7	x	√	x	x	x	√	√	
E	103	826	7	√	x	√	x	√	x	√	
R	104	813	7	x	√	x	√	x	√	√	
	105	851	7	x	x	x	√	x	x	√	
	106	870	6	x	√	x	x	x	x	x	
	107	764	6	√	x	x	x	x	x	√	
	108	783	6	x	x	x	√	x	x	x	
	109	822	5	x	x	x	x	x	x	x	
	110	847	5	x	x	x	x	x	x	√	
Total		11 Candidates		2	3	2	5	1	2	7	

Step 3: Draw a table of vertical columns and horizontal rows. Enter the serial number of the Response Sheets which you put while carrying out Step 1, one below the other in the first column. (The corresponding Roll No. of candidates can be given in the second column, if necessary.) Leave a gap of three rows each below the HAG, the MAG and the LAG.

Step 4: Enter item-wise scores in the horizontal row against each candidate (Col. V). When the item-wise scores of all candidates are entered, the total score of the sample on each item could be calculated by adding up scores in the vertical columns and the total score of each individual on the test could be calculated by adding scores along the horizontal row.

Determining the facility value: Facility value is generally presented **as a percentage.** In the case of an objective type item, it is calculated as the number of learners answering the item correctly divided by the number of learners attempting it. The fraction is multiplied by 100 to get the figure in percentage. In the case of a supply-type question, facility value is the average mark obtained by the sample on the question divided by the maximum mark allotted for the question. Here too the fraction is converted into a percentage figure.

To summarise, FV of an objective item =

$$\frac{\text{No. of learners answering the item correctly}}{\text{No. of learners taking the test}} \times 100$$

FV of a free-response question =

$$\frac{\text{Average score obtained by the sample of the question}}{\text{Max. mark allotted for the question}} \times 100$$

The facility value ranges from 0% to 100%. FV represents the fact that none of the sample has answered the item correctly and hence the item has no ‘facility’ whatsoever for the given sample. 100% FV represents the fact that everyone in the sample has answered the item satisfactorily and the item has no difficulty whatsoever with the given sample.

Determining the discrimination index: Discrimination index of an item is arrived at by deducting the facility value of the LAG **(**in the item from the facility value of the HAG on the same item, The DI is always presented in the form of decimal fraction and it may range from - 1.0 to + 1. .

In the case of an objective type item:

DI = FV of the item with HAG - FV of the item with LAG

$$= \frac{\text{No. of testees answering the item correctly in HAG}}{\text{No. of testees in HAG}}$$

$$= \frac{\text{No. of testees answering the item correctly in LAG}}{\text{No. of testees in LAG}}$$

(We should mention here that multiplication by 100 which is normally required **to** present FV’s in percentage is avoided here, because Dl’s are to be given in decimal fraction, not in percentage.)

To calculate DI in the case of a free-response item, you have to find out the mean score (MS) i.e., average of the scores on the item in the HAG (Let us represent this as MS-HAG) and the mean score on the same item in the LAG (Let us represent this as MS-LAG). The difference between the two mean scores (obtain by deducting the LAG mean score from the HAG mean score) divided by the maximum marks allotted for the item gives the D1 of the item.

$$\text{DI of an item} = \frac{\text{(MS-HAG) - (MS + LAG)}}{\text{Max. score fot the item}}$$

3.3.6 Interpreting Behavioural Characteristics

FV and DI can be helpful to us in determining:

- the quality of an item at the tryout stage in the development of a test or a question/item bank; and
- the quality of teaching/learning at the stage of actual **use** of a test.

What different values of facility and discrimination can signify with regard to the quality of an **item** in the context of item tryout are tabulated in tables 7 and 8:

Table 3.7: FV in the context of item try-out: an interpretation

FV ranging			
From	**to**	**would mean that ….**	**would require……**
0%	25%	the item **is** too hard	modification of the item (probably checking of dstracters)
25%	75%	the item is within the suitable range of facility	retention of the item
75%	100%	the item is too easy	rejection of the item (perhaps, checking of clues sometimes can help improve items and retain).

Table 3.8: DI in the context of item try-out: an interpretation

DI ranging			
from	**to**	**may mean that ...**	**may require**
- 1,00	+ 0.20	the upper sample is not doing better than the lower	suitable modification or rejection of item
+ 0.20	+ 1.00	the sample item makes satisfactory discrimination	the item is to be retained.

