CHAPTER-3

PLANNING AND PROCEDURE
CHAPTER-3

PLANNING AND PROCEDURE

3.1 Introduction
3.2 Research Design
3.3 Objective testing
   3.3.1 Subjective and Objective testing
   3.3.2 Objective Test
   3.3.3 Multiple-choice items
   3.3.4 The stem/the correct option
3.4 Vocabulary Tests
   3.4.1 Limited response
   3.4.2 Multiple choice completions
   3.4.3 Multiple choice paraphrases
   3.4.4 Simple Completion [Words]
3.5 Grammar Tests
   3.5.1 Limited Response Test
   3.5.2 Multiple-Choice Completion
   3.5.3 Simple Completion (Sentences)
   3.5.4 Cloze Procedure
3.6 Reading Comprehension Tests
   3.6.1 Alternate Forms of Limited Response Items
   3.6.2 Testing Sentence Comprehension
   3.6.3 Testing Passage Comprehension (Short and Long passages)
3.7 Test Administration
3.8 Sampling
   3.8.1 Definition and Purpose
   3.8.2 Methods of selecting a sampling
   3.8.3 Determination of sample size
   3.8.4 Avoiding of Sampling Bias
3.9 Item Analysis
  3.9.1 Difficulty Value
  3.9.2 Discrimination Index

3.10 Characteristics of a Standardized Test

3.11 Validity
  3.11.1 Concept of Validity
  3.11.2 Types of Validity

3.12 Reliability
  3.12.1 Concept of Reliability
  3.12.2 Types of Reliability

3.13 Norms

3.14 Conclusion
3.1 Introduction

After reviewing the literature the investigator has decided to work on proficiency test. For that planning is prerequisite thing. The planning of the test in the present investigation involved preparation of an outline of the specification of the language ability to be measured. It required the theoretical and scientific knowledge of the test construction and selection of items, their arrangement in sub tests, its administration and analysis.

3.2 Research Design

The design of the research work is based on the following three stages:

A. The Initial Stage:

After selecting the topic, the investigator started to be well familiar with language, language testing, language proficiency and proficiency test. She went through the related literature and reviewed them. From that she decided to test three areas viz. vocabulary, functional grammar and reading comprehension. It is a survey type research. To be more clear regarding the types of tests, the length of tests and the content of tests the investigator interviewed some experts and studied their views in detail and finalize the idea. Not only that but also the researcher went through the detail study of the English textbooks of Std. VI, VII and VIII. Thus keeping in mind the basic theory, various items were prepared and finally the test was constructed.

B. The Piloting:

To fulfill the objective of the piloting the investigator has tried out the piloting forms of the tests containing 300 items in two parts. After calculating the difficulty value and discrimination index the investigator has selected 100 items to construct final form of the test. The sample for piloting was 400 teachers of RMC.

C. The Final Run:

The final form of the test was tried out on the sample of 150 with a view to standardize the test. To standardize the test, the investigator has computed to estimate reliability and validity. The content and the construct validity were estimated. Moreover the reliability,
with using Test-Retest method and Kuder Richardson’s technique was estimated. Using the sample of 150 the investigator has established percentile norms and ‘T’ score.

To study the proficiency of the teachers working in the RMC the investigator has analyzed the data. She has tested all the hypothesis. The variables taken for the study were sex, age, educational qualification, experience if teaching and specialized subject. The sample is taken for the variable studies or to study the proficiency of the teachers working in the RMC was 150 teachers from Upper Primary Schools. After doing that the investigator has prepared a research report.

Thus it is a brief research design for the present study. All these stages are discussed in detail in the chapter IV.

### 3.3 Objective testing

#### 3.3.1 Subjective and Objective testing

“Subjective and objective are terms used to refer the scoring of tests. All test items, no matter how they are devised, require candidates to exercise a subjective judgment. In an essay test, for example, the testee must think of what to say and then express his ideas as well as possible; in a multiple-choice test he has to weigh up carefully all the alternatives and select the best one. Furthermore, all tests are constructed subjectively by the tester: he decides which areas of language to test, how to test those particular areas, and what kind of items to use for his purpose. Thus, it is only the scoring of a test that can be described as objective.” (Heaton, 1975)

This means that a testee will score the same mark no matter which examiner marks his or her test.

As objective tests usually have only one correct answer, they can be scored mechanically. The fact that objective tests can be marked by computer which is one of the important reason for their evident popularly among examining bodies responsible for testing large numbers of candidates.

Example A is subjective, while Example B is objective.
Example A: Write two sentences containing the word ‘while’

Example B: Complete the sentence by putting the correct word in the blank.

1. Shreya has been working in this school _____ 2000.

On the whole, objective test require far more careful preparation than subjective type of testing. In subjective testing the teacher tends to spend a relatively short time on setting the question but much time on marking. In an objective test the tester spends a great deal of time constructing each test items as carefully as she can, attempting to anticipate the various reactions of the testee at each stage.

3.3.2 Objective Test

Objective test are frequently criticized on the grounds that they are simpler to answer than subjective examinations. Items in an objective test, however, can be made just as easy or as difficult as the test constructor wishes. The fact that objective test may generally look easier but there is no indication at all that they are easier. The constructor of a standardized achievement or proficiency test not only selects and constructs his items carefully but also analyses student performance on each item and rewrites where necessary so that the final form of his test discriminates widely.

Another criticism is that objective tests of the multiple-choice type encourage guessing. However, 4 or 5 alternatives for each item are sufficient to reduce the possibility of guessing. Furthermore, experience shows, those candidates rarely make wild guesses: most base their guesses on partial knowledge.

A much wider sample of grammar, lexis and phonology can generally be included in an objective test than in a subjective test. In an essay, for example, the student can limit the range of the sentence structures and grammatical items she uses to those which is produced correctly. However, the fact that a test is apparently completely objective gives no guarantee that it will be a good test. It will be a very poor test if:

(1) The wrong features of the target language are tested,
(2) Irrelevant areas are emphasized in the test simply because they are ‘testable’, and
(3) The test items are poorly written.
It should never be claimed that objective tests can do those tasks which they are intended to do. They can never test ability to communicate in the target language nor can they evaluate actual performance. A good classroom test will probably contain subjective test items.

Although the examples have been confined to the testing of grammar, many of the test types can be used equally successful to test vocabulary as well as aspects of the reading, writing, listening and speaking skills.

3.3.3 Multiple-choice items

The multiple-choice item is now widely regarded as being one of the most useful of all the objective type items. Although it is amongst the most difficult of all objective item types to construct, but it is simple to score and administer.

In most public tests the optimum number of alternatives, or options, for each multiple choice item is five. Although a large number, say seven, would reduce even further the element of chance, it is extremely difficult and often impossible to construct as many as seven good options. Indeed, since it is often very difficult to construct items with even five options, four options are recommended for most classroom tests.

Before constructing any test items, first the test writer must determine the actual areas to be covered by multiple-choice items and the number of items to be included in the test. The test must be long enough to allow for a reliable assessment of the testee’s performance and short enough to be practicable. Too long a test is undesirable because of the administration difficulties often created and because of the mental strain and tension which may be caused among the students taking the test. The number of items included in a test will vary according to the level of difficulty, the nature of the areas being tested, and the purpose of the test. The teacher’s own experience will generally determine the length of a test for classroom use, while the length of a public test will be affected by various factors.

In each multiple-choice item the initial part is taken as the stem; the choices from which the student selects his answers are referred to as options/responses or alternatives. Out of it one option is the answer/correct option or key, while the other options are distracters. The task of a distracter is to distract the majority of poor students (i.e. those who do not know the answer) from the correct option.
Here is an example;

Sit here until she ____ you to come in. = stem

A. will tell
B. told = Options = distracters
C. is telling
D. tells = correct option/key

The following general principles should be observed when multiple choice items are constructed:

(1) Each multiple-choice item should have only one answer. Although this may seem an easy matter, it is sometimes difficult to construct an item having only one correct answer.

(2) Only one feature at a time should be tested. Few would wish to test both grammar and vocabulary at the same time, but sometimes word order and sequence of tenses are tested simultaneously. Such items are called impure items.

(3) It has to be noted that each option should be grammatically correct when placed in the stem, except of course in the case of specific grammar test items.

