11. Tests, testing and evaluation

'Testing to teach or teaching to test?'

11.1 The need for evaluation

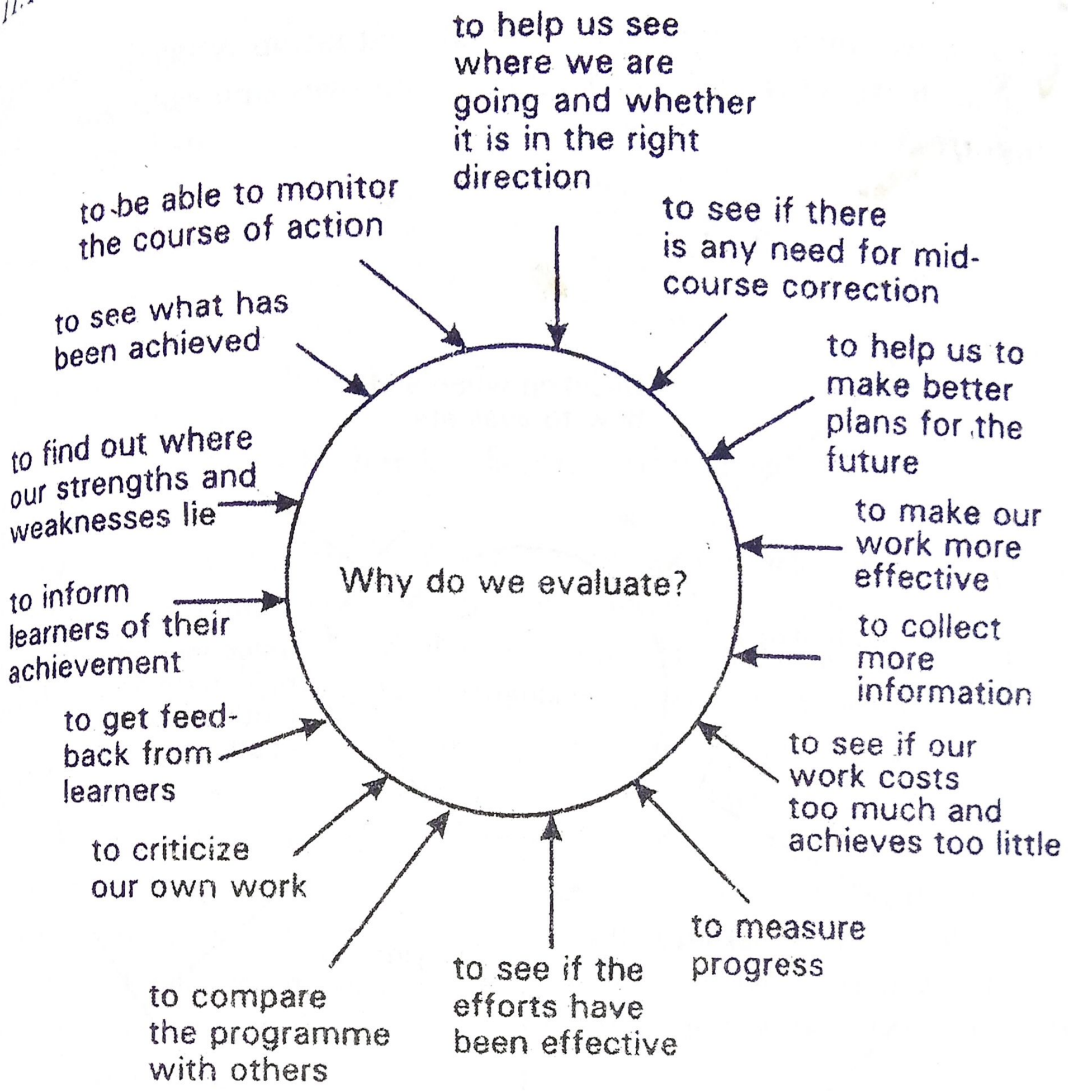
Testing and evaluating of language skills and competencies are very important components of language teaching; testing becomes an integral part of teaching because it provides significant information or inputs about the growth and achievement of learners' difficulties, styles of learning, anxiety levels, and such other feedback necessary for planning a programme and making mid-course corrections. Effective teaching and effective testing are two sides of the same coin.

