Dr. Suswati Hendriani, M. Pd., M. Pd
Nina Suzanne, M. Pd

LANGUAGE TESTING

STAIN Batusangkar Press
2013

Perpustakaan Nasional: Katalog dalam Terbitan (KDT)

Suswati Hendriani
Nina Suzanne

Language Testing

Cet 1.-Batusangkar: STAIN Batusangkar Press, 2013

v + 135 hlm. ; 21 x 15 cm

ISBN : **978-602-8887-69-4**

1. *Language Testing*

1. Judul

Cetakan Pertama, Agustus 2013

Language Testing

Penulis
Suswati Hendriani
Nina Suzanne

Perwajahan Isi & Penata Letak
Marhen

Desain Cover
Marhen

Penerbit

STAIN Batusangkar Press, 2013
Jl. Sudirman No. 137 Lima Kaum Batusangkar
Telp : (0752) 71150, 574221,574227,71890,71885.
Fax : (0752) 71879
Web : www.stainbatusangkar.ac.id
e-mail : press@stainbatusangkar.ac.id

PREFACE

This book offers some aspects in language testing that consists of definition, characteristics, kinds and forms of language testing. The purpose of this book is to give a straightforward guide to assessments for language students as candidate teachers and for anybody who wants to get more understanding about testing and assessment.

There are six chapters in this book:

1. Teaching, Assessment, and Testing

   This chapter will give an overview about the relationship of teaching, assessment, and testing.
2. Characteristics of a Good Test

   This chapter will consist of detail explanation about validity, reliability, and practicality.
3. Constructing and Administering Language Test

   The readers can find some principles and guidance to construct and administer the language testing in chapter III.
4. Types of Test

   There are four types of test that will be discussed in this chapter, they are: objective, performance, oral and essay.
5. Testing Language Skills and Components

   This chapter will offer some principles and kinds of tasks giver to test the language skills (listening, speaking, reading, and writing), and components (grammar and vocabulary).
6. Assigning Grades and Course Marks

   This chapter will give explanation about scoring, grading, and giving marks.

The writers are grateful to all people which cannot be mentioned one by one here that enabled them to work on this book.

The Writers

TABLE OF CONTENT

UNIT I

TEACHING, ASSESSMENT, AND TESTING

A. Introduction to Language Assessment and Testing

It is helpful for students especially those who take educational major (candidate teachers) to be knowledgeable about educational assessments because the results are used widely for selection, certification, diagnosis, or placement. Students should also have knowledge that tests, exams, quizzes, projects, assignments, or portfolios as a part of their teaching face development. The ability to develop worthwhile assessments does not come naturally but is a skill that can be acquired. That's why, candidate teachers should understand this well which can support their professional job field in the future. It has been estimated that teachers spend as much as one-third of their time in assessment-related activities.

1. Teaching and Assessment

Assessment plays an important role in teaching. The assessment may occur informally during the instruction (teaching and learning process) or at the end of instruction as formal assessment. Both assessments are useful to find out

the students' progress, achievement, and even career's prospects in the future.

High quality teaching or successful teaching demands accurate, comprehensive assessment of learning at the beginning, during the year, and at the end of the year. Obviously, the measurement objectives must coincide with instructional objectives. A teacher has a crucial role on this point.

Athanasou and Lamprianou (2002:3) in their book define assessment as the process or processes of collecting and combining information from tasks (e.g., tests on performance or learning) with a view to making a judgment about a person or making a comparison against an established criterion. From the definition, it can be understood that teachers are free to include many different types of tasks (assignments, exercises, projects, quizzes, etc) under the umbrella of assessment. The assessment itself is used daily by teachers and students to refer in a shorthand type of way to a particular task that may have been designed.

In all fields of education, the purpose of assessment is to decide about:

- Students (students' progression).
- Teaching and learning (curriculum decisions).

Increasingly assessments will be linked with certification competence and the validation of performance on job-related tasks.

For instance, assessments are given to find the answers for some questions, such as:

- How realistic are my teaching plans for these students?
- Are my students ready for the next unit?
- What learning difficulties are students facing?
- Which students are underachieving?
- How effective was my teaching?

- Which learners are advanced?
- Which learners are gifted or talented?
- Which learners require special assistance?

The assessment process is an integral aspect of education that helps teachers to make judgments about students' current levels, about the most appropriate method for teaching, or when to conclude teaching a topic.

Because of the importance of assessments, teachers or candidate teachers should prepare and construct the assessment well. Teachers should be able to create an effective and appropriate assessment, because it will help people or students to learn in a way that is meaningful and encourage their motivation to learn.

2. Assessment and Test

Assessment is a popular and sometimes misunderstood term in current educational practice. You might be tempted to think of testing and assessing as synonymous terms, but they are not.

In the last decade, the word "assessment" has taken over from terms such as "testing." Firstly, it was seen as a broader term than "test" because it encompasses many different educational practices, such as portfolios, case studies or presentations. Secondly, it also took into account divergent processes of assessment such as teacher assessment, self-assessment and peer-assessment. Thirdly, it gave some expression to more liberal views in education that were opposed to the oppressive, mechanical and unthinking use of tests.

Assessment is an ongoing process that encompasses a much wider domain. Tests, on the other hand, are a subset of assessment. They are prepared administrative procedures that occur at identifiable times in a curriculum when learners

muster all their faculties to offer peak performance, knowing that their responses are being measured and evaluated.

Tests are a subset of assessment; they are certainly not the only form of assessment that a teacher can make. Tests can be useful devices, but they are only one among many procedures and tasks that teachers can ultimately use to assess students.

In simple terms, a test is a method of measuring a person's ability, knowledge, or performance in a given domain. Let's look at the components of this definition.

- A test is a method: it is an instrument ─ a set of techniques, procedures, or items ─ that requires performance on the part of the test-taker.
- A test must measure: Some tests measure general ability, while others focus on very specific competencies or objectives.
- A test measures an individual's ability, knowledge, or performance.
- A test measures performance, but the results imply the test-taker's ability, or to use a concept common in the field of linguistic, competence.

3. Kinds of Assessment

Assessment can be divided into:

a. Informal and Formal Assessment

Informal assessment can take a number of forms, starting with incidental, unplanned comments and response, along with coaching and other impromptu feedback to the student. Example: "Nice job!" "Good work!" Did you say can or can't?" I think you meant broke the glass, not you break the glass", or putting (a smile character) on some homework or assignment.

On the other hand, formal assessments are exercises or procedures specifically designed to tap into a storehouse of skills and knowledge. They are systematic, planned sampling techniques constructed to give teacher and student an appraisal of student achievement.

Is formal assessment the same as a test? We can say that all tests are formal assessments, but not all formal assessment is testing. For example, you might use a student's journal or portfolio of materials as a formal assessment of the attainment of certain course objectives, but it is problematic to call those two procedures "tests." A systematic set of observations of a student's frequency of oral participation in class is certainly a formal assessment, but it too is hardly what anyone would call a test. Tests are usually relatively time-constrained (usually spanning a class period or at most several hours) and draw on a limited sample of behavior.

b. Formative and Summative Assessment

Most of our classroom assessment is formative assessment: evaluating students in the process of "forming" their competencies and skills with the goal of helping them to continue that growth process. The key to such formation is the delivery (by the teacher) and internalization (by the student) of appropriate feedback on performance, with an eye toward the future continuation (or formation) of learning.