(4) All multiple-choice items should be at a level appropriate to the linguistic ability of the testees. The context, itself, should be at a lower level than the actual problem, which the item is testing. A grammar test item should not contain other grammatical features as difficult as the area being tested, and a vocabulary item should not contain more difficult semantic features in the stem than the area being tested.

(5) Multiple-choice items should be as brief as possible (though it is often desirable to provide short contexts for grammar items).

(6) In many tests, items are generally arranged in rough order which increase difficulty. It is generally considered important to have one or two simple items to “lead in” the testees, especially if they are not too familiar with the kind of test being administered.
3.3.4 The stem/the correct option

The primary purpose of the stem is to present the problem clearly and concisely. The testee should be able to obtain from the stem a very general idea of the problem and the answer required. At the same time, the stem should not contain extraneous information or irrelevant clues, thereby confusing the problem being tested. The stem should be short; it should convey enough information to indicate the basis on which the correct option should be selected.

The stem should usually contain those words or phrases which would otherwise have to be repeated in each option. For example,

The word “astronauts” is used in the passages to refer to

A. Travelers in an ocean linear  
B. Travelers in submarine
C. Travelers in a space-ship  
D. Travelers in a balloon

Instead of it, the stem here should be re-written so that it reads like:

The word “astronauts” is used in the passage to refer to travelers in

A. an ocean linear  
B. a Submarine
C. a space-ship  
D. a balloon

Another thing is that, the correct option should appear each position (e.g. A, B, C, D or E) approximately the same number of times in a test or sub-test. This can usually be achieved if it is placed at random in a certain position or if all the options are placed in alphabetical order (i.e. according to the first letter of the first word in each option). However, if the options have a natural order (e.g. figures, dates), it is advisable to keep it in this order as:

Nehru started his career as a barrister in

A. 1921  
B. 1925  
C. 1926  
D. 1932
3.4 Vocabulary Tests

The purpose of vocabulary test is to measure the comprehension and production of words used in speaking or writing.

Four general kinds of vocabulary tests are presented. The first, limited responses, is for beginners. These test items require either a simple physical action like pointing at something or a very simple verbal answer such as “yes” or “no.” The second, multiple-choice completion is a test in which a sentence with a missing word is presented; students choose one of four vocabulary items given to complete the sentence. A third type, multiple choice paraphrases, is a test in which a sentence with one word underlined is given. Students choose out of four words which is the closest in meaning to the underlined item. A fourth kind of test, simple completion (words), in which students write in the missing part that appear in sentences.

3.4.1 Limited response

In testing children and beginning-level adults, we often use direct physical responses and visuals. We do this to avoid language skills that have not been mastered yet. For these tests, students don’t have to know how to read or write. In fact, they don’t even have to know how to speak.

Individual Testing

You can test beginning students by asking for only a very short answer-for example, “Is the book green?” There are of course other kinds of brief answers that can be used with visuals (“What color is the book?”/ “Green or Red” or is the boy sleeping or swimming?”/ “[He’s] swimming”).

Group Testing

For group testing, non verbal physical responses as well used to test the whole class at the same time. A good way is to draw or duplicate a sketch such as the one on the opposite page and give one to each person in the class. After an example, (such as, “Draw a circle around the boy”), you can say, “Now draw a circle around a tree.” Another command might be to “Put an ‘X’ on every tool.” Still another would be to have an empty clock face on the paper. You could then have students draw hands on the clock showing, for example, that it is 3:30.
Naturally, activities, visuals, and objects can be used with more advanced students as well: a botany student could name the part of flower, and a technical student could describe a pump. But we normally use other ways to test people who can read and write.

**Advantages of Limited Responses**

1. It causes less stress or nervousness than other types of tests.
2. It avoids skills such as reading and writing that have not yet been developed.
3. It can be scored easily and objectively.

**Limitations of Limited Responses**

1. It requires individual testing, which takes longer than group testing.
2. It is usually difficult to test abstract words with this technique.
3. Sketches are sometimes ambiguous (e.g. an orange may look like a ball; running may look like dancing or jumping)

**3.4.2 Multiple choice completions**

The following steps should be taken in writing multiple-choice completion items:

1. (1) Select the words to be tested.
2. (2) Get the right kind of sentence to put each word in (this sentence context is called the stem).
3. (3) Choose the several wrong words to put the right word with (these wrong words are called distracters). Three distracters plus the right word are enough for a written item.
4. (4) Finally, prepare clear and simple instruction.

**Alternate Forms of Multiple-Choice Completion**

Following are alternate ways to prepare vocabulary multiple-choice completion items:

1. **Definition.**
   
   To _____ someone means to save him or her from danger.

   A. praise       B. injure       *C. rescue       D. announce

2. **Phrase completion.**
   
   a. Idioms.
The committee _____ choosing you as president.

*A. wound up       B. buttoned up       C. wiped out       D. Sat in

b. Appropriateness to context.

The newspaper says, “A two-year old girl _____ today when struck by a bus.”

A. kicked the bucket       B. was eradicated       *C. was killed       D. departed

3. Phrasal context (not sentence completion).

He _____ his fingernails.

A. tailored       B. reduced       C. remodeled       *D. manicured

4. Multiple-choice cloze

Cloze tests are made from stories or essays by deleting words at regular intervals. Students have to write in each blank the word that they think to belong there. Multiple-choice cloze tests work like regular multiple-choice sentence completion. As,

After the capture of Troy, Ulysses set out for his (A. neighborhood       B.continent

*C.homeland       D. street) many miles away. But so many strange (A. Sights

*B.things       C.places       D.people) happened to him on his journey that ten

(*A. years       B.times       C. roads       D.cities) passed before he reached Ithaca!

Advantages of Multiple-Choice Completion

1. It helps students to see the full meaning of words by providing natural contexts. Also, it is a good influence on instruction: it discourages word-list memorization.
2. Scoring is easy and consistent.
3. It is a sensitive measure of achievement.

Limitations of Multiple-Choice Completion

1. It is rather difficult to prepare good sentence contexts that clearly show the meaning of the word being tested.
2. It is easy for students to cheat by copying what others have circled.
3.4.3 Multiple choice paraphrases

Multiple-choice paraphrase tests of vocabulary items offer much of the same advantage that multiple-choice completion tests do, and the contexts are much easier to prepare. Understanding is checked by the student’s having to choose the best synonyms or paraphrase of the vocabulary item. A sentence context is still used. However, choosing the right word depends more on knowing the key vocabulary item than on finding meaning in the sentence.

Vocabulary Choice and Context Preparation

In writing paraphrase items, we follow the same steps that we took to prepare completion items:

(1) Select the words to be tested;
(2) Prepare a sentence context;
(3) Choose distracters; and
(4) Write instructions.

The meaning comes more from the emphasized word than from its context. Here is a typical item.

He was *irate* when he heard about the new plans.

A. Interested  B. surprised  *C. angry*  D. sad

Alternate Forms of Multiple-Choice Paraphrase

There are two kinds of paraphrase questions: Both use a key word in a sentence. But one asks the student to find the best synonyms or related word, and in the other the student has to choose the phrase that is the best short definition or paraphrase of the key word. The following are four alternate ways to write paraphrase items.

1. Idioms and the other phrases.

The salesman seemed *quite down and out.*

*A. Poor*  B. disappointed  *C. lost*  D. Angry

2. Phrasal context.
A considerate little lady.

A. worried  B. tired  C. happy  *D. Kind

3. Reading passage context. (Vocabulary questions can be included with reading comprehension questions.)

Here is an example:

Just then we saw him run out of the side door. As he turned the corner, a slip of paper fluttered to the ground...With that paper, the police were able to trace the man’s whereabouts.

In this passage, trace means ______

A. copy  B. enter  *C. locate  D. Eliminate

4. Related-word identification. (These questions do not use synonyms; they use examples and no examples of the key word.)

He eats lots of vegetables.

A. bananas  B. peaches  C. oranges  *D. carrots

5. Unrelated-word identification.

He lives in a big house.

A. Attic  *B. car  C. basement  D. bedroom

Advantages of Multiple-Choice Paraphrase

1. Context preparation is rather easy.
2. Scoring is easy and consistent.
3. It is a sensitive measure of achievement.

Limitations of Multiple-Choice Paraphrase

1. It is difficult to find good synonyms.
2. It is easy for students to cheat.

3.4.4 Simple Completion [Words]
Word-formation items require students to fill in missing parts of words that appears in sentences. These missing parts are usually prefixes and suffixes, for example, the un- in untie or the –ful in thankful. A related task is to use words like this in a sentence and have students supply missing syllables of any kind.