How different values of FV and DI of objective type items can be interpreted in relation to the quality of teaching/learning is presented in tables 9 and 10:

Table 3.9: Interpretation of FV in the context of actual test-use

FV ranging		
from	**to**	**would mean that**
0%	25%	the topic has not been taught/learnt well and that the teaching techniques are to be reviewed to ascertain what has gone wrong.
25%	75%	the topic has been taught/learnt reasonably well.
75%	100%	the learners have gained an exceptionally good knowledge of the topic.

Table 3.10: Interpretation of DI in the context of actual test-use

DI ranging		
from	to	would mean that
-1.00	- 0.25	the weaker students have a better grasp of the topic than the 'good' students.
- 0.25	+ 0.25	all the students have an equal grasp of the topic.
+ 0.25	+ 1.00	there is too great a gap between the lower ability and the higher ability groups (perhaps, some remedial work for lower learners must be planned).

FV and DI of subjective items: Interpretation of FV and DI in the case of subjective items is slightly complicated. 'Md FV in the case of an essay-type or short answer type question need not be the index of the facility of the question alone. Since subjective assessment is involved in scoring and since it is the scores that we take as the data, the FV in such cases may not be reliable. In the absence of precise guidelines for scoring, the FV in these cases may reflect the lenient/severe attitude of the scorer, or perhaps the scorer's attitude and the item's facility (if the scorer's attitude remains constant).

The interpretation of DI is also difficult in the case of free-response tests. In such cases the number of questions that make up the test are fewer than the number of questions that make up an objective type test. Consequently the weight that each question carries in a subjective test is relatively high and each question contributes in considerably large chunks to make up the total score. As a result the correlation between the score on a question and the score on the test turns out to be undependable. A DI in such a case is likely to be spuriously high. The 'satisfactory' level of DI in such cases, therefore, has to be higher. It should be, say, more than + 0.50.

Check Your Progress 2

Notes: *a) You can work out your answer in the space given below.*
b) Compare your answer with those given at the end of this unit.

i) Distinguish between 'facility value' and 'discrimination index' of an item.

....

....

....

ii) What should be the range of 'facility value' and 'discrimination index' for an item to be retained in a test?

....

....

....

3.3.7 Use of Behavioural Indices

The facility and discrimination indices will be helpful to both test constructors and teachers. To test constructors they will help identify faulty items which they can modify and use, or reject. They will also help them in developing effective multiple choice items by providing feedback on the working of the 'key' and the 'distracters'. They could also help test constructors to develop test tools designed in order of progressive difficulty with items of weak discrimination followed by items of gradually increasing discrimination in favour of better students. The facility value obtained from different groups of learners taking the same test can provide teachers with a basis for comparison and can also help in defining and maintaining 'standards'. The discrimination indices locate topics to be addressed to all the learners and topics to which learners of lower ability are to be restricted.

3.4 GUIDELINES FOR THE USE OF AN EVALUATION TOOL

Most learning is a complex mix of physical and psychological activities. The proportion of the mix varies from one area of learning to another. But, invariably, in all areas of learning, the involved psychological processes make precise assessment of accurate learning. For the same reason, the evaluation of an educational test cannot be very accurate and will mostly be a subjective assessment.

3.4.1 Quality of a test: Some Focal Points

If we are to evaluate an educational test, with all its inherent complexities disallowing total precision and objectivity, how are we to go about it? Though the assessment is subjective, it can be regulated and guided to a certain extent by a careful consideration of the following features of a test-content. (Table3.11).

Table 3.11: Characteristics expected of a test-content

Feature of test content	**Aspects to check**
Presentation of item numbering sequencing, wording, punctuation	Clarity in expression and specificity in task requirement
Presentation of general and specific instructions—wording, punctuation, etc.	Clarity in expression and precision in the description of requirements
Sampling of items	Appropriacy in terms of coverage and allocation of due weight
Choice of item-format	Suitability to the chosen test-point
Choice of the form of measurement (written test, oral test, field work, laboratory work	Suitability to the general objective of the test
Scheme of evaluation (given to examiners)	Maintenance of concurrence with the set objective of the items and the test

These features are the focal operational points with regard to which even any little amount of slackening of care might impair an educational test meeting its stated or intended objective. The chief attributes of a good test—validity, reliability and usability—should be verified with these features to ascertain the quality of a given test. We shall now take up each one of these attributes and discuss them in greater detail.