A *curriculum* is what constitutes a total teaching-learning programme composed of overall aims, syllabuses, materials, methods and testing- in short it provides a framework of knowledge and capabilities, selected to be appropriate to a particular level. The *syllabus* provides a statement of purpose, means and standards against which one can check the effectiveness of the programme and the progress made by learners. *Tests* evaluate not only the progress and achievement of learners but also the effectiveness of the teaching materials and methods used; even tests can be evaluated to see if they do what they are supposed to do. It is useful to make some distinction between evaluation, and tests and examinations.

Evaluation is a systematic gathering of information for purposes of making decisions. It is the collection, analysis and interpretation of information about any aspect of a programme of education and training as part of a recognized process of judging its effectiveness, its efficiency, and any other outcomes it may have. Evaluation is defined as 'an attempt to understand what is going on to judge its worth and make decisions about it.'

Evaluation is an activity that is not restricted to only education; we do it every day, in many contexts and levels, formally and informally. We evaluate the food we eat; we evaluate and pass judgement on everything around us—maybe, a party that we attended, a lecture or a TV programme or the dress that someone is wearing! We constantly evaluate the people around us—whether someone is honest, reliable, bossy, cunning, etc. In a way, evaluation is a natural tendency!

In addition, teachers must also know why they are evaluating what they are evaluating! They are not just evaluating learners' progress and the programmes and materials; they are also evaluating the effectiveness of their teaching. The following chart is intended to show why we evaluate.

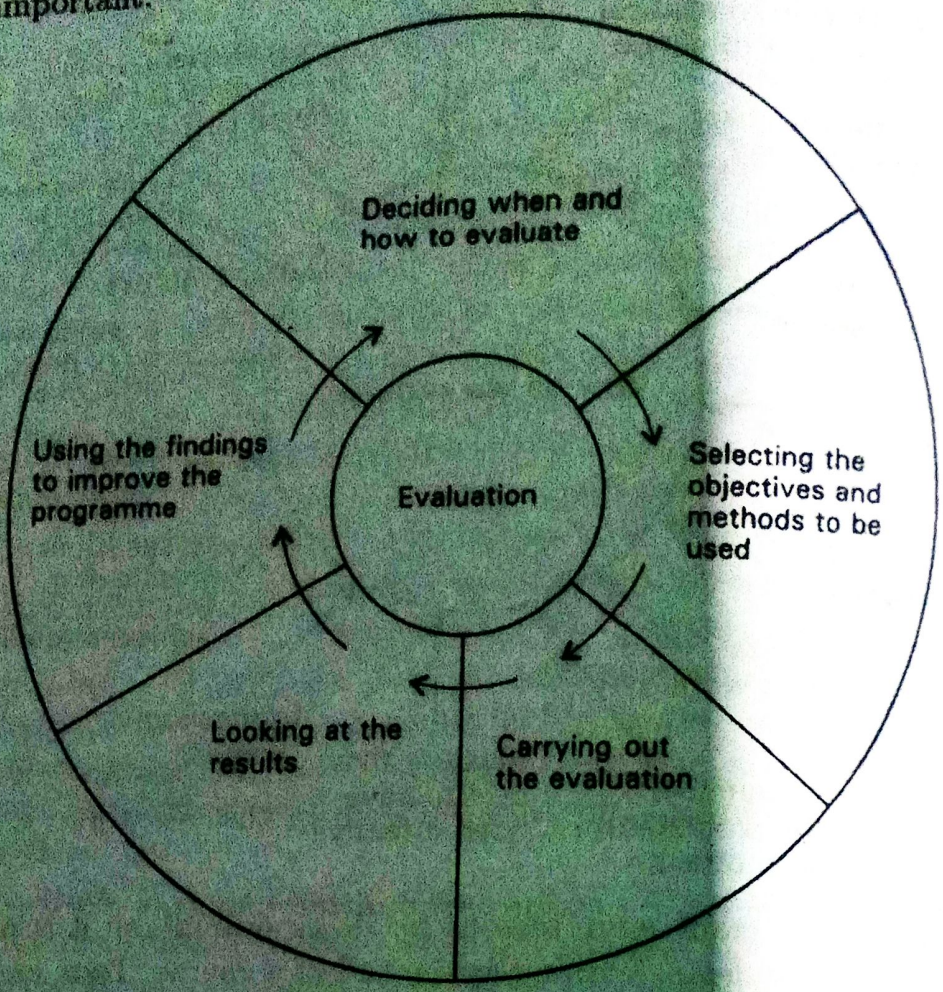


One can add many more reasons for doing evaluation; it is not just for measuring students' learning at the end of a course. Tests are useful in diagnosis, prediction, selection, grading, guidance, self-correction, etc.