For all practical purposes, virtually all kinds of informal assessment are (should be) formative. They have as their primary focus the ongoing development of the learner's language. So when you give a student a comment or a suggestion, or call attention to an error, that feedback is offered in order to improve the learner's language ability.

Summative assessment aims to measure, or summarize what a student has grasped, and typically occurs at the end of a course or unit of instruction. Final exams in a course and general proficiency exams are examples of summative assessment.

B. Purposes and Methods of Language Testing

Language tests have many uses in educational programs, and quite often the same test will be used for two or more related purposes. Generally, classroom tests are constructed, administered and scored by the teacher, so the objectives of the tests usually are based on the course objectives.

Before giving a language test, a teacher should decide the purpose or the function of the test. The following list summarizes the chief objectives of language testing which emphasized in measuring student's ability or achievement (Harris, 1969):

1. To determine readiness for instructional programs
2. To classify or place individuals in appropriate language classes.
3. To diagnose the individual's specific strengths and weaknesses.
4. To measure aptitude for learning.
5. To measure the extent of student achievement of the instructional goals.
6. To evaluate the effectiveness of instruction.

For simplicity, the foregoing six categories can be grouped under three headings:

a. Aptitude (category 4 above). An aptitude test serves to indicate an individual's facility for acquiring specific skills and learning.

b. General proficiency (categories 1 to 3). A general proficiency test indicates what an individual is capable of doing now (as the result of his cumulative learning experiences), though it may also serve as a basis for predicting future attainment.

c. Achievement (categories 5 and 6). An achievement test indicates the extent to which an individual has mastered the specific skills or body of information acquired in a formal learning situation.

In line with above quotation, Hughes (1989:7) mentions some purposes of testing as follow:

- ✓ To measure language proficiency regardless of any language courses that candidates may have followed.
- ✓ To discover how far students have achieved the objectives of a course of study.
- ✓ To diagnose students' strengths and weaknesses, to identify what they know and what they do not know.
- ✓ To assist placement of students by identifying the stage or part of a teaching program most appropriate to their ability.

Tests are used to obtain information. The information will vary from situation to situation. Tests can be categorized according to a small number of kinds of information being sought. Further information about the kinds of tests will be given below.

C. Direct versus Indirect Testing (approach to test construction)

Testing is said to be *direct* when it requires the candidate to perform precisely the skill which we wish to measure. If we want to know how well candidates can write compositions, we get them to write compositions. If we want

to know how well they pronounce a language, we get them to speak.

Direct testing is easier to carry out when it is intended to measure the productive skills of speaking and writing. The very acts of speaking and writing provide us with information about the candidate's ability. With listening and reading, however, it is necessary to get candidates not only to listen or read but also to demonstrate that they have done this successfully.

Direct testing has a number of attractions: a) provided that we are clear about just what abilities we want to assess, it is relatively straight forward to create the conditions which will elicit the behavior on which to base our judgments. b) At least in the case of the productive skills, the assessment and interpretation of students' performance is also quite straightforward. c) Since practice for the test involves practice of the skills that we wish to foster, there is likely to be a helpful backwash effect.

Indirect testing attempts to measure the abilities which underlie the skills in which we are interested. One section of the TOEFL, for example, was developed as an indirect measure of writing ability. It contains items where the candidate has to identify which of the underlined elements is erroneous or inappropriate in formal standard English.

The main problem with indirect tests is that the relationship between performance on them and performance of the skills in which we are usually more interested tends to be rather weak in strength and uncertain in nature.

D. Discrete Point versus Integrative Testing

This historical perspectives underscores two major approaches to language testing that were debated in the 1970s and early 1980s. these approaches still prevail today, even if

in mutated form: the choice between discrete-point and integrative testing methods.

What does an integrative test look like? Two types of tests have historically been claimed to be examples of integrative tests: cloze tests and dictations.

Discrete point testing refers to the testing of one element at a time, item by item. This might involve a series of items each testing a particular grammatical structure. Integrative testing, by contrast, requires the candidate to combine many language elements in the completion of a task. This might involve writing a composition, making notes while listening to a lecture, taking a dictation, or completing a cloze passage. Discrete point tests will almost always be indirect, while integrative tests will tend to be direct. However, some integrative testing methods, such as the cloze procedure, are indirect.

E. Norm-referenced versus criterion-referenced testing

It is important to differentiate between norm-referenced tests and criterion-referenced test. Norm-referenced relate one candidate's performance to that of other candidate. For example, a student took a test. Then, the student obtained a score that placed her or him in the top ten percent of all candidates who have taken the test. In norm-referenced tests, each test taker's score is interpreted in relation to a mean, median, standard deviation, and/or percentile rank. Typical of this kind of test is SAT (Scholastic Aptitude Test), and TOEFL (Test as a Foreign Language).

Criterion-referenced tests are designed to give test takers feedback, usually in the form of grades, on specific course or lesson objectives. Classroom tests involving the students in only one class, connected to a curriculum. This kind of test is designed to provide information about what a student can actually do in the language without comparing to

other candidates. The purpose of criterion-referenced tests is to classify people according to whether or not they are able to perform some task or set of tasks satisfactorily.

F. Communicative Language Testing

By the mid-1980s, the language testing field had abandoned arguments about the unitary trait hypothesis and had begun to focus on designing communicative language-testing tasks. Bachman and Palmer (1996, p.9) include among "fundamental" principles of language testing the need for a correspondence between language test performance and language use.

And so a quest for authenticity was launched, as test designers centered on communicative performance. Communicative testing presented challenges to test designers. Test constructors began to identify the kinds of real-world tasks that language learners were called upon to perform.

Communicative language testing presented challenges to test designers. Test constructors began to identify the kinds of real-world tasks that language learners were called upon to perform. The assessment field became more and more concerned with the authenticity of tasks and the genuineness of texts.

UNIT II

CHARACTERISTICS OF A GOOD TEST

All measuring instruments possess, to some degree, three important qualities (characteristics): (1) validity, (2) reliability, and (3) usability (practicality). Of course, other qualities are important, but these three are the essentials without which a testing instrument would be useless.

A. Validity

Of the three qualities mentioned above, validity is the most important. The term validity refers to an instrument's truthfulness, appropriateness, or accuracy. A valid instrument is truthful because it measures what the person using the instrument wishes or attempts to measure.

In *the Standards for Educational and Psychological Testing* in Athanasou and Lamprianou (2002: 167), the definition of validity is "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test." The degree of validity is an inference that requires several lines of evidence.

Furthermore, validity is specific, since an instrument has validity only for the purpose for which it was intended.

This can be illustrated by the fact that an English grammar test probably has virtually no validity for measuring algebra.

The purpose and the educational objectives to be assessed determine the choice of measurement instrument. Thus, when the wrong instrument is selected, validity will be low. Even when the correct test is selected, validity will be low if the test is poorly constructed or if it has a large sampling error because it consists of too few test items or of unrepresentative test items.