3.4.2 Validity of Tests

What is validity?

The concept of the validity of a test is primarily a concern for the 'basic honesty' of the test 'honesty' in the sense of 'doing' what an item promises to do. It is a concern for the relationship , on the one hand, between the purpose to be achieved and on the other hand, between the efforts taken, the means employed and what those efforts and means actually achieve. There is always a gap, whatever be the size, between the purpose of a test and the extent of realisation of the purpose in practice. Hence absolute validity is ideal in educational testing. Perfection in terms of validity—a perfect match between the purpose and practice—is hard to achieve. This is due to:

i) the nature of 'learning', which is the subject of measurement, and

ii) the nature of each of a number of factors that become involved in the measurement learning.

(These factors, we have discussed briefly in Section 3.4 above).

The less variant that a test turns out in practice from its stated purpose, the more valid it is. Hence validity is a measure of the degree of success with which a test accomplishes what it sets out to accomplish. It is an attempt to answer how close a test is in its operation to the purpose in its plan-design. To be precise, a test is valid to the extent to which it measures what it purports to measure.

Types of validity:

`Purpose' and 'practice' then, are the two dimensions of considerations involved in the concept of validity. 'Practice' is conditioned mainly by three operant forces—the test, the testee and the examiner. If the demands made by the test, the performance offered by the testee and the valuation (of the testees' performance) done by the examiner are all directed to be symmetrical with the given/set purpose of the test, then validity is ensured. Thus the test purpose is the constant point of reference for validity. Consequently, several types of validity are conceived of to suit the specific purposes that tests are designed to serve. We shall take up four types of them for our discussion. They are:

i) Content validity

ii) Criterion-related validity

   a) Concurrent validity, and

   b) Predictive validity

iii) Construct validity

iv) Face validity

Content validity

Content validity is the most important criterion for the usefulness of a test, especially of an achievement test. It is a measure of the match between the content of a test and the content of the 'teaching' that preceded it. The measure is represented subjectively after a careful process of inspection comparing the content of the test with the objective of the course of instruction, The key aspect in content validity is that of **sampling.** Every achievement test has a content area and an ability range specified for its operation. Given the limited human endurance in taking a test (say three hours at a stretch) and hence a limited test-duration, no single test can ever make **a** total representation of any considerable length of a content area. A test, therefore, is always a sample of many questions that can be asked. It is a concern of content-validity to determine whether the sample is representative of the larger universe it is supposed to represent.

A table of specifications with a careful allocation of weight to different units of the content area and the several abilities, keeping in view the set objectives of the course and the relative significance of each of these, can help a test-constructor as a road-map in the construction of items and the development of a test. A careful scrutiny of the table of specification (if any has been used) and the loyalty with which it has been adhered to while developing the test can help you assess the adequacy and appropriateness of sampling. Where such a table of satisfaction is not available, you may have to develop a 'concept-mapping' of the content area to check the sampling of content represented by a test.

We should note here that a test that is content valid for one purpose may be completely inappropriate for another. We should also note that the assessment of content-validity has to be subjective basically as it depends on the assessor's estimate of the degree of correspondence between what is taught (or what should be taught) and what is tested it requires a careful examination of the stated objectives of the course in terms of course content and target abilities and a study of the size and depth of realisation of their coverage. Such examinations lead to 'estimates' and not to 'measurements'. That is; the observations of such examinations tend to be subjective statements. They cannot be expressed in terms of objective numerical indices.

Check Your Progress 3

Notes: *a) You can work out your answer in the space given below.*
b) Check your answer with the one given at the end of this unit.

We have said that an achievement test of high content validity cannot be a content valid test for diagnostic purposes. Why?

....

....

....

....

....

....

....

....

....

....

Criterion-related validity

Unlike content validity, criterion-related validity can be objectively measured and declared in terms of numerical indices. The concept of criterion-related validity focuses on a set ‘external’ criterion as its yardstick of measurement. The ‘external’ criterion may be a data of ‘concurrent’ information or of a future performance.

The ‘concurrent’ criterion is provided by a data-base of learner-performance obtained on a test, whose validity has been pre-established. ‘Concurrent’ here implies the following characteristics;

i) the two tests—the one whose validity is being examined and the one with proven validity (which is taken as the criterion)—are supposed to cover the same content area at a given level and the same objectives;

ii) the population for both the tests remains the same and the two tests are administered in an apparently similar environment; and

iii) the performance data on both the tests are obtainable almost simultaneously (which is not possible in the case of ‘predictive’ criterion).