There is internal as well as external evaluation in most educational institutions; both are useful to arrive at an overall picture! Each one has its own advantages and disadvantages. Some are given below:

	Internal evaluation	**External evaluation**
1.	Knows the programme and the examinees well	Can take a fresh look at the programme
2.	Can't be objective	Can be objective
3.	Part of the power structure and internal politics	Not a part of the power structure
4.	May not be trained in evaluation procedures	May be trained in evaluation
5.	Can interpret with familiarity	May not be familiar with the programme and the people involved
6.	Not much fear or anxiety; so less threatening	May cause anxiety to the participants

Finally, we must understand that all evaluation is like a wheel. All parts must fit together so that the process moves smoothly. All aspects are important.



Similarly, in education and language teaching, the terms 'evaluation', 'assessment', and testing are used for measurement and judgement. Generally, we use the term 'examination' for measurements that are more serious like the end of the year/semester/term tests; the term *test* is used for periodical measurements that are intended to show that what has been taught is learnt. *Testing* refers to procedures that are based on tests. *Measurement* is a broader term that includes many other factors like attendance records, internal assessment done by teachers, other extra-curricular and co-curricular activities; measurement is a quantitative description described in terms of numbers; evaluation is still broader because it includes quantitative as well as qualitative description expressed in words; for example, qualitative information may be available in interviews, casual meetings, group discussions, diaries, or even impressions that are subjective (e.g. extraordinary, poor) These distinctions may be somewhat hazy and blurred, but they are useful in the field of education.

11.2 Types of tests

Language tests may be broadly classified into two types:

Testing skills	Testing knowledge of content
Skills such as listening, speaking, reading, writing and sub-skills such as comprehension, vocabulary, grammar, spelling, punctuation, etc.	Knowledge of literary forms, literary techniques, literary history, social history, culture, etc.

Generally, at the higher levels knowledge is tested, and at the school level skills are tested.

There are various kinds of tests; tests can be ***subjective*** or ***objective.*** Measuring language competence in everyday situations, in interviews, in speech-making or essay-writing, etc. is subjective. Multiple-choice tests, matching exercises, true/false items are examples of objective tests. Subjective procedures are more impressionistic, even if the evaluation is done by trained professionals; there can be variations in the evaluation even by experts in the field. For example, an essay or a precis can be marked by three or four trained people; variations are bound to be found in their evaluation. Objective tests can be marked by machines with an answer-key that gives the right answer. Objective tests are not based on impressions and, generally, for each item in the test there is only one correct answer.

There are tests that measure ***productive abilities*** like speaking and writing and ***receptive abilities*** like listening and reading. Tests can also be classified according to the mode or medium: *oral* tests and *written* tests; there can be *bimodal* tests like dictation or note-taking (i.e. writing after listening).

Tests can be based on language ***components*** such as pronunciation, spelling and punctuation, grammar, composition, vocabulary, etc. The term '*discrete point*' refers to anything isolated—maybe, one skill at a time or at one level (phonology, grammar, etc.) or one bit (sentences, word, etc.) at a time. ***Discrete point tests*** emphasize isolated bits of language. ***Integrative tests emphasize*** the use of all the skills and components and test communication skills; they evaluate the ability to communicate effectively in speech as well as writing and they have become important in areas like information technology and management.

Tests can be ***direct*** or ***indirect.*** A direct or performance-based test measures one's communicative competence in almost real-life situations. An indirect test does it in an indirect way like testing one's reading

or speaking ability. Direct tests are difficult to administer; that is why, indirect tests are generally used.

Non-referencing tests measure an examinee's performance with reference to the performance of others in a group, like gaining grades; *criterion-referencing* tests measure the examinee's progress towards a specific objective regardless of what others do; for example, specifying a minimum score like 35% or 40% for passing in a paper is *criterion-referencing.*

Tests are also classified as follows: (a) aptitude tests (b) proficiency tests (c) achievement tests (d) diagnostic tests.

Aptitude tests give us guidelines to show if learners have the ability to learn a given subject. Aptitude tests in areas like music, painting, scientific reasoning, mathematical ability, etc. can be constructed to see if one has a tendency towards fine arts, science, language learning, etc. They are difficult to construct and their validity and reliability are to be carefully evaluated since they provide the inputs for deciding one's future course of action in life. Aptitude tests are not much in use in India but, in countries like the USA, they are widely used.