There are two errors, which teachers may make, can reduce the validity of measurement: (error in the choice of an instrument and (2) error in item sampling.

There have been generally been two approaches to the determination of the relative validity of achievement tests:

1. The curricular approach
   In the curricular approach, which is actually a rational approach, it is assumed that the curriculum in the specific field as represented by textbooks, courses of study, and expert opinion is valid, and the test content is compared against these criteria to determine its validity. A teacher, as the expert, examines his course outline and his teaching objectives to determine the degree to which they coincide with test content.
2. The statistical approach
   When the statistical approach is used to validate a test, a criterion which is presumed to have validity is selected, and the test scores of a group of examinees are correlated with that criterion. Criteria that are used in this validation process include school marks, judgment of experts, and acceptable tests. Statistical validation is not often used for teacher-made tests; nevertheless, a teacher certainly must note such data as other school marks and performance on standardized tests for comparison with the performance of examinees on his own tests.

1. Types of Validity

There are several types of validity, but among them there are three which are commonly described: contents, construct, and criterion validity. Each type of validity seeks to answer a separate question, such as:

- ✓ Content validity: does the assessment match the content and learning outcomes of the subject?
- ✓ Construct validity: does the assessment really involve the particular behaviors, thought processes or talents that are said to be assessed?
- ✓ Criterion validity: does the assessment provide evidence of future achievement or performance in other related subjects?

Here are the explanations for each type of validity:

a. Content Validity

A teacher's primary concern for a classroom test is with content validity. Content validity refers to the degree to which an instrument samples the subject matter in the area to be measured or the degree to which it coincides with the instructional objectives which are to be measured in a given field. In this term, the test-takers are required to perform the behavior that is being measured. For example, assessing a person's ability to speak a foreign language in a conversational setting is by asking him/her to speak within some sort of authentic context can achieve the content validity. But, asking the learner to answer paper-and-pencil multiple-choice questions requiring grammatical judgments doesn't achieve content validity.

Content validity is sometimes confused with what is called "face" validity. Face validity is the appearance

of the test and its relevance to the person taking the test. For instance, test-takers are perplexed when the content does not appear to match the subject area. Face validity is not content validity but is related.

There are two essential aspects that need to be considered by testers related to content validity; *first*, testers need to ensure that all questions are asked in a way that is familiar for the learner and consistent with the subject; *two*, testers need to judge whether the assessment adequately samples the topics and learning outcomes in the subject.

b. Criterion-related validity

Criterion-related validity is the extent to which the "criterion" of the test has actually been reached. In the case of teacher-made classroom assessments, criterion-related evidence is best demonstrated through a comparison of results of an assessment with results of some other measure of the same criterion. Criterion validity is commonly indicated by correlation coefficients.

There are two categories of criterion-related evidence:

1). Predictive Validity

Predictive validity is present in an evaluative instrument or technique if relative success of the student can be predicted accurately from the score or rating obtained. It is important in the case of placement tests, language aptitude tests, and the like. It is used to assess and predict a test-taker's likelihood of future success.

The predictive validity of results indicates how well the assessment results can predict future performance of the students/testee. In education, there is little evidence of the predictive validity of our assessments since this involves long-term follow-up studies of pupils who have completed schooling, students who have graduated or people who have completed courses.

Example of predictive validity:

- A teacher wants to determine whether the results of a computer programming selection test predicts success in his/her class. The end-of-year test results for a certificate course in computing are used as the predictive criterion.

2). Concurrent Validity

Concurrent validity is a characteristic that the measuring instrument or evaluative technique must have in order to determine the current status of a pupil; it is concerned with present student behavior. A test has concurrent validity if its results are supported by other concurrent performance beyond the assessment itself.

Concurrent validity is the extent to which results estimate present performance on some other task or assessment. Different from predictive validity, information on the concurrent validity of assessments is easier to obtain since it involves examining the performance of students on other subjects in a course, or comparison with work placements or on-the-job training.

c. Construct validity

Construct validity indicates the qualities a test measures. Constructs that may adversely influence the performance of certain pupils include desire to achieve, response set, reading ability, competitiveness, and poor test psychology. The pupil who has little desire to achieve, or who has a poor test psychology, or who is uncompetitive may perform much below his optimum level. Furthermore, a pupil who knows the answers to some questions but misunderstands them, possibly because of his culture bias or his poor reading ability, will probably answer incorrectly.

Construct-related validity refers to the theoretical evidence for what we are measuring. How can you be sure that a final exam in marketing is really assessing "marketing" and not scholastic aptitude or examination ability?

Construct validity studies can involve:

- Internal consistency analysis of the questions in a test to see whether a single aspect is being assessed.
- Analysis of results over time to trace changes in student development of knowledge, skills or attitudes.
- Checks to see whether graduates or workers in an occupation perform better than novices or students.
- Factors analysis of the items.
- The correlation between assessments results and other related assessments of knowledge, skills or attitudes.

In construct validation, testers have to identify and describe the general characteristics they are assessing (e.g., numerical skills, literacy, communication competence, mechanical competence, word processing skills, design creativity). Then, testers also need to find evidence to support the claim that the assessment is directed to what is intended.

d. Face Validity

Face validity refers to the degree to which a test looks right, and appears to measure the knowledge or ability it claims to measure, based on the subjective judgment of the examinees.

Face validity will likely be high if learners encounter:

- A well-constructed, expected format with familiar tasks.
- A test that is clearly doable within the allotted time limit.
- Items those are clear and uncomplicated.
- Directions those are crystal clear.
- Tasks that relate to their course work (content validity)
- A difficulty level that presents a reasonable challenge.

This type of validity is not something that can empirically tested by a teacher or expert. It is purely a factor of the "eye of the beholder", how the test takers intuitively perceive the instrument.

2. Factors Reducing Validity

The most important factor for a classroom teacher is to ensure that the assessment given has content

validity. The validation of test results should be a continuing process. There are some questions to do it:

- Does the content sample the topics in my syllabus?
- Is every learning outcome covered by a question?
- Is the relative importance of a topic and learning outcome reflected in the number and types of questions?
- Is the format appropriate for the students?
- Is the reading level of the questions appropriate for the group?
- Are the directions clear to the students?
- Are the questions clear and unambiguous?
- Is the scoring accurate?
- Is the grading of the results appropriate to the learning outcomes?

The list of the questions can prevent us from the inappropriateness of assessment given. However, it is not uncommon for major errors to occur in examinations.

Green (1975:138) mentions that there are two errors that can reduce the validity of measurement:

a. Error in the choice of an instrument.
   A writing test by asking the students to compose a text has high validity for measuring writing ability, probably has no validity for measuring speaking. When the wrong instrument is selected, validity will be low.
b. Error in item sampling.
   It occurs when the test has a large sampling error. For example, there are too few items or several unrepresentative test items.

B. Reliability

The second important quality of measuring instrument is its reliability. Reliability is the next important characteristic of assessment results after validity. Testers should be certain that the assessment has a high degree of validity and must also be reliable.