The ‘predictive’ criterion is provided by the performance-data of the group obtained on a course/career subsequent to the test which is administered to the group and whose validity is under scrutiny.

The validity of a given test is established when ‘concurrent’ criterion correlates highly (i.e. agrees closely) with its own data.

Validity established by correlation with ‘concurrent criterion’ yields **concurrent validity** and similarly validity established against the scale of ‘predictive’ criterion is called **predictive validity**. The former resolves the validity of tests serving the purpose of measuring proficiency; the latter resolves the validity of tests meant for predictive function. The ‘concurrent’ criterion has been widely used in the validation of psychological tests, especially tests of intelligence. The general practice is that one or two standardised tests of intelligence with proven quality are used to validate the item. It is crucial in all selection and placement tests. For example, when a Banking Recruitment Board selects candidates for the post of clerks on the basis of a clerical aptitude-cum-intelligence test, the selection will be purposeful only if a high correlation is established between the test results of the candidates and their performance ability, subsequently, in the clerical position. The higher the predictive validity, the more emphatic this assertion will be. In all cases of criterion-related validity, an index of the degree of correspondence between the tests being examined can be obtained. This index of agreement is known as correlation coefficient in the statistical parlance.

Construct validity

The word ‘construct’ means the ideas developed in one’s mind to define, identify or explain objects/phenomena. Let us suppose that a person is interested in the study of intelligence. He/she hypothesises that the third or fourth generation learners will have a higher IQ than the first generation learners. On the basis of his/her observations he/she may build a theory specifying the degree of difference in the IQ of the two groups of learners.

If a test is constructed, then, to measure the difference in the levels of intelligence of first generation learners and third/fourth generation learners, the test would be considered to have construct validity to the extent that its scores correspond to judgements made from the observations derived by the scorer about the intelligence of the two groups of learners. If the expected level of difference is not established by the test scores, then the construct validity of the assumption that the test measures the difference in the levels of intelligence is not supported. Thus, a test will be described to have construct validity if its scores vary in ways suggested by the theory underlying the construct. In other words construct validity is the degree to which one can infer certain constructs in a psychological theory from the test score.

Construct validity is an important concept to those who are engaged in theoretical research on various constructs.

Face validity

Before we take up a detailed scrutiny of a test for any of the above validity-types, we generally tend to make an impressionistic assessment so as to develop some propositions which may guide our approach to the assessment of validity. Such propositions are developed on a facial understanding of the extent to which a test looks like a valid test or the extent to which the test seems logically related to what is being tested. These propositions constitute what is known as face validity.

Face validity may not be dependable. A test may look right without being rational or even useful. For instance, a terminal examination in a course of 10 units may appear to have reasonable face validity until you come to realise that it contains questions on the first five units only and therefore lacks content validity. Sometimes there may be situations where a test may appear to have low face validity, but in practice it may turn out to be a sound one. In such cases, the testees too may not know what is being tested and ipso facto the measure may provide far more effective assessment. For instance, the ability to react quickly to a flash of light may be a good test of potential as a football player. Such a test of reaction time may have content validity even if it doesn't have much face validity.

(Incidentally the 'idea' that prompts using this reaction time test for identifying a potential football player is a construct. The idea or the construct perhaps may be explained as that one who is able to react speedily and correctly to the sudden darts of an object can be a potential football player).

Check Your Progress 4

Notes: *a) Write your answer in the space given below.*
b) Check your answer with the one given at the end of this unit.

We talked about four types of validity-content validity, criterion-related validity, and face validity. Given below are three situations. Identify which relates to what type of validity.

i) A test given at the end of a semester of a degree course to measure, how much the learners have achieved of the given course-

....

ii) An entrance test to an engineering course to select suitable candidates.

....

iii) A test designed to be parallel in structure and content to another test of proven worth.

....

3.4.3 Reliability of Tests

The discussion of the concept of validity relates to the question of what to test. The concept of reliability which we shall discuss in this section relates to the question of accuracy' with which the 'what' is measured.

If a student were to take the same test twice, the logical expectation would be for the student to get more or less the same score both the times. But this does not happen practically on most occasions, Differences in scores do occur and they are likely with every repetition of the test.