Proficiency tests measure the ability, for example, in English, to undertake a particular programme or course. For instance, placement tests are proficiency tests; they are, in a way, diagnostic, in the sense they tell us whether a learner is ready to undertake a course at a particular level, and the candidate's areas of strengths and weaknesses. TOEFL (Testing of English as a Foreign Language) or IELTS (International English Language Testing System) are proficiency tests. Some institutions stream students on the basis of proficiency tests and decide if a student can be placed in Stream A (i.e. above average), Stream B (i.e. average), or Stream C (i.e. below average). In a way, tests for admission to courses, jobs, etc. are of this type.

Achievement tests, on the other hand, indicate the level of achievement or the level of progress towards a specified goal. Most examinations and tests conducted in schools and colleges are achievement tests.

Diagnostic tests are useful in planning remedial activities because they give the necessary inputs about the nature of the problem.

The categories mentioned above are not watertight compartments. Teachers, generally, evaluate students in a number of ways—homework assignments, in-class exercises (oral as well as written), team activities, weekly tests, projects, class participation, etc. In some institutions there is internal assessment that carries weight up to 40 to 50 per cent of the total marks. The final examination is based on the candidate's written performance.

11.3 Characteristics of a test

In order to develop a proper test, one must be familiar with the important aspects of test-construction. Given below are some of the aspects of a test:

i. the purpose of a test (i.e. What is it supposed to do?)
ii. a knowledge of the examinees and their background
iii. the accuracy of measurement (i.e. its validity)
iv. the suitability of the format and the familiarity of the examinees with the test format (i.e. multiple-choice, cloze, etc.)
v. the nature of scoring and reporting
vi. test economy (i.e. cost effectiveness)
vii. test acceptability (i.e. suitable to the demands of the society, the institution and individuals)

There are three important characteristics of a good test; they are ***validity, reliability,*** and ***feasibility.***

Validity means the test actually measures what it is intended to measure; *reliability* shows consistency and that the test will function the same way each time it is administered with different examinees; *feasibility* includes several factors like economy (not time-consuming and not expensive with respect to development, administration, and scoring), availability (i.e. easily available), acceptability (i.e. culturally and socially acceptable), relevance, interpretability and applicability.

A reliable test must go through the following stages:

- Identify the population (i.e. the target group)
- Identify the objective or purpose
- Prepare a pool of items
- Do a pilot try-out
- Do an item analysis
- Check for the Difficulty Value (DV)
- Check the Discriminatory Power (DP)
- Administer the test
- Give the test to a large population
- Standardize the test

On the basis of DP and DV, items are retained or deleted after the trial run.

DV: If the number of items in the test is 100, those items that are very easy and very difficult are to be deleted. If, for example, 92% and above get the answer right for a given item, that is considered too easy; similarly, if 30% and below get the correct answer for an item, it must be difficult. So, only items for which the scoring is between 31% and 91% can be included in the test pool.

DP: The discriminatory power of an item is calculated on the basis of the top 25% and bottom 25% of the scores for an item. If an item does not discriminate properly, it is to be dropped.

Similarly, improper wording can cause a test to become unreliable. Instructions are an essential part of any test; instructions must be clear, precise and grammatically correct. In writing instructions, one must be concise; it is better to check them over as if the one who constructs the text is the candidate to be tested; only then can the difficulties and problems be seen. Check the spelling. Instructions and sub-instructions are to be spaced out properly to avoid any resulting confusion.

It is also very essential to create a proper 'climate' for testing. We can reduce the fear and anxiety of the students towards tests and examinations, by developing a positive attitude; this can be done by conducting more informal tests and short quizzes; teachers too can reduce the excessive evaluation work by using informal testing techniques. Informal oral evaluation, open-book tests, announcing a battery of questions in advance, familiarizing students in objective type questions, using short assignments at frequent intervals and other such techniques can reduce the tension of the learners as well as the teachers and promote a more healthy attitude towards tests.

Tests can be used to enhance learning; they are effective teaching devices: By giving advance notice of the area to be covered in a test, teachers can virtually ensure that the learning is done before hand. That is why we can say that most of the teaching activities can be considered testing instruments and most tests can be used for teaching.

11.4 The question-bank

Making examinations more and more transparent is one way of reducing examination-tension or what is popularly called 'exam fever'; test-tension is one of the main factors that affect a candidate's performance in the test or examination.