Reliability refers to its consistency. A reliable instrument is one which is consistent enough that subsequent measurements give approximately the same numerical status to the thing or person being measured. If a reliable test is given two or three times to the same group, each person in the group should get approximately the same score on all tests.

The Standard for Educational and Psychological Testing in Athanasou and Lamprianou (2002:175) define reliability as "the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual test taker." Reliability relates to questions about the stability and consistency of results for a group. The questions that need to be answers are:

- Would the ranking of students' results on an assessment be similar when it is repeated?
- Do two versions of an assessment produce the same results?
- Can I increase the reliability of results by lengthening a test?
- Are all the responses to the items or tasks homogeneous and consistent with each other?

In a sense reliability is a part of validity, for the test with high validity should measure that quality with

consistency and accuracy. It would be possible to have an extremely reliable test that had very little validity for the purposes for which it was being used. For instance, an algebra test might be reliable but lack validity if it were used in an English class to measure achievement in grammar.

Reliability or the stability of results are necessary but not sufficient for validity. Like validity, reliability refers to the results and not necessarily to a particular assessment. In other words, we cannot say that a particular test is reliable, but we can say that a set of results is reliable. Unlike validity which is based on evidence or inference, reliability is largely a statistical approach and is reported mainly as a correlation coefficient.

Unfortunately, in educational assessment, there will always be a margin of error. The major source of error or unreliability is people (test-takers). This is because people are naturally inconsistent in their response, answers and reaction to situations.

The accuracy or consistency of measurement is subject to three types of errors:

(1)Errors inherent in the instrument

For example: if one wished to measure the dimensions of a room and had a choice of using a yardstick or a 50-foot steel tape, he could error in the choice of instrument by selecting the yardstick.

(2)Errors in the use of the instrument

For example: related to the case above, even if he were to select the steel tape, he might still have an error of several inches in his measurement of the room if he lacked skill in using the tape.

(3)Errors emanating from the subject's response

For example: Errors emanating from subject responses have a variety of causes, including poor motivation,

lack of interest, improper test environment, poor emotional set, and illness.

Methods of Checking Test Reliability

There are three methods of checking test reliability: (1) the test-retest method, (2) the alternate-forms method, and (3) the split-half method. Each of these methods relies on the correlation technique, wherein two sets of scores are compared statistically to determine the extent of their relationship. Correlations are statistical indicators of a relationship.

The coefficient of correlation can range from a +1.00, a perfect positive relationship, to ─1.00, a perfect negative relationship. A coefficient of relationship of 0.00 shows no relationship between the two sets of scores (zero relationship). Reliability of good tests is generally in the range of +0.80 to above +0.95. As a rule, a negative correlation is unacceptable.

The following is the explanation for each method:

1. Test-retest method

   With test-retest method, only one form of the test is necessary, since the test is given to a group of pupils and after a reasonable lapse of time ─one to six months─ the same test is given to the same group. Then, testers should correlate the scores from the first occasion with those from the second occasion.

   Test-retest method gives *a coefficient of stability* because it is based on the stability of the performance of a group of students in a retest situation after a lapse of time following the first administration of the test.

   A test-retest correlation is normally calculated for a group. We would assume that if the reported test-retest correlations are high for a particular assessment

then we could have some confidence in the stability of results over time for most people who take this assessment. However, the test-retest method is not used for most classroom tests because it is an inconvenient approach.

2. Alternate-forms method

With the alternate-forms method two equivalent forms of a test are made up, and a group of pupils is given both forms in order that their scores on each form can be compared. Obviously, the higher the correlation, the more reliable the test will be. A comparison of the results for the groups determines equivalence and this should provide you with moderate (0.5 to 0.6) to large (0.7 to 0.8) reliability coefficients.

The alternate-forms method gives *a coefficient of equivalency* if both forms of the test are administered to the same group of pupils at the same time, since the purpose of the method is to develop two forms on which pupils obtain equivalent scores and equivalent rank within their group. This approach is much easier to implement for classroom tests.

3. Split-half method

The split-half method of determining reliability is used more than any other method because it requires only one group of pupils, one test form, and one test administration.

This technique actually gives *a coefficient of internal consistency*. It is based on giving the test once to a group; scoring two equivalent halves of the test (e.g., odd and even numbered questions), correlating the scores on the two halves and finally correcting for the shortened test length.

The split-half method can be used for classroom tests. Even though it is time consuming in adding up scores for each half but it is not difficult. These measures are simply a measure of mean inter-item correlation, or how well items correlate with each other.

Internal reliability coefficients are affected by some factors:

- Number of items on the test: The more the items, the more reliability the test has.
- Variation in item difficulty: Reliability just can increase with equal difficulty.
- Dispersion of scores: No spread of scores will reduce reliability.
- Level of item difficulty: Items with facility values of 0.5 maximize item variance, and so increase test reliability.

Teacher-made assessment commonly has reliabilities somewhere between 0.60 and 0.85 and for the most part this is satisfactory level for the instructional decisions made by teachers. Lower reliability is tolerable when: it is minor decision, dealing with the early stages of a decision, a decision is reversible, or achievement is able to be confirmed by other information, or the decisions that are made concern a class rather than an individual, or the effects of any decision are temporary.

Factors Affecting Reliability

Controllable factors that operate to increase or decrease the reliability of a measuring instrument include:

a. The test itself
   It includes:
   - Length of the instrument

Tests must be short enough to be practical and yet long enough to be reliable.
Example: short daily quizzes, such as a 10-item true-false quiz are completely unreliable as measuring devices.

b. The testee/test-taker
It includes:
- Variability of the tested group
Test reliability is greatest with an unselected heterogeneous group. If a test is administered to an extremely homogeneous group of pupils, the differences among individuals could be so small that a very slight increase or decrease in score would change one person's group rank significantly. Furthermore, if the test is too easy, the variability of scores is reduced, and reliability may be affected.
- Age range of the tested group
Although the age range is a factor in standardized tests, teacher-made tests that are administered to one class or grade level are little affected by it because ordinarily there is not a great variability in age range.

c. Tester/test maker (scoring process)
It includes:
- Objectivity of scoring
A test can be inherently reliable but have its reliability reduced either through error or subjectivity in the scoring. The more objective the test, the greater the scoring reliability.

d. Test administration
It includes:
- Testing conditions

Testing conditions affect reliability in that poor testing environment, including such factors as poor light, ventilation, or heat, can affect the reliability by increasing the human error.

C. Practicality

Practicality refers to its usability. It is the third desirable quality of tests. A test should be applicable to our particular situation. For a test to have a high degree of usability, it should:

- be easy to administer
- be easy to score
- be economical to use, both in terms of teacher time and of materials required
- have good format
- have meaningful norms.

The usability of a teacher-made test can be ensured by observing the following rules:

1. Have the test typed and duplicated so that each pupil will have a copy.
2. Directions to the pupil should accompany each part of the test.
3. The test should be designed to fit the time limits of the class period.
4. The test should be set up so that it can be readily scored.
5. Care should be exercised in planning the test to make it economical in terms of time required for test construction, duplication, and scoring.
6. Norms of pupil performance should be established from test results.