The difference may be due to many reasons:

i) the characteristic (say, intelligence) which is being measured may change across time (we can call this 'trait instability');

ii) the particular questions that a test-constructor chooses to be representative of the quantum of knowledge which the test has to deal with may affect the score (we can call this 'sampling inconsistency);

iii) any variation in the communication of instruction or in test-timing or in the rapport with the administrator could lead to score variability. Let us call this 'administrator inconsistency');

iv) lack of objectivity in terms of one or more of the following could also affect the score:

   a) the item (it may not have the same significance to all the students or to a student, every time it is repeated),

   b) the response which the item permits (it may not be limited in number and may permit various levels of adequacy), and

   c) the scoring method used (it may leave room for free play or subjective judgement of the scores),

We can see (a) and (h) as constituting 'item inconsistency' and (c) as 'scoring inconsistency' and

v) personal characteristics of an individual like fluctuations in memory, effort, attention, fatigue, emotional strain and similar factors could cause score variability (Let **us** call this 'human inconsistency'),

The concept of error score and the concept of reliability: The sources of variation that we have talked about are called 'sources of error' and the variation in a person's score is called 'error variance'. The assumption here is that all scores given to learners are affected positively or negatively by one or more of the factors mentioned above. The 'true score' reflecting a learner's ability is hypothetical and is never determinable as the sources of error are not totally extricable from any measurement of learning. But the 'true score' is supposed to be constant (assuming the 'learning' that it reflects is stable). The 'error score' is not constant as the error-sources themselves are variable.

Thus, the 'true score' being supposed to be constant and the 'error score' inconstant, the combination of the two which we get in the form of scores on test-papers (i.e. 'observed score') varies in proportion to the difference in the 'error score'. That is, whatever variation we find in the 'observed score' of a learner on different administrations of the same test, are due to errors in assessment brought about by the variables operating on different occasions—the 'true score' of the learner remaining the same all through.

You should note here that among errors two types are possible: 'systematic error' and random error'. When a weight bridge consistently weights 50 kg. less of a lorry-load repeatedly, the error is constant and we call it systematic. We refer to those errors as random errors which do not remain the same on every occasion of measurement.

Some points of further clarifications: Before we proceed further to the methods of determining reliability it is worth observing here some points of further clarification:

i) Reliability refers to the results obtained with an evaluation instrument and not to the instrument itself. An instrument may have a number of different reliabilities depending on the groups of subjects and situations of use. Hence it is more appropriate to speak of the reliability of 'test scores' or of 'the measurement' than of 'the test' or of 'the instrument'.

ii) Test scores are not reliable in general. An estimate of reliability always refers to a particular type of consistency—say, consistency of scores over a period of time ('stability') or consistency of scores over different samples of question ('equivalence') or consistency of scores across scoring on different occasions ('scorer reliability') and the like. The scores of a given test may be consistent in one of the above respects and not in another. The appropriate types of consistency in a given case are dictated by the use to be made of the results.

iii) Reliability is a necessary but not a sufficient condition for validity. While low reliability can restrict the degree of validity that is obtained, high reliability provides no assurance for a satisfactory degree of validity. (A balance that always weighs 10 gms. in excess of the true weight of an object may be highly consistent in recording the same weight for the same object every time it is weighed. Nevertheless measurement suffers in respect of validity in as much it fails to give the real weight of the object).

iv) Reliability is primarily statistical in nature; it may be expressed in terms of shifts in relative standing (in respect of scores) of persons in the group or in terms of the amount of variation to be expected in a specific individual's score. In the former case it is reported by means of a correlation coefficient called a 'reliability efficient' and in the latter case it is reported by means of the 'standard error of measurement'.

Check Your Progress 5

Notes : *a) Indicate your answers in the boxes given against each item*
b) Check your answers with those given at the end of this unit.

i) From the sources given below identify those that are responsible for measurement.

a) the purpose of a test may not the same for all situations □

b) the scoring standards of test-responses may vary from person to person □

c) the representation of content may not be the same for any two tests on the same subject area □

d) the same group of testees are not always available for two consecutive test-administrations □

e) the situations of administration may not be the same for any two tests

f) an item may not give the same idea (about the task to be performed) to different testees □

g*)* a testee may not be equally alert and active on two occasions.

ii) What do you call the score given by an examiner on the cover-page of an answer book?

a) True score □

b) Observed score □

c) Error score □

d) Random score □

Estimates of reliability

The methods used to measure reliability differ according to the source of error under consideration. The most common approaches to estimates of reliability are:

i) Measures of stability,

ii) Measures of equivalence,

iii) Measures of stability and equivalence,

iv) Measures of internal consistency, and

v) Scorer reliability.