The steps in preparing a simple-completion vocabulary test are similar to those followed in the two previous types, but with one difference: now no distracters are needed.

**Alternate Forms of Simple-Completion Items**

Here are four additional ways to prepare simple-completion questions.

1. Stem-first procedure. (An advantage is that many words need spelling changes when suffixes are added. This allows for such changes.)
   
   *(beauty)* She has a *beautiful* new dress.

2. Phrasal context (Note that grammatical clues are sometimes given to less advanced students.)
   
   An *in* convenient delay.

3. Compounds.

   He found the bedroom, but he couldn’t open the door to the clothes *closet*.

4. Inflectional cloze.

**Advantages of Simple Completion [Words]**

1. It reflects teaching approaches.
2. It is generally faster and easier to construct than are items with distracters.

**Limitations of Simple Completion [Words]**

1. Fewer words can be tested this way than with multiple choice.
2. There is some difficulty in avoiding ambiguous contexts.
3.5 Grammar Tests

Grammar tests are designed to measure students’ proficiency in matters ranging from inflections (bottle-bottles, bake-baked) to syntax. Syntax involves the relationship of words in a sentence, including matters such as word order, use of the negative, question forms, and connectives. Each of the four sections of this is discussed here. The first approach, limited response, is especially useful for students at the beginning level. The remaining sections discuss multiple-choice completion, simple completion, and cloze tests.

Of course, in testing grammar, we don’t pretend to measure actual communication. But we can do a good job of measuring progress in a grammar class, and we can diagnose students’ needs in this area.

3.5.1 Limited Response Test

The grammar of students with very little ability in English can be checked without having them speak or write anything. This can be done by means of mentioned physical responses and visuals. This type of test presents two basic ways to measure the grammar skill of these beginning-level students: (1) testing them one at a time and (2) testing them in groups.

Individual Testing

One can test students individually by using oral requests. These requests can ask for easy spoken replies or simply for nonverbal actions. When teaching students who know almost no English, one can even permit answers in their native language.

Example: (students hear in English)

“How many books are there on the table?”

(Students answer in their language)

“[There are] six.”

Group Testing

One can also test students in groups by using directed physical responses. The following “drawing” activity can test prepositions of place: First, explain and illustrate any new
vocabulary words. Then have students make a drawing according to your spoken instructions: “Draw an airplane in the middle of the paper. [Pause while students draw.]” Now draw a house below the airplane. [Pause] next draw a cloud in front of airplane.”

**Advantage of Limited Response Test**

1. It puts student at ease and avoids unnecessary stress.
2. It avoids skills such as reading and writing that have not yet been developed.
3. It can be scored easily and objectively.

**Limitations of Limited Response Test**

1. Individual testing takes longer than group testing.
2. It is difficult to find suitable pictures (although the teacher can make needed sketches).
3. Only a limited number of grammatical structures can be tested.

**3.5.2 Multiple-Choice Completion**

This type of test includes an incomplete sentence stem followed by four multiple-choice options for completing the sentences.

Preparing multiple-choice completion *grammar* items follow the procedure like: about the same procedure as that described in this for writing multiple-choice completion *vocabulary* items: (1) Choose the grammar points that you need to test; (2) prepare the right kind of sentence context (or stem) for the grammar structure; (3) select three logical distracters; and (4) prepare clear, simple instructions.

It is good not to confuse or tire your students by having them reread unnecessary material. Take out any repeated words from the distracters and put these in stem.

(Poor) If I had a new fur coat, _____

A. I showed it to everyone. *B. I’d show it to every-one C. I’ve shown it to everyone.

D. I’ll show it to everyone.

(Better) If I had a new fur coat, _____ it to everyone.
A. I showed  *B. I’d show  C. I’ve shown  D. I’ll show

Alternate forms of Multiple-Choice Completion

1. **Space saver.**

The next example simply puts distracters inside the stem:

Tom lives (A. At,  *B. on,  C. in,  D. inside) centre street.

2. **Dialogue context.**

“Did she ask you to go with her?”

“No, she asked someone _____ instead.”

*A. Else  B. another  C. other  D. ----

3. **Error identification.**

Unlike previous test items in this section, error identification does not require students’ to complete a sentence. Instead, they have to find the part containing an error. This kind of test question is particularly useful in testing grammar points for which there are few logical options.

I Visited Netherlands/ for a week last year/while/ you were in school. (No error)

* A  B  C  D  E

Advantages of Multiple-Choice Completion

1. It is impossible for students to avoid the grammar point to being evaluated.
2. Scoring is easy and reliable.
3. This is a sensitive measure of achievement (and like other multiple-choice language tests, it allows teachers to diagnose specific problems of students).

Limitations of Multiple-Choice Completion

1. Preparing good items is not easy.
2. It is easy for students to cheat. (It is possible to create a second form of the test by rearranging the items, but this is time consuming for the teacher.)
3. It doesn’t appear to measure students’ ability to reproduce language structures (although in actual fact this kind of test is a good measure of the grammar sub skill).

4. This can have a negative influence on class work if used exclusively. (Students may see no need to practice writing if tests are objective.)

3.5.3 Simple Completion (Sentences)

Simple completion items used for testing grammar consist of a sentence from which a grammatical element has been removed. An elementary item would be “He went to school.” An advanced open-ended item would be “I would have gone if he had invited me.” Students may be asked to decide from the context what word or phrase to write in the blank; or they may be asked to write in an option from a list, or to change the form of a key word (such as write to wrote).

There are three steps to follow in preparing simple-completion grammar tests: (1) Select the grammar points that need to be tested; (2) provide an appropriate context; and (3) write good instructions. But it is also necessary to decide what kind of simple-completion question to use. Some are easier than multiple-choice completion, and some are much more difficult. Most of this section will deal with the three basic kinds of simple-completion grammar tests: (1) the option form, (2) the inflection form, and (3) the free-response form.

(1) The Option Form

The easiest simple-completion items are like multiple-choice questions with only two options.

Directions: Complete the following sentences with “do” or “make.”

1. He made a lot of money last year.
2. I always do my best.

But generally three or four options are preferable for a good test.

(2) The Inflection Form

Testing the mastery of inflections provides for a productive response.
1. He’s the tallest person in the class.
2. They were in Colorado last week.

When students have to write in their own answer like this, you have to be careful about context. For example, new teachers might think that if they write a sentence like “He _____ (sing) a song,” only “is singing” will fit. If they’re testing the progressive, they may be disappointed to find that several other answers are possible, such as sings, sang, has been singing, had been singing, will sing etc. This problem can be solved by giving part of the verb or adding more context.

1. He is singing (sing). (or) He is singing now (Add one word.)
2. “What’s Tom doing now?” “Oh, he is singing (sing).”

(3) The Free-Response Form

Sometimes a few simple terms can be used, if everybody in the class knows what they mean. The free-response form illustrates how that common terminology can occasionally be used. Here are some sentences.

Directions: Write in the missing part of the two-word verb.

Example: “What time did he get up this morning?”

This example illustrates free response with a minimum amount of contextual control.

“You would be better sooner if ______.”

These take longer to correct than other completion types, and they also take more language skill to evaluate properly. Consider a few acceptable ways that students could complete example “if you dressed warmer,” “if you’d see a doctor,” “if mother were here,” “if we had some medicine for you.” Obviously, this kind of simple-completion question requires the most real productivity of all. It also provides flexibility; and it is perhaps the most communicative.

Advantages of Simple Completion [Sentences]

1. These are generally easier to prepare than are multiple-choice items.
2. These give the appearance of measuring productive skills because some items permit flexibility and original expression.
3. There is no exposure to incorrect grammatical forms.
4. These provide a sensitive measure of achievement.

Limitations of Simple Completion [Sentences]

1. These are usually more time consuming to correct than are multiple-choice questions. Not only can poor penmanship be a problem but also “irrelevant” errors beyond those being tested.
2. Occasionally students can unexpectedly avoid the structure being tested.

3.5.4 Cloze Procedure

Cloze test are prose passages, usually a paragraph or more in length, from words have been deleted. The student relies on the context in order to supply the missing words.