In the preparation of a new test or the adoption of an existing one, we must keep in our mind a number of very practical considerations.

1. Economy
   A test should be economy in cost and time. In writing or selecting a test, we should certainly pay some attention to how long the administering and scoring of it will take. This point is of particular importance when the test must be administered in the classroom and scored by the classroom teacher.
2. Ease of administration and scoring
   Other considerations of test usability involve the ease with which the test can be administered. Scoring procedures, too, can have a significant effect on the practicality of a given instrument.
3. Ease of interpretation
   If a standard test is being adopted, it is important that we examine and take into account the data which the publisher provides. If we plan to use the test over a long period of time, we shall almost certainly wish to develop local norms of our own.

UNIT III

CONSTRUCTING AND ADMINISTERING LANGUAGE TEST and TABLE OF SPECIFICATION

The construction of a good test does not happen by chance. It involves extensive thought, planning, and work. At least, there are three steps are essential in the development of a high-quality testing instrument: 1) planning the test, 2) constructing the test, and 3) evaluating the test.

Before constructing the test, there are some important steps that should be followed in planning the test. Teachers should consider the taxonomies of educational objectives in planning their instructional objectives, balance the emphasis that he places on various types of pupil learning, and also select the type of measurement that is appropriate for their selected objectives.

The taxonomies of educational objectives are well known as Taxonomy Bloom. Here is the outline:

1. Knowledge
   Knowledge involves recall of specific facts and universal concepts, structures, patterns, etc. The objectives are generally stated in terms of the pupil's

being expected "to know", "to demonstrate knowledge of", "to define", and "to be familiar with."

2. Comprehension
   This is the one that is generally of greatest concern in schools. It involves understanding of communications in oral and written form, either verbal or symbolic. The objectives are generally stated in terms of the pupils' ability "to understand", "to comprehend," "to interpret", "to translate", and "to interpolate".
3. Application
   Application requires problem solving using acquired knowledge and understandings. Objectives are frequently stated in terms of pupils' ability "to apply", "to predict", and "to use."
4. Analysis
   Analysis requires the ability to breakdown material into its constituent parts to determine the relationship and organizing principles among those parts. The objectives are often stated in terms of pupils' ability "to distinguish", "to analyze", "to discriminate", and "to infer".
5. Synthesis
   Synthesis requires putting together the elements and parts to form an entire entity. Objectives can be stated in terms of the ability "to write", "to put together", "to create", "to make", and "to produce".
6. Evaluation
   This category requires the pupil to make a judgment of the relative value of ideas, products, documents, paintings, etc. The objectives are written in terms of requiring the pupil "to judge", "to appraise", "to assess", and "to evaluate".

The six terms above are called cognitive domain. The second major learning domain is the affective domain. It deals with a person's values, feeling, and attitudes. This affective domain is much more difficult to teach and to evaluate than the more concrete cognitive domain.

A. Constructing language Test

The construction of an educational test includes the following steps:

1. Planning the test
2. Preparing the test items and directions
3. Submitting the test material to review and revising on the basis of review
4. Pretesting the material and analyzing the results
5. Assembling the final form of the test
6. Reproducing the test

Each step above will be described as follows:

Planning the Test

Effective testing requires careful planning. Here are the steps:

Step 1. Determining the General Course Objectives

In the preparation of an achievement test, one may base the test objectives directly on the objectives of the course. The teacher should consider the course content. The following are the example of course content:

1. Textbook lessons, each consisting of
   a. Short reading selection
   b. Dialogue
   c. Pronunciation drill
   d. Grammar drill

e. Word study

2. Laboratory practice, including drill on dialogue and pronunciation points keyed to the textbook.
3. Weekly compositions based on topics related to the textbook reading.

From the course coverage it is clear that the general objectives of the course are:

1. To increase skill in listening comprehension
2. To increase skill in oral production
3. To develop skill in reading simple descriptive and expository prose.
4. To develop skill in writing simple description and exposition

Our basic objectives, then, are to measure the extent to which students have acquired or improved their control of these skills.

Step 2. Dividing the General Course Objectives into Their Components

The objectives in step 1 were extremely broad. As our next step, then, we need to break them down into their specific components, after which we may determine which of these components we shall measure in our final examination.

Step 3. Establishing the General Design of the Test

At this point, two extremely important factors must be considered: the time to be provided for testing, and the degree of speediness we wish to build into our test.

Let us assume that a maximum of two hours has been scheduled for the final examination. Of the total 120 minutes, we should reserve at least 10 minutes for administrative procedures: seating the students and handing out the materials, giving general directions, collecting the materials at the end of the testing period, handling unanticipated

problems, etc. We are thus left with 110 minutes for actual testing.

Preparing the Test Items and Directions

a. Additional notes on the preparation of items
   In the preparation of multiple-choice or short-answer (supply) items, it is always necessary to write more items than will be needed in the final form of the test.
b. Writing test directions
   Test directions should be brief, simple to understand, and free from possible ambiguities. They should be accompanied by sufficient examples to ensure that even the slow learner or least skilled examinee understands the problem type. It is also advisable too to indicate the length of time which will be allowed for the test or its several parts.

Reviewing the Items

When the items have all been written, they should be set aside for a few days before being reviewed by the writer. Once he is satisfied with his material, it should be submitted to at least one colleague with experience in the subject matter field-as, in this case, a teacher of English.

Pretesting the Material

Standard objective tests consist of pretested materials. That is to say, all the items have first been tried out on a fairly large number of subjects of the same kind as those for whom the test is being designed. Only those items which prove statistically *satisfactory* in the pretest are included in the final version of the test. Items are said to be statistically satisfactory if they meet two requirements:

- If they are of a suitable level of difficulty—neither too hard nor too easy for the population being tested.
- If they discriminate between those examinees who know the material or have the skills or abilities being tested, and those who do not.

In the case of informal classroom tests, pretesting is seldom practicable, and in the preparation of such tests this step, and item analysis, may need to be omitted. Pretesting is, however, essential for any test that is to be administered to large numbers of examinees and used to make important decisions about these subjects—as, for instance, a test designed to screen university applicants or to place such students in appropriate classes.

Analyzing the Pretest Result (Item Analysis)

After the pretest answer sheets have been accumulated, the items should be analyzed to determine their effectiveness in terms of the two criteria listed in the preceding section.

a. Determining item difficulty
   A very satisfactory method is simply to ascertain the percent of the sample who answered each item correctly.
b. Determining item discrimination
   This step is to determine how well each item discriminates between high-and low-level examinees, for each item in a test should help to separate the proficient subjects from those who lack the tested skills or learning.

   How?

   Step 1. Separate the highest and the lowest 25 percent of the papers.

Step 2. For each item, subtract the number of "lows" who answered the item correctly for the number of "highs" who answered correctly.

Step 3. Divide the result of step 2 by the number of papers in each group, "highs" and "lows", to obtain the "item discrimination index."

c. Determining the Effectiveness of Distracters

One further step in the analysis of multiple-choice items is highly desirable, and that is to inspect the way each item distracter functioned.

d. Recording item analysis data

It is most convenient to record item analysis data on an "item analysis slip" which contains (1) the item, written out in full, (2) an identification of the pretest in which the item was tried out, (3) the position of the item in the pretest, (4) the item difficulty and discrimination indices, and, in the case of multiple-choice items, (5) a tabulation of how the "highs" and "lows" responded to the several choices.