Measures of stability: Measures of stability are known as 'test-retest estimates of reliability'. They are obtained by administering a test twice to the same group with a considerable time-interval between the two administrations and correlating the two sets of scores thus obtained.

In this type of estimate we do not get specific information as to which of the sources of error contribute(s) to the variance in the score. It gives only the measure of consistency, over a stretch of time of a person's performances on the test.

The estimate of reliability in the case will vary according to the length of time-interval allowed between the two administrations. The intervening period can be relatively long, if the test is designed to measure relative stable traits and the testees are not subject enduring the period between the two tests administrations to experiences which tend to affect the characteristic being measured. The intervening time should be shorter when the conditions are not satisfied. But it should not be so short as to allow 'memory' or practice effects' to inflate the relationship between the two performances.

Measure of equivalence: In contrast to the test-retest estimate of reliability which measures change in performance from one time to another the estimate of reliability with equivalent forms of tests measures changes due to the specificity of knowledge within a domain. Instead of repeating the same test twice with an intervening time-gap, the latter procedure administers two forms ('parallel' in terms of content and difficulty) of a test to the same group on the same day (i.e. with negligible time-gap) and correlates the two sets of scores obtained thereon.

The two methods of estimating reliability are quite different and can yield different results. The choice between the two depends on the purpose for which you administer the test. If your purpose is long-term prediction about the reliability of the test, you can choose the procedure of retest reliability estimation. If your purpose, on the other hand, is to infer one's knowledge in a subject matter area, you will have to depend on equivalent forms of estimate of reliability.

Measures of stability and equivalence: When one is concerned with both long-range prediction and inferences to the domain of knowledge, one should obtain measures of both equivalence and stability. This could be done by administering two similar (parallel) forms of a test with considerable time-gap between the two administrations. The correlation between the two sets of scores thus obtained by the same group of individuals will give the coefficient of stability and equivalence. The estimate of reliability thus obtained will be generally lower than the one obtained in either of the two other procedures.

Measures of internal consistency: The three methods discussed above are concerned with consistency between two sets of scores obtained on two different test administrations. The methods that we are to discuss, hereafter collectively called 'measures of internal consistency, arrive at reliability estimate taking into consideration the scores obtained on a single test-administration. 'The estimate of reliability obtained through these methods is mostly indices of homogeneity of items in the test, or of the extent of overlap between the responses to an item and the total test score. The three types of measures of internal consistency are discussed below.

Split-half estimates**:** Theoretically the split-half method of estimating reliability is the same as the equivalent forms methods. Yet the split-half method requires only one test administration but while scoring the items, a sub score for each of the two halves of the test is obtained and the two sub scores are correlated to get the reliability estimate of half the length of the test. To estimate the reliability of the scores on the full length test, the following formula is used:

$$\text{Reliability on full test} = \frac{2 \times \text{reliability on 1/2 test}}{1 + \text{reliability on 1/2 test}}$$

The application of this formula assumes that the variances of the two halves are equal. That is to say that the items in one half are supposed to match in respect of content and difficulty with the corresponding items in the other. The question then is how the tests can be split into two halves. Different methods are followed, but ordinarily it is done by a preconceived plan (say, assigning the odd numbered items to one half and the even numbered items to the other) without obvious statistical measures to make them equivalent.

Kuder-Richardson estimates: This method of estimating the reliability of test scores from a single administration of a single form of a test by means of formulae KR 20 and KR 21 was developed by Kuder and Richardson. With the help of these two formulae we can estimate whether the items in the test are homogeneous, that is, whether each test item measures the same quality or characteristics as every other, In other words, these formulae provide a measure of *internal consistency* but do not require splitting the test in half for scoring purposes.

The formulae are:

$$\text{KR 20} = \frac{\text{n}}{\text{n-1}}\left(1\text{-}\frac{\sum pq}{\sigma t^2}\right)$$

Where n = number of items in the test,

$\sigma\tau$ = standard deviation of the test scores,

P = Proportion of the group answering item *correctly,*

q = 1–P = proportion of the group answering a test item *incorrectly.*

We use KR 20, we have to

1) Compute the standard deviation of that rest (i.e., σt)

2) Compute p and q for each item,

3) Multiply P and q to obtain the value of pq for each item,

4) Add the value of all the items to get ∑pq

5) Using the formula KR 20 we can calculate the reliability of a test.