At the present time, no single test format is more popular than cloze procedure. It is easy to prepare and rather easy to score. Teachers like it too because it is integrative – that is, it requires students to process the components of language simultaneously, much like what happens when people communicate. Moreover, studies have shown that it relates well to various language measures - from listening comprehension to overall performance on a battery of language tests. In brief, it is a good measure of overall proficiency.

Before examining how to prepare this kind of test, let’s consider briefly the principle that underlines it. The cloze is simply a story or essay from which a number of words have been deleted. We fill in the missing words much as we do while conversing. In a noisy restaurant, we guess at the words that we don’t hear by relying on the whole conversation. So in cloze tests, the overall meaning and surrounding grammar help us to replace the missing parts.

Preparing a Cloze Test

The steps in preparing a cloze test are simple:

1. Select an appropriate passage (e.g. from the reading material in your ESL class);
2. Decide on the ratio of words to take out;
3. Write the instructions and prepare an example. Each of these steps will be explained in the paragraphs that follow.
The first and most important step is to choose a story or essay on the right level. If your class uses an ESL reader, choosing a passage that is rather difficult for your students will simply frustrate them. So choose a passage that they can read with little or no difficulty. You can even use something that has already been read and discussed in class. They will not be able to answer the test from memory. The length of the selection depends on the number of blanks you plan to have. But most are not longer than 300 to 400 words. This means that you will often have to use only part of an article or story.

Adapting the Cloze to Test Grammar

For a measure of general grammatical ability, take out function words (pronouns, articles and determiners, auxiliary verbs, prepositions, and conjunctions). Here is a short passage from an ESL text that illustrates this procedure:

Everyday thousands of people jog, why has jogging-running slowly for long distances-becomes so popular? Donald Robbins, who is forty-two years old (1) works in an office, began jogging (2) few years ago because he felt (3) was too fat. At first he (4) only run about 100 yards, and (5) took him almost three months to (6) able to run a full mile. (7) two years later, he ran in (8) eastern marathon race-over twenty-six miles. (9) you jog too? If you decide (10), be sure to ask your doctor (11) advice.

Does jogging cost much? No, (12) costs almost nothing. But most agree (13) good running shoes are very important. (14) protect your feet and legs from (15) shock of running on hard surfaces.

If you start jogging, it could make your heart stronger and also help you to feel better about yourself.

(Key: (1) and, (2) a, (3) he, (4) could, (5) it (anticipatory pronoun), (6) be, (7) But, (8) an, (9) Should, (10) to, (11) for, (12) it (personal pronoun), (13) that, (14) They, (15) the.

It would be better if you give four options for each correct answer.

Advantages of Cloze

1. It is easy to prepare and quite easy to score.
2. It is a good measure of integrative English skills.
3. Standard cloze is a good measure of overall ability in English.

Limitations of Cloze

1. It is not a sensitive measure of short-term gains.
2. It is difficult for teachers who are non-native English speakers to choose acceptable equivalent words.

3.6 Reading Comprehension Tests

Tests of reading come in a wide variety of forms and evaluate broad spectrum activities. These range from pre-reading concerns (learning the Roman alphabet, for example, or word-attack skills) to reading comprehension, reading speed, and skimming techniques. Advanced and more specialized applications include translation, reading aloud, and reading literature. Reading speed is especially important for students with lots of out-of-class reading to do. Skimming is handy for people who need to hunt for information in print: This includes reading a newspaper as well as doing research in a library.

3.6.1 Alternate Forms of Limited Response Items

1. **Advanced odd items.** The student still looks for the word that is different from the others. But now all three can be different, or all three can be the same. If only one word is different, he circles, the odd word. If each word is different, he checks the “All Different” column. If all three are the same, he checks the “All the same” column.

<table>
<thead>
<tr>
<th>Items</th>
<th>All Different</th>
<th>All The Same</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speak Peeks</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Peace Peace</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Into bed In the bed</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. **Key word and odd option.** There is only one difference between this alternate form and the original. Originally, we found the word that matched the key item. Here is word that does not match the key item, in short, the different word:
**Pleasure:** Pleasing  pleasure  pleasure  pleasure

**Want to go:** want to go  want to go  want to go  won’t go

**Advantages of Limited-Response Items**

1. These are quite easy to construct and score.
2. Only the recognition of letters is required, making this a simple task for beginning students.

**Limitations of Limited-Response Items**

1. This is not interrogative skill involving actual reading.
2. Overemphasis on this technique could reduce reading speed.

**3.6.2 Testing Sentence Comprehension**

There is no need to test comprehension of an essay if students still have difficulty understanding a sentence. Sentence comprehension must precede essay comprehension. Some sentence-level comprehension items are good for beginning students. Some elementary questions simply ask for the right picture to be circled. There are other questions that ask only for a one-word response.

**Elicitation Techniques**

1. **Picture Cues:** We will start with questions that simply ask for the right picture to be circled.

2. **Phrase and Sentence Cues:** one of the simplest forms of sentence comprehension is the true-false sentence. It is used with beginners. The student looks for truths, untruths, or impossibilities:
   - The sun sets in the east.  T  F
   - She is my brother.  T  F

**Advantages of Sentence-Comprehension Items**

1. It is rather easy to write true–false items on pictures.
2. These are good for testing the skills of near beginning students.
3. This is a rapid way to test reading comprehension.
Limitations of Sentence-Comprehension Items

1. Finding good pictures can be rather time consuming.
2. Not all reading skills are converted in sentence-comprehension questions.

3.6.3 Testing Passage Comprehension (Short and Long passages)

The most integrative and challenging kind of reading test type is passage comprehension.

Context Selection

Sometimes these unusual passages appear in the modern ESL readers and sometimes teachers—particularly teachers of secondary school students—collect particular reading items from the community for the use in class. Besides the usual articles and stories, these may include advertisements, want ads, business and social letters, driver’s license and loan applications, bank statements, rental and sales agreements, and mail order catalogs. Other reading materials may include texts from other classes at school such as biology, history, or chemistry.

Question Techniques for More Advanced Students

Standard Multiple-Choice

There are many ways to test reading. One of the best is a reading passage followed by multiple-choice questions.

The number of passages and the length of each depend on your particular test. Let’s assume the whole exam is on reading. Multiple-choice questions can be asked on very short passages of 35 to 75 words. Quite a few of these can be answered in one period. Student level and passage difficulty naturally influence how many can be done.

Usually longer passages will run from 100 to 300 words. This is sufficient since more than one passage will appear on a single test.

Plan to use a variety of types of questions on your reading test. One very important type is the paraphrase. Look at the following example:

Karate is a science of unarmed self-defense and counterattack. It is a sort of “weapon in an empty hand.” In many U.S. cities thousands of young people are developing their minds as well as their bodies by learning karate.
The key portion that we will use for our paraphrase question is “In many U.S. cities, thousands of young people are ...learning karate.” The paraphrase of this is “Karate is being taught to many young Americans.” Every word but “karate” is different. Here is the resulting question:

In this passage we learn that karate ______

*A. Is being taught to many young Americans.

B. And training for the mind and both being taught.

C. An remove a weapon from someone’s hand

D. Is used to start a fight.

**Advantages of Passage Comprehension**

1. This is most integrative type of reading test.
2. It is objective and easy to score.
3. It can evaluate students at every level of reading development.

**Limitation of Passage Comprehension**

1. Passage comprehension is more time consuming to take than other kinds of tests.
2. One pitfall in preparing this kind of test is utilizing questions that deal with trivial details.
3. Passage-comprehension tests which use questions on trivial details encourage word-by-word reading.

**3.7 Test Administration**

It is important to remember that a test must be practicable: in other words, it must be fairly straightforward to administer.

Peter (1977) suggests three basic principles of test administration:

(1) It is essential that a suitable testing situation be arranged to enable the subjects to do justice to them.

(2) It is also important that the subjects be suitably motivated.
(3) In the case of standardized test, the test should be administered accordingly.

The length of time available for the administration of the test is frequently misjudged even by experienced test-writers, especially if the complete test consists of a number of sub tests. In such cases sufficient time may not be allowed for administration of the test, the collection of the answer sheets, the reading the test instructions, etc. In the case of all large scale tests, the time to be allowed should be decided on as a result of a pilot administration of the test.

Yet another practical consideration concerns the answer sheets and the stationary used. Many tests require the testee to enter his answer on the actual question paper (e.g. circling the letter of the correct option), thereby unfortunately reducing the speed of the scoring and preventing the question paper from being used a second time. In some tests, the candidate is presented with a separate answer, but too often insufficient thought has been given to errors arising from the (mental) transfer of the answer from the context of the item on the question paper to the answer sheet itself. The use of separate answer sheets, however, greatly facilitates marking (through the use of a mask or key) and is strongly recommended when large numbers of students are being tested.