Some institutions use the 'question-bank' idea; they are different ways of using the 'bank or pool' idea! The Union Public Service Commission

asks professionals in different areas to contribute different test-items to their question-bank; these items are carefully scrutinized, edited, classified according to the level of difficulty and type (i.e. subjective, objective, etc.), and kept in the bank. Question paper-setters are invited and asked to select items from the question bank for a particular examination conducted by the UPSC. Close monitoring helps the process of paper-setting.

Another possible way is to have 'open' question-bank examinations. In a country like India where there is enormous student enrollment (like the CBSE examinations, Intermediate and Higher Secondary examinations in some States), it is possible to invite questions from the teachers of various subjects (like English, Mathematics, Chemistry, etc.); these question-banks (i.e. booklets) can be published by the CBSE, Intermediate Board, etc. The booklet may have about 5000 or 6000 questions of various types—multiple-choice, true/false, short answers, etc.

The agency concerned (CBSE, Intermediate Board, etc.) can price these publications and sell them to the students. Since the number is so enormous, the agencies can run the 'question-banks' on commercial basis. Students can be told that questions will be asked only from the question-banks. If students can prepare answers for about 5000 or 6000 questions in a particular subject, say English, the results will be very effective. At the time of question paper setting, the random selection method can be used and about 50 or 60 questions selected. (If computers can be used for this purpose, random selection can be done in no time.) This way of selecting questions will eliminate the costly process of setting up two or three sets of question-papers and all the secrecy that goes with it. At some stage, candidates can be allowed to bring the question-bank booklet to the examination centre and only the numbers of the questions communicated to them with a proper networking of the computers. This practice will be cost-effective, transparent and commercially self-financing for the agencies concerned.

Even individual educational institutions can set up their own banks and use it for their own teaching and testing. At the regional level universities and states can effectively use the question-bank concept.

11.5 Question types

Question papers that are easiest to set are the most difficult to grade and vice versa. Essay-type questions that can be made up in a few minutes require more time to grade; objective type questions, which take more time to construct, can be corrected at a greater speed. Short-answer questions fall somewhere between these two extremes.

Questions are often selected mostly in terms of class size, using objective-type tests for large classes or examinations involving a large number of students; the number of students to be examined is an important factor in selecting question types but the aims and objectives of the syllabus and the educational goals are as important, if not more, as the number of students to be tested. This implies that different types of questions, problems, and questions requiring analysis, integration and application, and long-answer questions will have to find a place in a balanced question paper. A balanced test design must be properly contextualized, integrative, interactive and pragmatic.

There are several types of questions.

(a) **'Limited response'** format where the response is limited like *yes/no* questions or the *true/false* format or asking the learner to do only one thing at a time.

(b) The **'Multiple-choice'** format (with four suggested responses)— Care must be taken in setting multiple-choice questions. *Distracters* are the incorrect options but they are to be carefully selected; they should not be too easy or too vague. For example, in

Of the two toys, the child chose_____.

(a) the less expensive

(b) the one most expensive

(c) the least expensive

(d) the most expensive of them

there are very good distracters for the correct answer (a). The other items are likely to distract the one taking the test. Similarly, in a vocabulary test, for example, in

The little boy had a long day; he was feeling *drowsy*

for the word 'drowsy', *lazy, sad, exhausted* are good distracters but not *stupid*.

Multiple-choice items are easy to score but difficult to set.

(c) Simple completion format like filling in blanks or completing incomplete sentences.

(d) Multiple-choice completion: Example:
Sahitya has been waiting here _____ half an hour.
(i) during (ii) for (iii) while (iv) since

(e) Rearranging mixed items, maybe words, sentences or even paragraphs

(f) The matching format: Matching items in Column A with the ones in Column B; matching halves (called 'photo-fit')

(g) Cloze procedures: A text in which every nth word is deleted and the testee tries to restore the text by filling in the words, using his/her linguistic and textual knowledge and the knowledge of the world.

(h) Editing tests like precis writing, removing errors or changing the style or the mode of narration, etc.

(i) Dictation: Different variations of the traditional dictation that was used only as a spelling test.

(j) Short-answer questions

(k) Essay-type questions.

(l) Translation

(m) Information-transfer format: Texts are to be read carefully and the information to be arranged in tables/chart, etc. or the reverse of it. (i.e. from tables/charts into paragraphs)

(n) Communicative testing: Actual use of language in real-life settings which are based on the needs of the learners.

11.6 Testing the four skills

In testing, as in teaching, it is useful to think in terms of the four skills though in practice they are combined and integrated. Ease and convenience in constructing and scoring have retained skill-based testing.