The computation of KR 20 is more extensive but accurate. A less accurate but simpler formula to compute the reliability of a test was proposed by Kuder and Richardson Known as KR 21.

$$KR\,21 = \frac{n\sigma t^2 - (Mn - M)}{\sigma t2}$$

Where σt^2 = standard deviation of the test scores

n = number of test items in the test

and M = the mean of the test scores.

Crombach alpha: Kuder-Richardson estimates are possible when the scoring of items is dichotomous. When the scoring is not dichotomous as in a test consisting of essay questions, the formula developed by Cronbach can be used to get the reliability estimate. This formula known as Cronbach alpha is the same as KR 20 except for the fact that is replaced by, where is the variance of a single item. The formula is:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sigma si^2}{\sigma st^2}\right]$$

Where n = number of items in a test

σsi^2 = variance of a single item

σst^2 = variance of total number of items in a test

[*'Variance'* is a statistical term. It measures how much the individual scores of a group of learners vary from the average score of the group. It is the mean of the squared deviation of the scores from their mean. Please refer to block 3 of ES-315 for more detail about variance]

The measure actually compares the variance for any single item with the variance for the entire test. It is, therefore, suggested that there should be at least five questions in the test to make this measure meaningful. Having given you some idea about the three types of measures of internal consistency, we shall now talk about yet another variety of reliability.

Inter-scorer-reliability: The question of estimating inter-scorer reliability does not arise in the context of objective tests. But where the scoring is subjective, it is necessary to determine the likely error component in scores which may be there due to the person(s) that scored the performance, and establish objectivity in evaluation. In determining inter-scorer-reliability, the same procedure of correlating two or more steps of scores as done in test-retest or equivalent forms estimate is followed.

This would give the reliability estimate of a single reader (i.e. the scorer who reads through the response). If we want to know the reliability of the sum or average of scores of two or more readers we could use the Spearman -Brown prophecy formula.

$$r_{xx} = \frac{nr}{1 + (n-1)r}$$

Where r_{xx} = reliability coefficient of a test

r = reliability estimate of a single reader

n = number of readers

Check Your Progress 6

Notes: *a) Space is given below for your answer.*

b) Check your answer with the one given at the end of this unit.

Write down the three types of measures of internal consistency and mention in what situation each of these are used.

....

....

....

....

....

....

....

....

Let us now sum-up the procedures of the different methods of estimating reliability in a tabular form.

Table 3.12: Different methods of estimating reliability

Methods of estimating reliability	**Types of reliability measure**	**Procedure**
1) Test-retest method	Measure of stability	Give the same test twice to same group with a considerable time-gap between the two administrations.
2) Equivalent forms method	Measure of equivalence	Given two forms of a test of the same group in succession.
3) Retest method using equivalence forms	Measure of stability and equivalence	Given two forms of a test to the same group with between the two.
4) Split-half method	Measure of internal consistency	Administer a test once. Get sub-scores for items of two equivalent halves of the test. Use Spearman-Brown formula to obtain reliability estimate for the whole test.
5) Kuder-Richardson method	Measure of internal consistency	Administer a test (of objective type) once. Score the test and apply Kuder-Richardson formula.
6) Cronbach alpha method	Measure of internal consistency	Administer a test (of subjective type) once. Score the test and apply Cronbach alpha formula.

7) Multiple scorer method	Measure of scorer reliability	Administer a test once. Let it be scored by two or more scorers. Correlate the sets of scores to measure the reliability of the scores of one scorer. Apply Spearman-Brown Prohecy formula to obtain the reliability of the sum (or average) of the scores of two or more scores.

Comparison of methods

As noted earlier each type of reliability measure represents different source(s). A summary of this information is given in Table 9. Note that more sources of error are represented by measures of equivalence and stability than by any other type of measure. Naturally, reliability estimates obtained on measures of equivalence and stability are likely to be lower. This should caution you to take into account the type of measures used to report reliability estimate, especially when you attempt to choose a test from among standardised tests guided by reliability estimates.

The 'X' mark in Table 13 indicates the sources of error represented by the reliability measures.