Before a test, one must make certain that the necessary equipment is available in each centre and that there is a high degree of standardization in the test administration. For example, it is useless to record talks or dialogues on tape if certain test centre does not have a tape recorder. What is not so obvious, however, is the potential unreliability of an auditory test resulting from the different sizes of the rooms where the test is administered and the different degrees of interference caused by extraneous noise.

A final point concerns the presentation of the test paper itself. Where possible, it should be printed or type written and appear neat, tidy and aesthetically pleasing. Nothing is worse and more disconcerting to the testee than an untidy test-paper, full of misspellings, omissions and corrections.
3.8 Sampling

3.8.1 Definition and Purpose

“Sampling is the process of selecting a number of individuals for a study in such a way that the individuals represent the larger group from which they were selected.”

The purpose of sampling is to use a sample to gain information about a population.

Definition of a population:

“A population is the group to which a researcher would like the results of a study to be generalizable.”

A defined population has at least one characteristic that differentiates it from the other group. The population that the researcher would ideally like to generalize results is referred as the target population; the population that researcher realistically selects from is referred to as the accessible or available population.

3.8.2 Methods of selecting a sampling

Regards of the specific technique used, the steps in sampling include identification of the population, determination of recognized sample size and selection of the sample. The degree to which the selected samples represent the population is the degree to which results are generalizable.

(1) Cluster Sampling

Cluster sampling is sampling in which groups, not individuals are randomly selected. Any intact group of similar characteristic is a cluster. The steps in cluster sampling are similar to those in random sampling except that random selection of group clusters is involved, not individuals.

(2) Systematic Sampling

In systematic sampling individuals are selected from a list by taking every Kth name, where K equals the number of individuals on the list divided by the number of subjects described for the sample.
Random sampling is the process of selecting in such a way that all individuals in the defined population have equal and independent chance of being selected for the sample. In other words, every individual has the same probability of being selected and selection of one in no way affects selection of another individual.

Random sampling is the best single way to obtain a representative sample. No technique, not even random sampling guarantees a representative sample, but the probability is higher for this procedure than for any other. Differences between the sample and population should be small and unsystematic. For example, you would not expect the ratio of males and females in a sample as in a population; random sampling, however, assures that the ratio will be close and that probability of having too many females is the same as the probability of having too many males. In any event, differences are a function of chance and are not the result of any conscious or unconscious bias on the part of researcher.

Another point in favor of random sampling is that it is required by inferential statistics. This is very important since inferential statistics permit the researcher to make inference about population based on the behavior of samples. If samples are not randomly selected, then one of the major assumptions of inferential statistics is violated, and inferences are correspondingly tenuous.

Steps in random sampling

In general, random sampling involves defining the population, identifying each member of the population, and selecting individuals for the sample on a completely chance basis. One way to do this is to write each individual’s name on a separate slip of paper, place all the slips in a hat or other container, shake the container, and select slips from the container until the desired number of individuals is selected. This procedure is not exactly satisfactory, if a population had 1,000 members. At that time more satisfactory approach is to use table of random numbers. In it, a table of random numbers selects the sample for you. Each member is being selected on a purely random basis. Such tables are included in the appendix of most statistics books and some educational research books. They usually consist of columns of five-digit numbers which have been randomly generated by a
computer. Using the data of random numbers to select a sample involves the following steps.

1. Identify and define the population.

2. Determine the desired sample size.

3. List all numbers of population.

4. Assign all individuals on the list a consecutive number from zero to the required number. For example, 000-249 or 00-89.

5. Select an arbitrary number in the table of random numbers.

6. For the selected number, look at only appropriate number of digits. For example, if a population has 800 members, you only need to use the last 3 digits of the numbers; if a population has 90 members, you only need to use the last 2 digits.

(4) Stratified Sampling

Stratified sampling is the process of selecting a sample in such a way that identified subgroups in the population are represented in the sample in the same proportion that they exist in the population. It can also be used to select equal sized samples from each of a number of subgroups if subgroup comparisons are desired. On the other hand, equal-sized samples would be desired if you wanted to compare the performance of different subgroups.

Suppose, for example, that you were interested in comparing the performance of students of different IQ levels (Say high, medium and low) by following two different methods of mathematics instruction. Simply randomly selecting a sample and assigning one-half of the samples of each of the methods would not guarantee equal representation of each of the IQ levels in each method. In fact, just by chance, one of the methods might not have any students from one of the levels. However randomly selecting students from level, and then assigning half of each selected group to each of the methods, would guarantee equal representation of each IQ level in each method (That is the purpose of stratified sampling, to guarantee desired representation of relevant subgroups).
Steps in Stratified Sampling

The steps in stratified sampling are very similar to those in random sampling except that selection is from subgroups in the population rather than the population as a whole. Stratified sampling involves the following steps.

1. Identify and define the population.
2. Determine desired sample size.
3. Identify the variable and subgroups (strata) for which you want to guarantee appropriate representation (either proportional or equal).
4. Classify all members of the population as members of one of the identified subgroups.
5. Randomly select (using a table of random numbers) an “appropriate” number of individuals from each of subgroups “appropriate” meaning either a proportional number of individuals or an equal number of individuals.

3.8.3 Determination of sample size

The sample would be as large as possible in general, the large the sample, the more representative it is likely to be, and the more generalizable the results of the study are likely to be.

Minimum, acceptable sample sizes depend on the type of research: descriptive research 10% of the population is enough. Even large sample can lead to erroneous conclusion if they are not well selected.

3.8.4 Avoiding of Sampling Bias

Sampling bias does not result from random, chance difference between sampling and populations; sampling bias is systematic and is generally the fault of the researcher. Two major sources of bias are the use of volunteers and the use of available groups. Any sampling bias presenting a study should be fully described in the final research report.
3.9 Item Analysis

As a matter of fact, item analysis is a long process but that cannot be avoided. If one wants to make each item consistent with the total test he must pass through the way of item analysis. Guilford (1954) explains the purpose of item analysis as under:

“The most common use of the item analysis is the selection of best items to compose the final test form”

Item analysis also helps to improve the test as Analysis (1968) reports:

“Item analysis makes it possible to shorten a test and at the same time increases its reliability and validity.”

Selection of appropriate language items is not enough by itself to ensure a good test. Each question needs to function properly; otherwise, it can weaken the exam. Fortunately, there are some rather simple statistical ways of checking individual items. This procedure is called “item analysis.” It is most often used with multiple-choice questions. An item analysis tells us basically three things: how difficult each item is, whether or not the question “discriminates” or tells the difference between high and low students, and which distracters are working as they should. An analysis like this is used with any important exam- for example, review tests and tests given at the end of a school term or course.

To prepare for the item analysis, first score all of the tests. Then arrange them in order from the one with the highest score to the one with the lowest. Next, divide the papers into three equal groups: those with the highest scores in one stack and the lowest in another. The classical procedure is to choose the top 27 percent and the bottom 27 percent of the papers analysis. But since language classes are same in results and allow us to use a few more papers in the analysis. The middle group can be put aside for awhile.

In the present study also, investigator passed through the process of item analysis which had the following steps. Item analysis helps in deciding the facts.

1. Calculation of difficulty value; and

2. Calculation of discrimination Index.
3.9.1 Difficulty Value

As said earlier, the calculation of difficulty value is a part of item analysis. It is also known as difficulty index. Difficulty value of an item is not only the property of that item but it also reflects the ability of the person, responding to the item. Such reflection of ability is the proficiency of a person which is of more interest.

Lindeman (1930) writing about the importance of difficulty index says.

“The difficulty level of a test item provides some indication of the extent to which the item is doing its job.”

After accepting the importance of difficulty value, it will be interesting to know the method of determining difficulty index.

Garrett (1967) gives three ways of determining difficulty index but he further says,

“Difficulty index by the number of examinations in the group who got the item correct is the standard method for determining the difficulty in objective tests.”

In this method, “total of correct number in both the groups is divided by double of the total respondents of both groups.” The same method and the following formula were used for calculation of present study.