Assembling the Final Form

In assembling multiple-choice items in the final form, the test maker must take care not only to order the items according to increasing level of difficulty but also to ensure that (1) each answer position is used about the same number of times and (2) the answer positions do not form any observable pattern. For example, if we have a test consisting of one hundred 4-choice items, each answer position A, B, C, and D should occur about twenty-five times, the answers having the appearance of a completely random sequence.

Reproducing the Test

1. It is essential that test materials be reproduced as clearly as possible, for poor reproduction of a test will almost certainly affect student performance.
2. Test material should be spaced so as to provide maximum readability.
3. No multiple-choice item should be begun on one page and continued on the next.
4. When blanks are left for the completion of short-answer items, a guideline should be provided on which the examinee may write his response.
5. It is advisable to indicate at the bottom of each page whether the examinee is to continue on to the next page or stop his work.
6. If each part of the test is separately timed, the directions for each part should occupy a right-hand page of the book.
7. The use of a separate cover sheet will prevent examinees from looking at the test material before the actual administration begins.

B. Administering Language Test

Preparing for the Test

1. Selecting the testing room.
   Testing must be conducted in a quiet, well-lighted room located where there will be a minimum of outside noise and interference.
2. Checking materials and equipment.
   At least a day in advance of the test-and considerably earlier if materials must be ordered from an outside source-the examiner must take a careful count of all testing materials to ensure that an adequate supply is

available. The examiner must also obtain an accurate watch for timing the test.

3. Reading the test materials in advance.
   The examiner must familiarize himself with all the test materials in advance of the administration.

Conducting the Testing Preliminaries

1. Seating the examinees.
   As the examinees arrive at the testing room, the proctors should have them take alternate seats within rows.
2. Calling the examination to order
   When the examinees have all been seated, the examiner should call the group to order and briefly introduce the test.
3. Distributing the test materials.
   Materials should be given individually to each examinee.
4. Instructing examinees on filling out the answer sheet.
   When all materials have been distributed, the examiner should explain how the personal data portion of the answer sheet is to be filled out.
5. Giving test directions
   Clear, complete directions to examinees should be included in the test itself.

C. Table of Specification

According to Fulcher and Davidson (2007:52), Test specifications (specs) are generative explanatory documents for the creation of test tasks. Specs tell the test makers of how to phrase the test items, how to structure the test layout, how to locate the passages, and how to make a host of difficult choices as we prepare test materials.

Test specifications for classroom use can be a simple and practical outline of a teacher-made test. The table of specifications may consist of:

1). A broad outline of the test,
2). Skills to be tested, and
3). The items will be look like.

The time span of the test should follow the curriculum, and the four skills need to be tested because suggested in the curriculum also. While, considering the items is much more complicated and should be based on the skills and its purposes.

The following is the sample of table of specification:

KISI-KISI PENULISAN SOAL MID SEMESTER I

Sekolah : SMPN …..
Mata Pelajaran : Bahasa Inggris
Kelas : VII
Jumlah Soal : 20 butir
Alokasi Waktu : 30 menit
Tahun Pelajaran : 2013/2014
Penulis : ……..

No	SK	KD	Mat	Indikator Penc.	Indikator soal	Bentuk soal	Level of think	No soal	Wkt

Ket:

No	: Nomor
SK	: Standar Kompetensi
KD	: Kompetensi Dasar
Mat	: Materi
Indikator Penc.	: Indikator Pencapaian
Indikator Soal	: Indikator Soal
Bentuk soal	: Bentuk soal
Level ot think	: Level of thinking
No. Soal	: Nomor soal
Wkt	: Waktu

STANDAR KOMPETENSI

5. Memahami makna dalam teks tulis fungsional pendek sangat sederhana yang berkaitan dengan lingkungan terdekat.

KOMPETENSI DASAR

5.2 Merespon makna dalam teks tulis fungsional pendek sederhana secara akurat, lancar dan berterima yang berkaitan dengan lingkungan sekitar.

MATERI

Functional Text

- Invitation card
- Short messages
- Announcement
- Greeting Card

INDIKATOR PENCAPAIAN

- Mengidentifikasi makna teks tulis fungsional berbentuk invitation card.
- Mengidentifikasi makna teks tulis fungsional berbentuk short message.
- Mengidentifikasi makna teks tulis fungsional berbentuk announcement.
- Mengidentifikasi makna teks tulis fungsional berbentuk greeting card.

INDIKATOR SOAL

1. Invitation Card
   - Menentukan gambaran umum teks
   - Menentukan topik teks.
   - Menentukan informasi rinci tersurat dalam teks.
   - Menentukan informasi rinci tersirat dalam teks.
   - Menentukan rujukan kata dalam teks
   - Menentukan makna kata berdasar konteks

2. Short Messages
   - Menentukan gambaran umum teks
   - Menentukan topik teks.
   - Menentukan informasi rinci tersurat dalam teks.
   - Menentukan informasi rinci tersirat dalam teks.
   - Menentukan rujukan kata dalam teks
   - Menentukan makna kata berdasar konteks

3.Announcement
   - Menentukan gambaran umum teks
   - Menentukan topik teks.
   - Menentukan informasi rinci tersurat dalam teks.
   - Menentukan informasi rinci tersirat dalam teks.
   - Menentukan rujukan kata dalam teks
   - Menentukan makna kata berdasar konteks

4. Greeting Cards
   - Menentukan gambaran umum teks
   - Menentukan topik teks.
   - Menentukan informasi rinci tersurat dalam teks.
   - Menentukan informasi rinci tersirat dalam teks.
   - Menentukan rujukan kata dalam teks
   - Menentukan makna kata berdasar konteks

BENTUK SOAL

- Pilihan ganda
- Essay

UNIT IV

TYPES OF TESTS

There are some kinds of tests that are divided into some categories. According to a small number of kinds of information being sought, the tests are divided into as follows (Hughes: 9):

1. Proficiency Tests

   This kind of test is designed to measure people's ability in a language regardless of any training they may have had in that language. The content of this test is based on a specification of what candidates have to be able to do in the language in order to be considered proficient. So, it is not based on the content or objectives of language courses which people taking the test may have followed.

   In the case of some proficiency tests, "proficient" means having sufficient command of the language for a particular purpose. Example: a test designed to discover whether someone can function successfully as a United Nations translator, a test used to determine whether a student's English is good enough to follow a course of

study at a British university, Cambridge examinations, and the Oxford EFL examinations.

Despite differences between them of content and level of difficulty, all proficiency tests have in common the fact that they are not based on courses that candidates may have previously taken.

2. Achievement Tests

Most teachers are unlikely to be responsible for proficiency tests. It is much more probable that they will be involved in the preparation and use of achievement tests. In contrast to proficiency tests, achievement tests are directly related to language courses, their purpose being to establish how successful individual students, groups of students, or the courses themselves have been in achieving objectives.