I. Testing receptive skills

A. Testing reading comprehension and vocabulary

Most textbooks have comprehension questions after a reading passage and it is taken for granted that reading comprehension is the easiest to evaluate. However, as it was pointed out earlier (See Ch.8), comprehension takes place at various levels involving different operations. Comprehension can be *literal*, *analytical*, and *inferential*. Analytical comprehension involves classification, categorization, and selection; inferential comprehension, on the other hand, requires reasoning abilities which are more complex. Comprehension may be referential, based on the ability to recover factual information. It can be *local* or *global* showing the ability to perceive the overall organization of a given text.

At the higher levels, evaluation or critical comprehension may be included since it will test the ability to judge the author's statements, attributes, emotions, etc.

When a passage is taken from a textbook, which has already been used for intensive reading in the class, the teacher will be testing only recollection and not comprehension. Seen passages do not provide a test of reading comprehension; unseen passages are better for testing reading comprehension because the memory element is eliminated.

Selecting texts: Texts for comprehension should be from authentic sources from standard newspapers, magazines, books, etc. The texts selected should conform to the types of texts included in the syllabus. It is not good to select texts simply because they are readily available or easy to read. A relatively difficult or complex text should be shorter than a more straightforward one. The length of the text may vary from 50 to 100 words at the elementary level, 200 to 300 words at the intermediate level, and 400 to 600 words at the advanced level. Passages for scanning may be longer; detailed reading can be tested using shorter texts. So, we should be clear about the objectives of testing reading comprehension. Texts dealing with a single idea or one theme are not suitable; it is better to select passages with a series of events, a collection of facts, or different options and attitudes. In addition, the selected passage should cover the structures and words included in the syllabus.

Instructions and questions for testing comprehension should be in simple English so that students understand them. It is better to use a variety of questions and not questions that can be answered by copying some words from the passage. Questions can be set according to the length and nature of the text selected. Some examples are given below.

Sentence Comprehension: Multiple-choice paraphrases of even simple signs can be useful.

Smoking is injurious to health.

It means (a) Smoking is prohibited. (b) Smoking will harm the smoker's health. (c) It is better not to smoke. (d) Smoking will cause injury.

Passage comprehension: Rearranging exercises, yes/no type questions, true/false questions, multiple-choice exercises, wh-questions, inferential questions, simple completion tests, combining sentences, joining split sentences, fill-in-the blank questions, information-transfer exercises, questions on points of view, etc. can be asked. At a higher level questions that test the emotions and feelings expressed, their implications and suggestions, questions on the organization of the text or on the author's intentions, etc. can be asked. (See also 9.2)

Gapped texts (i.e. cloze passages) can also be used for testing reading comprehension; students cannot supply the missing words if they do not understand what they read. The passage can be about a page in

length with about twenty or twenty-five deletions; the first and the last sentences are always given in full without deletions. The deletion of words can be regular, random or by deliberate choice.

In 'pure' cloze texts every fifth or seventh word is deleted. *Multiple-choice cloze* tests give a choice of three or four words and the student has to select one that best fits the gap.

Example:

'Master' he cried, 'As I was—(sleeping, crawling, walking, flying) through the crowded market, a woman —(struck, pushed, kissed, came) against me and nearly fell.'

The test can be *banked cloze* in which all the words are shown, but out of order, at the bottom of the passage; that is the *word-bank* from which the selection is made. For a passage of twenty blanks, the bank may offer thirty words.

Scrambled sentences may be given and students asked to arrange them in a proper order. This is commonly used in many competitive examinations.

Most reading tests involve some writing and the understanding of words too; in a sense, they become a part of reading comprehension.

B. Testing listening comprehension

Except for listening to their teachers in the class, our learners of English are neither exposed to any listening nor testing of listening comprehension in the Indian context; but, in tests like the TOEFL that is required for admission to universities in the USA, or in IELTS (International English Language Testing System) required for admission to universities in the UK, Australia, and New Zealand, listening comprehension is tested. It is becoming more and more important even in job selection, telephone interviews in areas like information technology and management.

Generally, listening tests do involve reading, writing and, sometimes, even speaking. Yet, it is possible to test listening alone in terms of (a) discrimination of sounds, (b) recognizing specific elements like words, and (c) overall comprehension.

As in the case of testing reading, a test constructor must keep in mind the length of the material, coverage, authenticity and the level of difficulty in testing listening.

Listening can be done by giving instructions about the following:

- ticking the correct option
- circling the correct option
- underlining the correct/odd unit