Formula used is,

\[
\text{D.I.} = \frac{U + L}{2N}
\]

Where D.I = Difficulty Index.

\[
U = \text{No. of correct responses in upper group}
\]

\[
L = \text{No. of correct responses in lower group}
\]

\[
N = \text{Total No. of respondents in each group.}
\]
With the help of this formula, difficulty index for all the items were calculated which ranged from 0 to 0.98

Now for the selection of items the test constructor adopted the following procedure.

Trivedi and Parekh (1972) suggest that:

“The items having difficulty value either 0.5 or nearer to it should be considered good items; however there is no definite formula for it.”

3.9.2 Discrimination Index

The discriminatory index of an item is of considerable importance in the process of item selection. Discrimination index discriminates higher and lower students from the group of measuring the proficiency.

Rebecca M. Vallette (1967) clarifies the term as,

“Discrimination power tells us how well item performance in separating the better students from the poorer students.”

In short, if a test is to tell something about an individual it must be capable of discriminating among the groups of which he is a member.

There are about 60 methods of determining discrimination index but for calculation of the present study, setting of two extremes groups of high scores and low scores was done. The results of correct answers of both the groups were also used here.

The most commonly formula given by Johnson (1965) was followed.

\[
ULI = D.I. = \frac{U - L}{F}
\]


\[
U = \text{No. of correct responses in upper group}
\]

\[
L = \text{No. of correct responses in lower group}
\]

\[
F = \text{Total No. of respondents in each group.}
\]
The D.I. of all the items was calculated with this formula which ranged from -0.19 to 0.81.

The most common and practical suggestion for selection of the item is given by Trivedi and Parikh(1972)

Scoring to them, following criterion is given.

0.40 and more - Very good items

0.39 to .030  - good items

0.20 to 0.29  - fair items, require improvement

0.19 to less  - weak items to be rejected.

With the said criteria, the investigator selected only those items which has their D.I. greater than 0.30 and they were accepted for the final form of the proficiency test.

3.10 Characteristics of a Standardized Test

There are a number of characteristics shared by standardized tests. While these characteristics are desirable for all tests, they are much more likely to be in evidence for a standardized test. Regardless of the type of test being sought, there are certain kinds of information that should be available about any standardized tests. This information, or lack of it, forms the basis for test selection.

Standardization is a process which makes the test standard in many respects for measuring different aspects. Standardization is a procedure of refining a measuring instrument through scientific processes. Noll (1965) defines standardized test as,

"One that has been carefully constructed by experts into the light of acceptable objectives of purposes; procedures for administering, scoring, and interpreting scores are specified in detail so that no matter who gives the test or where it may be given, the results should be comparable; and norms and averages for different age or grade levels have been predetermined."
Standardized tests are typically developed by experts and are therefore well constructed. Individual test items are analyzed and revised until they meet given standards of quantity. Direction for administrating scoring and interpreting standardized tests are carefully specified on resulting characteristic. Another major characteristic of standardized tests is the existence of validity and reliability data. Validity is the most important quality of any test. Validity means the instrument measures what it is supposed to measure, and nothing else. Reliability refers to the consistency with which a test measures whatever it measures.

Specification of conditions of administration is a very important characteristic of a standardized test. This insures that if directions are carefully followed, the test will always be administered the same way and will always yield the same kind of data. If directions are carefully spelled put, even at beginning researcher should be able to properly administer most tests. The directions usually include instruction to be read out to the test taker, restriction on time, if any and guidelines concerning the amount and nature of communication permitted between the test administrator and the test takers. Last but not the least; standardized tests generally include directions for scoring and guidance for interpretation of scores.

Directions for scoring include specifications such as the criteria for acceptable responses, when such appropriate, the number of points to be assigned to various responses, and procedure for computing total scores. Guidelines for test score interpretation generally include a table of norms. Typically, a test is administered to a large number of appropriately defined individuals, and resulting test scores are analyzed. A table of norms presents law scores and one or more equivalent transformations, such as corresponding percentage ranks, which facilitate interpretation of an individual’s score with respect to the performance of the group since researchers generally work with law scores. This type of information is generally of made value.

As stated previously, the most important characteristic of a standardized test, or any test for that matter, is validity. Validity is totally indispensable; there is no quantity or virtue of a test that can compensate for inadequate validity.
3.11 Validity

3.11.1 Concept of Validity

The most simplistic definition of validity is that it is the degree to which a test measures what it is supposed to measure. As discussed before, validity is one of the essential characteristics of a good test. A very simple question is always asked to check the validity of a test. The simple question is: "Does the test measure what it claims to measure?" If it does, it is a valid test.

Robert Lado (1951) explains as:

"Validity is essentially a matter of relevance."

This suggests that for a test to be valid, it is expected that the content and conditions be relevant. Test validity is always concerned with not only what a test measures, but also how well it does so. Considering this point, one can say that what a test measures can be determined only by an examination of the procedures employed in finding out the validity of the test, and especially by the nature of the criterion. A test, therefore, is said to be valid to the extent that it measures what it purports to measure.

Since tests are designed for a variety of purposes, and since validity can be evaluated only in terms of purpose, it is not surprising that there are several different types of validity: content, construct, concurrent and predictive.

3.11.2 Types of Validity

(1) Content Validity

Content validity is the degree to which a test measures an intended content area. Content validity requires both item validity and sampling validity. Item validity is concerned with whether the test items represent measurement in the intended content area, and sampling validity is concerned with how well the test samples content area.

The term “face validity” is sometimes used in describing tests. While its meaning is somewhat ambiguous, basically face validity refers to the degree to which a test appears to measure what it purports to measure. While determining face validity is not a psychometrically sound way of estimating validity, the process is sometimes as an initial screening procedure in test selection.
Content validity is of prime importance for achievement tests. A test score cannot accurately reflect a student’s achievement if it does not measure what the student was supposed to learn. While this seems obvious, content validity has been a problem in a number of research studies.

Content validity is determined by expert judgment. There is no formula by which it can be computed and there is no way to express if quantitatively. Usually experts in the area covered by test are asked to assess its validity. These experts carefully review the process used in developing the test as well as the test itself and make a judgment is based on whether all subjects areas have been included, and in the correct proportion.

(2) Construct Validity

Construct validity is the degree to which a test measures an intended hypothetical construct. A construct is a non observable trait. Such as intelligence, that explains behavior. You cannot see a construct, but you can only observe its effect. In fact construct were “invented” to explain behavior. We cannot prove they exist; we cannot perform brain surgery on a person and “see” his or her intelligence. Constructs, however, do an amazingly good job of explaining certain differences between individuals. For example, it was always observed that some students learn faster than others, learn more, and retain longer. To explain these differences, a theory of intelligence was developed, and it was hypothesized that there is something called intelligence which is related to learning and which everyone possesses to a greater or lesser degree.

Research studies that involve construct, either as an independent or a dependent variable, are only valid to the extent that the measure of the construct involves is valid. Anxiety, for example can be an independent or a dependent variable. A study might be designed to determine whether high-anxiety students perform better on difficult tasks than low anxiety students. A test of anxiety would need to be administered to the students in the study in order to classify them as “high anxiety” or “low-anxiety”. If the test did not really measure anxiety, conclusions based on a study utilized it would be meaningless. When selecting a test of a given construct, the researcher must look for and critically evaluate evidence presented related to the construct validity of the instrument.

The process of validating a test of a construct is by no means an easy task. Basically, it involves testing hypothesis deduced from a theory concerning the construct. If, for
example, a theory of anxiety hypothesized that high-anxiety persons will work longer on a problem than low-anxiety persons, than if persons was called high on the test under consideration did indeed work longer on a subsequent task. This would be evidence to support the construct validity of the test. Of course if the high-anxiety persons did not, as hypothesized, work longer, then it would not necessarily mean that the test did no measure anxiety; the hypothesis related to the behavior of high-anxiety persons might be incorrect. Generally, a number of independent studies are required to establish the credibility of a test of a construct.

(3) Concurrent Validity

Concurrent validity is the degree to which the scores on a test are related to the scores on another, already established test which is administered at the same time or to some other valid criterion available at the same time. If the developed test claims to do the same job of some other tests, easier or faster that is concurrent validity of the new test.

Concurrent validity is determined by establishing relationship or discrimination. The relationship method involves determining the relationship between scores on the test and scores on some other established test or criterion (e.g. GPA). In this case, the steps involved in determining concurrent validity are as follows.