There are two kinds of achievement tests: final achievement tests and progress achievement tests. *Final achievement* tests are those administered at the end of a course of study. They may be written and administered by ministries of education, official examining boards, or by members of teaching institutions. Clearly the content of these tests must be related to the courses with which they are concerned, but the nature of this relationship is a matter of disagreement among language testers.

In the view of some testers, the content of a final achievement test should be based directly on a detailed course syllabus or on the books and other material used (syllabus-content approach).

The disadvantage; if the syllabus is badly designed, or the book or other materials are badly chosen, and then the result of a test can be very misleading.

Progress achievement tests are intended to measure the progress that students are making. Since "progress" is towards the achievement of course objectives, these tests too should relate to objectives. Then if the syllabus and teaching are appropriate to these objectives, progress tests based on short-term objectives will fit well with what has been taught. If not, there will be pressure to create a better fit.

3. Diagnostic Tests

Diagnostic tests are used to identify students' strengths and weaknesses. They are intended primarily to ascertain what further teaching is necessary. Testers can create tests that will tell them that a student is particularly weak in, say, speaking as opposed to reading in a language.

The lack of good diagnostic tests is unfortunate. They could be extremely useful for individualized instruction or self-instruction.

4. Placement Tests

Placement tests are intended to provide information which will help to place students at the stage (or in the part) of the teaching program most appropriate to their abilities. Typically, they are used to assign students to classes at different levels.

The placement tests which are most successful are those constructed for particular situations. They depend on the identification of the key features at different levels of teaching in the institution.

According to their functions, tests fall into three categories:

1. Instructional tests.

This test is used in formative evaluation which is designed to let the pupils know about their strengths and weaknesses. When it is graded by teachers, the test is handed back to the students for class discussion.

2. Mastery tests

   The function of mastery test is to assess pupils competencies before proceeding to the next level of learning. This kind of test is useful in such basic skill areas such as grammar and mathematics. Evaluation of pupil competence on this performance basis is *criterion-reference* evaluation and relies on the individual performance.

3. Measurement tests

   This is a kind of standardized tests which established national or regional norms, or carefully constructed teacher-made tests. This test is given when the teacher needs to obtain *norm-referenced* or *group-performance.*

   Measurement test has a high difficulty level with the purpose to give a precise measurement of each pupil's achievement during the school year, during the semester, or at the end of a unit of teaching.

There are four types of tests based on its forms, they are: oral, objective, essay, and performance test. The explanation for each is given below:

A. Objective Test

Objective tests can be generally classified as formal or informal tests. *The formal tests* are the published, standardized tests that have been highly refined (1) through the process of careful analysis of objectives, of course

content, and of texts to determine test content; (2) through administration to a representative sample of the target population; (3) through item analysis and item editing and revision to eliminate faulty items; (4) through the establishment of standardized scores which represent the levels of performance of the sample of the population or the criterion group to whom the test was administered.

The informal objective tests are tests constructed by the teacher for use in his classroom. These tests are most useful in measuring pupils' knowledge and understanding in various subject fields and levels of instruction.

The advantages of objective tests are: 1) permits evaluative sampling of an extensive domain of content, 2) can be refined through item analysis for reuse with the same or other groups. In the cognitive domain, objective tests function best in the knowledge, comprehension, and application levels.

Forms of Objective tests

Green (1975: 58) divides the forms of objective tests into four. The explanation for each is given below.

1. Short-answer form

This may include questions, incomplete sentences, definitions, or identification items. This form is particularly well adapted to the measurement of factual recall. When the objectives to be measured indicate that the pupil is to know, to recall, to recognize, or to identify, then certainly the short-answer is an appropriate test form to use.

The short-answer question has wide application in education. It also requires a constructed response. Short answer questions are useful for eliciting necessary facts, information, declarative or procedural knowledge.

An assessment that uses short answer questions restricts the answer to a paragraph, a sentence or even a phrase. The short answer questions have clearly identified, correct answers and are scored objectively.

The examples of short-answer form:

- ✓ Question
  What is the type of the essay?
- ✓ Incomplete sentence
  The name of the first president of Indonesia was _____.
- ✓ Definition
  Define practicality: _____
- ✓ Identification
  Identify the core part of the following sentence.
  The man who is standing in front of the office is his father.

  _____

Rules for construction:

a. Only significant words should be omitted in incomplete statement items.
b. Grammatical clues to the correct answer should be avoided.
c. Blanks should be kept uniform in length so that the pupil is not given the unnecessary clues of a short blank for a short response and a long blank for a long response.
d. One point should be allowed for each blank.
e. Blanks should be arranged in a manner convenient to score.
f. Limit use of this test form to measure only the recall important information.

g. When the question form is used, questions should be explicit enough to evoke the correct type of response.
h. Limit the length of the responses to single words or short phrases.
i. Verbatim quotes from the textbook should be avoided.
j. A scoring key, which contains all acceptable answers, should be prepared.

***Short-answer forms have several strengths, as follow*:**

1) It is relatively easy to prepare and construct.
2) It is most effective in measuring recall.
3) It can be used in many fields to provide an extensive sample of factual information.
4) Providing better coverage of topics than an essay question.
5) Being easy to score.
6) Having high reliability of scoring, especially with model answer.
7) Allowing more scope than other objective questions to show whether a student has greater knowledge of a subject.

***On the other side, this test form also has some weaknesses as follow*:**

1) It is difficult to construct items that call for only one correct answer.
2) Completion-type items stress rote recall and encourage pupils to spend their time memorizing trivial details rather than seeking important understandings.

3) Completion items are somewhat unrealistic, since life problems generally offer a variety of possible solutions rather than one "key-word" solution.

2. Alternate-Response Form

The alternate-response or true-false form is probably used more frequently by teachers than any other test form. The alternate-response form can be constructed to measure fairly complex understanding. It consists of a statement to be judged true or false/correct or incorrect.

The average test difficulty is normally about the 75-percent level and pupil has a 50-50 chance of guessing the correct response. It means that, with a well-constructed, but difficult, true-false test, the average pupil will answer correctly approximately 75 percent of the items.

True-false questions are used widely in education and training because these questions can be directed to the essential structure of a subject's knowledge. They are helpful when there is a need to assess knowledge of the basic facts or ideas in a subject area.

Example:

- ✓ Most Americans stop working at age 65 or 70 and retire.

  T F

- ✓ Is downtown the central business area?

  Yes No

- ✓ Make every false statement true by suggesting a substitute for the underlined word.

  There is some <u>similar</u> between you and your brother.

***Rules for Construction**:*

a. Each statement should be entirely true or entirely false.
b. Trivial details should not make a statement false.
c. The statement should be concise without more elaboration than is necessary to give clear meaning.
d. Exact statements should not be quoted from the textbook.
e. Quantitative terms should be used instead of qualitative terms, whenever possible.
f. Specific determiners, which give a cue to the answer, should be avoided.
g. Negative statements should be avoided.
h. When a controversial statement is used, authority should be quoted.
i. A pattern of answers should be avoided.

Strengths of true-false questions:

1) Its versatility or adaptability to several fields.
2) Its usefulness in measuring both knowledge and understanding objectives.
3) Its relative ease of construction.
4) Its usefulness in sampling a broad range of material.