Table 13: Representation of sources of error

		Types of Reliability Measures			
Sources of error	**Stability Scorer reliability**	**Equivalence**	**Equivalence & stability**	**Internal consistency**	
Trait instability	X		X		
Sampling error		X	X	X	
Administrator error	X	X		X	
Random error within the test	X	X	X	X	X
Scoring error					X

3.4.4 Usability

We have discussed in detail the two chief criteria of test-validation— validity and reliability. What remains to be seen are the considerations of 'Usability'. 'Usability' mostly raises questions of feasibility with regard to test-construction, administration, evaluation, interpretation and pedagogical application.

While judging feasibility we should remember that the tests are usually administered and interpreted by teachers without the desirable amount of training in the procedures of measurement. Time available for testing and the cost of testing also deserve attention.

Besides these, attributes like the case of administration, which has little possibilities for error in giving directions, timing, etc., case and economy of scoring without sacrificing accuracy, ease of interpretation and application so as to contribute to intellectual educational decisions, are factors pertinent to the usability of tests.

3.5 LET US SUM UP

- We began this unit with a discussion of the different stages of test tool construction. There are five principles for constructing a tool of evaluation.
- We followed it with a discussion on item analysis that makes it complex to measure an educational achievement and to judge the soundness of such a measurement. Then to facilitate the judgement of an educational measurement, we identified some features of the item analysis and the important qualities to verify each one of them.
- We focused upon the two characteristics of a good tool namely, validity and reliability. We discussed the different approaches within each, the contexts in which each of these approaches becomes relevant and the procedures by which the evidence of validity and reliability of a test is to be established. Finally we referred to the aspects of a test that you should check to ascertain its usability.

3.6 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress 1

Y’, question (i.e., II) is better than X’s Reasons

a) Task-objective relationship

Question 1 of X demands that the learners write sentences of their own. This shifts the focus of the question from identifying the appropriate form of the vocabulary items to a more complex skill of constructing sentences. Besides, the learner also has to ‘invent’ contexts to suit the different forms of the given vocabulary item.

Question is of Y is better because it restricts the learner—task to the set objective of using the appropriate form—by providing the contexts and fully structured sentences.

b) Clarity, precision and adequacy of test specification

Question I is vague, because it does not specify how many different forms are to be attempted. It does not also restrict us as to which types of forms are to be used.

The infinitive forms (like to be relieved, to have relieved, to have been relieved, etc.) or the *participal forms* (like relieving, relieved, being relieved, having been relieved, etc.) the noun forms (relief and reliever).

There is no specification about the nature (simple/complex, etc.) of the sentences to be given and the kinds of context to be presented.

A sentence may be at different levels of complexity. Similarly, a context may be presented with different depths of inventive or imaginative intensity. There is, of course, no easy and accurate way of putting down the restrictions within which sentences and contexts are to be conceived and presented.

The format of question II avoids these problems of task specification, while still being functionally relevant to the objective. The elaborate instruction and the appropriate format fix the 'frame' within which the learner is to act to satisfy the set objectives of the teacher.

Check Your Progress 2

i) Facility value (Fv) is a measure of the 'difficulty' of a given question, whereas discrimination Index (D1) in a measure of the extent to which a question discriminates between the more able and the less able learners.

ii) An item can be retained in a test if the range of its Facility Value (FV) is from 25% to 75% and the range of its DI is from +0.20 to +1.00-

Check Your Progress 3

Achievement tests and diagnostic tests differ in their purpose and also in their treatment of the learning-content they are supposed to deal with.

An achievement test may enjoy a high content validity if it makes an adequately representative sampling of the learning content that it is concerned with (since its purpose is to measure the extent of an achievement, not to identity the specific lapses).

But for a diagnostic test, if meant for identifying the specific lapses in learning on the part of individual learners, to be credited with content representation of the learning-content is possible.

Check Your Progress 4

i) Content validity

ii) Predictive validity

iii) Concurrent validity

Check Your Progress 5

i) b, c, e, f and g.

ii) b

Check Your Progress 6

The three measures of internal consistency are:

i) Split-half estimates

ii) Kuder-Richardson estimates, and

iii) Cronbach alpha.

Split-half estimates are used when we want to know the extent of equivalence in 'content and difficulty' of the two halves of a test. Kuder-Richardson estimates are used when the items of a test are homogeneous and do not require splitting the test in half for scoring purposes. Cronbach alpha measure compares the variance for any single item with the variance for the entire test.