1. Administer the new test to a defined group of individuals.

2. Administer a previously established, valid test (or acquire such scores if already available) to the same group, at the same, or shortly thereafter.

3. Correlate two sets of scores.

4. Evaluate the results.

The resulting number, or validity coefficient, indicates the concurrent validity of the new test. If the coefficient is high, the test has good concurrent validity.

(4) Predictive Validity

Predictive validity is the degree to which a test can predict how well an individual will do in a future situation. An algebra aptitude test that has high predictive validity will fairly accurately predict which students will come in algebra and which students will not. Predictive validity is extremely important for tests that are used to classify or select
individuals. An example with which you all are familiar is the use of Graduate Record Examination (GRE) scores to select students for admission, often 1,000 in the belief that students who achieve that score have higher probability of succeeding in graduate school. The predictive validity of the GRE has been the subject of many research studies. Results seem to indicate that the GRE has higher predictive validity for certain areas of graduate study than for others.

As the GRE example illustrates, the predictive validity of a given instrument varies with a number of factors. The predictive validity of an instrument may vary depending upon such factors as the curriculum involved, textbooks used, and geography location.

No test, of course, has perfect predictive validity. Therefore, predictions based on the scores of any test will be imperfect. However, predictions based on a combination of several test scores will invalidity be more accurate than predictions based on the scores of any test. Therefore, when important classification or selection decisions are to be made, they should be based on data from more than one indicator.

The procedure for determining predictive validity is as follows:

1. Administer the test, the predictor variable, to a group.
2. Wait until the behavior to be predicted, the criterion variable, occurs.
3. Obtain measures of the criterion for the same group.
4. Correlate the two sets of scores.
5. Evaluate the results.

Closely related to the concept of validity is the concept of reliability, which deals with the question of score consistency.

### 3.12 Reliability

#### 3.12.1 Concept of Reliability

There are several meanings attached to the term 'reliability. It includes consistency, trustworthiness, dependability, objectivity and so on. Each signifies something different as applied to measurement. A test is said to be reliable if it gives the same result on
different occasions of the abilities of those to whom it is applied, it means the self-consistency with which it works at all times.

Rammers and Cage (1967) define it as:

"Reliability is the consistency with which a test yields the same result of measuring whatever it does measure."

David P. Harris (1969) explains the term as follows:

"By reliability it is meant the stability of test scores. It is an essential aspect of a useful test."

Robert Lado (1967) supporting the above concept writes as,

"Reliability has to do with the stability of scores for the same individuals. If the scores of the students are stable, the test is reliable, if the scores tend to fluctuate for no apparent reason, the test is unreliable."

W.A. Mehrens (1969) defines as:

"Reliability can be defined as the degree of consistency between two measures of the same thing."

Guilford (1954) writes:

"Reliability is the proportion of true variance in obtained test scores."

Anastasi (1970) explains the term as:

"Reliability refers to the consistency of scores obtained by the same individual when re-examined with the same test on different occasions."

Frank Freeman (1965) says:

"Reliability refers to the extent to which a measuring tool yields consist results upon different occasions."

Thus, it can be said that the test must be reliable, and a test is said to be reliable only if the correlation between the test scores at both the times of its administration is high. Otherwise it remains a poor test, a weak test. It is said that if a test lacks its reliability, it
is useless as a thermometer that gives different readings when the temperature of the air is the same.

Reliability is expressed numerically, usually as a coefficient; a high coefficient indicates high reliability. If test were perfectly reliable, the coefficient would be 1.00; this would mean that a student’s score perfectly reflected her or his true status with respect to the variable being measured. However, alas and alack, no test is perfectly reliable. Score are invariably affected by errors of measurement resulting from a variety of cause. High reliability indicates minimum error variance; if test has high reliability, then the effect of errors of measurement has been reduced. Errors of measurement affect score in a random fashion; some scores may be increased while other are decreased. Errors of measurement can be caused by characteristics of the test itself (ambiguous test items, for example, that some students just happen to interpret correctly), by conditions of administration or for example, directions not properly followed, by the current status of persons taking the test (some may be tired, others unmotivated), or by a combination of any of the above. High reliability indicates that these sources of error have been eliminated as much as possible.

Errors of measurement that affect reliability are random errors; systematic or constant errors affect validity. If an achievement test was too difficult for a given group of students, all scores would be systematically, lowered; the test would have low validity for the group (remember “valid whom”). The test might, however, yield consistent scores, i.e. might be reliable; in other words, the scores might be systematically lowered in the same way every time. A given student whose “true” achievement score was 80 and who scored 60 of the test (invalidity) might score 60 every time he took the test (reliability). This illustrates an interesting relationship between validity and reliability. A valid test is always reliable but a reliable test is not necessarily valid. In other words, if a test is measuring what it is supposed to be measuring, it will be reliable and do so every time, but a reliable test can be consistently measure the wrong thing and be invalid.

Reliability is much easier to assess than validity. There are a number of different types of reliability; each is determined in a different manner and each deal with a different kind of consistency.

3.12.2 Types of Reliability

(1) Test Retest Reliability:
Test-retest reliability is the degree to which scores are consistent over time. It indicates score variation that occurs from testing session to testing sessions as a result of errors of measurement. In other words, we are integrated in evidence that the score a person obtains on a test at some moment in time is the same score, or close to the same score, that the person would get if the test were administrated some other time. We want to know how consistently the test measures whatever it measures. This type of reliability is especially important for tests used as predictors, aptitude tests, for example.

Determination of test-retest reliability is appropriate when alternate (equivalent) forms of a test are not available, and when it is unlikely that persons taking the test the second time will remember responses made on the test the first time. Test takers are more likely to remember items from a test with a lot of history facts, for example, than from a test with algebra problems. The procedure for determining test-retest reliability is basically quite simple:

1. Administer the test to an appropriate group.
2. After some time has passed, say a week, administer the same test to the same group.
3. Correlate the two sets of scores.
4. Evaluate the results.

If the resulting coefficient, referred to as the coefficient of stability, is high, the test has good test-retest reliability. A major problem with this type of reliability is the difficulty of knowing how much time should elapse between the testing sessions. If the interval is too short, the chances of student’s remembering responses made on the test for the first time are increased, and the estimated reliability tends to be artificially high, if the interval is too long, student’s ability to do well on the test may increase due to intervening learning or maturation, and the estimate of reliability tends to be artificially low.

Thus, when test-retest information is given concerning a test, the time interval between testing should be given as well as the actual coefficient. Although it is difficult to say precisely, what, in general, the ideal time interval should be, especially since depends somewhat on the kind of test involved, one day will generally be too short and one month
too long. The problems associated with test-retest reliability are taken care of by equivalent forms reliability.

(2) **Equivalent-forms Reliability:**

Equivalent-forms of a test are two tests that are identical in every way except for the actual items included. The two forms measure the same variable; have the same number of items, the same structure, the same difficulty level, and the same directions for administration, scoring interpretation. In fact, if the same group takes both the tests, the average score as well as the degree of score variability should be essentially the same on both tests. Only the specific items are not the same, although they do measure the same items, or objectives. In essence we are selecting, or sampling, different items from the same behavior domain. We are interested in whether scores depend upon the particular set of items or whether performance of one set of items is general to other sets. If items are well selected, and if each set adequately represents the domain of interest, the latter should be true.

Equivalent-forms of reliability, also referred to as alternate forms reliability, indicates score variation that occur from to form, and is appropriate when it is likely that takers will recall responses made during the first session and, of course, when two different forms of a test are available. When alternate forms are available, it is important to know the equivalent-forms, it is reassuring to know that person’s score will not be greatly affected by which form is administered. Also, sometimes in research studies forms of a test are administered to the same group, on as a pretest and the other as a post test. It is crucial, if the effects of the intervening activities are to be validity assessed, that the two tests be measuring essentially by the same things.

The procedure for determining equivalent-forms reliability is very similar to that for determining test-retest reliability:

1. Administer one form of the test to an appropriate group.

2. At the same session, or shortly thereafter, administer the second form of the test to the same group.

3. Correlate the two sets of scores.

4. Evaluate the results.
If the resulting coefficient (referred to as the coefficient of equivalence) is high, the test has good equivalent-forms reliability. If the two forms of the test are administered at two different times (the best of all possible world’s the resulting) coefficient is referred to as the coefficient of stability and equivalence. In essence this approach represents a combination of test-retest and equivalent forms reliability and thus assesses stability of scores over time as well as the generalizability of the sets of items. Since more sources of measurement error are possible than with either method alone, the resulting coefficient is likely to be somewhat lower. Thus the coefficient of stability and equivalence represents a conservative estimate of reliability.

Equivalent-form reliability is the single most acceptable and most commonly used estimate of reliability for most tests used in research. The major problem involved with this method of estimating reliability is the difficulty of constructing two forms that are essentially equivalent lack of equivalence is a source of measurement error. Even though equivalent-forms reliability is considered to be best estimate of reliability, it is not always feasible to administer two different forms of the same, or even the same test twice. Imagine telling your students that they had to take two final examinations! Imagine someone telling you take the GRE or SAT twice! Fortunately, there are other methods of estimating reliability that require administering a test only once.

(3) Split-half Reliability:

A common type of internal consistency reliability is referred to as split-half reliability. Since split-half reliability procedures require only one administration of a test, certain sources of errors of measurement are eliminated, such as differences in testing conditions, which can occur in establishing test-retest reliability. Split-half reliability is especially appropriate when a test is very long.

The procedure for determining split-half reliability is as follows:

1. Administer to total test to a group.
2. Divide the test into comparable halves, or sub tests-the most common approach is to include all odd items in one half and all even items in the other half.
3. Compute each subject’s score on the two halves-each subject will consequently have two scores, a score for the odd items and score for the even items.
4. Correlate the two sets of scores.

5. Evaluate the results.

If the coefficient is high, the test has good split-half reliability. A number of logical and statistical methods can be used to divide a test in half, random selection of half of the items, for example. In reality, the odd-even strategy, however, is most often used. Actually, this approach works out rather well regardless of how a test is organized. Suppose, for example, a test is a 20-item power test and the items get progressively more difficult. Items 1,3,5,7,9,11,13,15,17 and 19 as group should be approximately as difficult as items 2, 4,6,8,10,12,14,16,18 and 20. Items 1 and 2 will easy, 3 and 4 will be more difficult and so forth so forth. Or, suppose a test is organized by topic so that items 1-10 will be circles and items 11-0 deals with quadrilaterals. In this case the odd items will contain items on circles and quadrilaterals will contain even items. Thus, regardless of how the test is organized, an odd-even split should produce essentially equivalent halves.

In fact, what we are doing, in artificially creating two equivalent forms of a test and computing equivalent forms reliability; the equivalent forms just happens to be in the same test. Thus it also labels ‘internal consistency reliability’.

Since longer test tend to be more reliable, and since split-half reliability represents the reliability of a test only half as the actual test, a correction formula must be applied to the coefficient. The correction formula which is used is the Spearman-Brown prophecy formula. For example, suppose the split-half reliability coefficient for a 50 item test were 80. It would be based on the correlation between score on 25 even items and 25 odd items and would therefore be an estimate of the reliability of a 25-item test, not a 50-item test. The Spearman Brown formula would need to be applied to estimate the reliability (r) of the 50-item test. The formula is a simple one, even for those of you who are not mathematically inclined:

\[
\Gamma_{\text{total test}} = \frac{2 \Gamma_{\text{split half}}}{1 + \Gamma_{\text{split half}}}
\]

Applying the formula for example:
\[ \Gamma_{\text{total test}} = \frac{2 \times (0.80)}{1 + 0.80} \]
\[ = \frac{1.60}{1.80} \]
\[ = 0.89 \]

Thus, the split-half estimate of 80 was corrected to an estimate of 0.89. One problem with the correction formula is that it tends to give a higher estimate of reliability than it would be obtained using the procedures.

Another approach to determining internal consistency is the method of rationale equivalence.

**(4) Rationale Equivalence Reliability** *(Kuder Richardson’s Method):*

Rationale equivalence reliability is not established through correlation but rather estimates internal consistency by determining how all items on a test relate to all other items of the test. Rationale equivalence reliability is determined through application of one of the Kuder-Richardson formula, usually formula 20 or 21 (KR-20 or KR-21). Application of a Kuder-Richardson formula result in an estimate of reliability that is essentially equivalent to the average of the split-half reliabilities computed for all the possible halves. Use of formula 21 require less time than any other method of estimating reliability. Its application also usually results in a more conservative estimate of reliability. Especially if more than one trait is being measured.

The formula is as follows:

\[ (1) \Gamma_{\text{total test}} = \frac{(K)(SD^2) - X(K-X)}{(SD^2)(K-1)} \]

Where,
- \( K \) = the number of items in the test
- \( SD \) = the standard deviation of the scores
- \( X \) = the mean of the scores
(2) \[ r_{\text{total test}} = \frac{n}{n-1} \times \frac{\sigma t^2}{\sum pq} \]

Where,

- \( r_{\text{total test}} \) = reliability of coefficient of the whole test
- \( n \) = number of items in the test
- \( \sigma t \) = SD of the test score
- \( p \) = the proportion of the group answering a test item correctly
- \( q = (1-p) \) the proportion of the group answering a test item incorrectly.

Summarized this, (1) test administration indicates either that only one test is administered (split-half and KR-21 reliability) or that two tests are administered at essentially the same time (equivalent-forms reliability). Two (2) administration times indicates a time interval exists (say a week) between the two administrations (test-retest and stability and equivalence reliability). You should keep in mind when reviewing test information that while the size on the reliability coefficient is of prime importance, the method used to calculate it should also be considered.

(5) **Score/ Rater Reliability:**

There are other situations for which reliability must be investigated. Such situation occur when the scoring of tests involves subjectivity, such as with essay test short answer tests involving more than a one-word response, rating scales, and observation instruments. In such situations we are concerned with inter judge (inter scorer, interpreter, observer) reliability and/or intra judge reliability. Inter judge reliability refer the reliability of two (or more) independent scorers; intra judge reliability refers to ability of the scoring of individual scorers. Scoring and rating are sources of errors of measurement, and it is important to estimate the consistency of scores’ assents. Estimate of inter judge or inter judge reliability are usually obtained using correlation techniques, as has already been discussed, but can also be expressed simply as present agreement. While such reliabilities are unfortunately usually not very good, a number of standardized instruments have been
developed to the point where inter judge and intra judge reliability appears to be relatively quite good.

(6) **Reliability Coefficients:**

What constitutes an acceptable level of reliability is to some degree determined by the type of test although, of course, a coefficient over 90 would be acceptable for the test. The question really is concerned with what constitutes a minimum level of acceptability. For achievement and aptitude tests, there is generally no good reason for selecting a test whose reliability is not at least 90. There are number of achievement and aptitude tests available that report such reliabilities and it is therefore not usually necessary to settle for less. Personality measures do not typically report such high reliability (although certainly some do) and one would therefore be satisfied with reliability in the eighties and might even accept reliability in the seventies.

### 3.13 Norms

Establishment of norms is also an essential step, because obtained row scores of any test are meaningless. Raw score is a numerical expression of performance on the part of an individual. These raw scores cannot be interpreted easily. For easy interpretation and for easy comparison, test scores should be made relative.

Norms means the average performance of average group of individuals. Bruce Tuckmen (1972) defines the terms ‘Norms’ in the following words:

“Norms are information about the relative performance of a specific group of people that is variable.”

One more opinion will help to clear concept W.S. Manoe (1957) opines as under:

“In fact our very language and thinking involves and requires such terms as ‘typical’, ‘much’ etc, imply the recognition of norms. It is obvious then that norms are of basic importance.”

Referring to such various definitions of norms it can be easily conducted that, Norms are level of performance in a test attained by defined group of population.
To use norms as a standard comparison the investigator is expected to convert raw scores into percentile ranks, standard scores or t-scores. Hence these norms are:

1. Percentile Norms
2. Standard score Norms or
3. T Score Norms.

To use norms in the sense of control tendency of the score of a group, separate scores for different specific groups armed by difference in sex, class or any other variable. Hence there following norms are known as:

1. Sex norms
2. 100 grade norms
3. Age norms
4. Area norms etc.

In the present study the investigator has computed the Percentile Norms and T scores to establish norms based on the total scores achieved by the sample. The investigator has given the details about the Percentile Norms and T scores in the next chapter.

### 3.14 Conclusion

After discussing the theoretical aspects of test construction including the planning and procedures of conducting the entire study, the investigator is now in a position to take up the discussion of the actual test construction.
References


Ibid, p.349.