***The weaknesses are*:**

1) It encourages pupils to guess.
2) It is often poorly constructed.

3. Multiple-Choice Form

The multiple-choice test is considered by most test experts to be the best type of objective test for measuring a variety of educational objectives. The test is versatile, and it

requires some discriminatory thinking on the part of the pupil. Multiple-choice items have a premise, which consists of an incomplete statement or question followed by several choices which include one correct answer and several distracters.

Multiple choices is widely applicable forms of questioning that is developed around World War I as an economic means of administering tests to large groups. The aim of multiple choice forms is to assess a person's knowledge as well as his/her ability to discriminate among several possible alternatives. It is adaptable to most subject areas.

Many standardized and commercial tests such as TOEFL use only multiple-choice questions. It is important to have some knowledge of their use and construction because there are many misconceptions about this form of questioning.

A multiple-choice questions consist of a direct question or an incomplete statement, in which the main part of the question is called *the stem,* and the suggested solutions are called *alternatives/choices or options.* Typically, the student is requested to read the stem and consider the alternatives in order to select the best option. The incorrect alternatives are called *distracters*.

Example:

- ✓ What does the term *consistent* mean?
  a. steady
  b. unsteady
  c. compatible
  d. fluid
  e. changeable
- ✓ The word "auctions" in line 16 is closest in meaning to
  a. meetings

b. deliveries
c. sales
d. demonstrations

***Rules for Construction*:**

a. The central problem of the item should be stated in the premise so as to make only one choice justifiable.
b. All choices in the item should be grammatically consistent.
c. The choices should be as brief as possible.
d. A pattern of answers should be avoided.
e. Negatively stated items should be avoided.
f. Authority should be quoted when the item contains controversial opinion.
g. Ambiguous items should be avoided.
h. All choices should be plausible.
i. Specific determiners should be avoided.
j. Each item should contain an independent problem which gives no clues to the answers of other items.

Strengths of multiple-choice forms are:

1) It minimizes the guessing factors.
2) It can be used to measure higher types of learning than the other forms.
3) It is adaptability to numerous fields and the possibility of extensive measurement sampling.
4) It is more difficult to remember exam content than with other questions.
5) Items can be re-used with less concern for security of the exam.
6) Multiple choices are not as physically exhausting as a written exam.
7) Item responses are easily analyzed.

8) The potential guessing in high scores is minimal.
9) Around one multiple-choice item can be answered per minute.
10) The results are more reliable than with a comparable essay.
11) The sampling of subject content is greater than with essays.
12) Higher level thinking can also be measured by the multiple-choice item.

***Weaknesses*:**

1) It is difficult to construct good and effective questions.
2) The formats are restricted to abstracted or verbally presented content rather than real situations in context.
3) They rely largely upon the recognition of the right answer and this is a very specific cognitive process.

4. Matching Form

The matching examination is most useful for measuring recognition and recall. It can be constructed in such a manner that the pupil has virtually no chance of guessing the correct responses.

Example:

All the names on the premise (on the left) are the capital city of the countries on the right.

1. Kualalumpur	a. Indonesia
2. Singapore	b. Malaysia
3. Jakarta	c. Myanmar

4. Bangkok	d. Philippines
5. Manila	e. Singapore
	f. Thailand

***Rules for Construction*:**

a. Only homogeneous premises and homogeneous responses should be grouped in a single matching set.
b. Relatively short lists of responses—not less than 5, not more than 15—should be used.
c. Premises should be arranged for maximum clarity and convenience to the examinee.
d. Response options should be arranged alphabetically or chronologically.
e. Directions should clearly indicate the intended basis for matching.
f. Providing perfect one-to-one matching between premises and responses should not be attempted.
g. All of the choices for each matching set should be included on a single page.
h. More responses than premises should be used in a set, or a single response should be used to answer several premises.

***Strengths*:**

1) It is easy to construct.
2) It takes very little time to administer and to correct.
3) It is to measure the pupil's ability to recognize or recall these facts

Weaknesses:

1) Its inability to measure the higher levels of learning.

B. Performance Test

Performance of any task is a complex of many factors, but it includes two measurable aspects: (1) the procedure, skill, or technique, and (2) the product or result. When the procedure is assessed, the examiner is attempting to determine how skillfully the subjects perform the desired procedure, while the assessment of the product stresses the end result through an examination of the quality of the product.

A teacher must select a measurement approach adaptable to his purpose. The three general approaches are:

1. Object or identification tests
   It is a test that emphasizes knowledge of the product, wherein the pupil is asked to identify the nature and function of various components of the product.

2. Procedure evaluation methods
   It includes:
   a. The evaluation of actual job performance
      This would be an ideal situation for evaluation, but it's too time-consuming unfeasible.
   b. The use of simulated-conditions tests to measure performance under conditions that approximate actual job conditions.
      This approach emphasizes the procedure. The actual conditions of the job to be measured are duplicated as nearly as possible.
   c. The use of work-sample tests to measure specific task or skill performance that is essential to competent job performance.
      This approach emphasizes on both the product and the procedure. Pupils are required to perform some aspects of the total job that is being measured.

The quality of the pupil's performance of the tasks then gives evidence of his likely success or failure in the job in question.

The procedure can be evaluated by using two forms, they are:

1). A check list

It is used to determine whether or not each of the important elements is represented in the subject's procedure without any attempt to evaluate the level at which each was performed.

2). Rating scale

It evaluates the level of performance on each element of the procedure. The evaluator must both identify the specific elements performed by the subject and determine how well he/she completed those performances.

3. Product evaluation methods

It includes:

a. Quality scales

It is often used to evaluate the products that pupil complete or construct. Example: drawings, writing composition, etc.

The steps are:

(1). Select 5 to 10 samples of pupils' work which representatives of various quality levels.

(2). Arrange the samples in a sequence from best to worst with approximately equal increments of improvement between each.

(3). Establish numerical scores or grade equivalents for samples at various levels of the scale.

b. Consensus ratings

It is a group-rating method in which a number of people rate the product or the procedure. For example, scores from a teacher and peers. The average of these ratings is the score that is assigned.

c. Rank orderings

It is another means of comparatively evaluating products. The products are arranged in order from the best to the poorest.

d. Paired comparisons

In this method each product is compared against others in all possible pairings to determine which of each pair is better.

C. Essay Test

Essay tests should be used to measure such objectives as understandings, attitudes, interests, creativity, and verbal expression. With reference to the taxonomy, this test form is useful to evaluate learning in the upper levels of the cognitive domain, notably: application, analysis, synthesis, and evaluation.

The essay examination fits well with the teaching style that encourages divergent thinking and creativity, that stresses the acquisition and application of large concepts, and that deals with controversies and problem solving.

Types of Essay Tests

There are two types of essay tests:

1. Restricted-response items

It is much more easily graded than the extended-response type, for the pupil's answers are closely circumscribed.

Example:

a. Discuss three techniques of instruction—lecture, demonstration, and class discussion.
b. Define the term *inference*.

2. Extended-response items

The extended-response question is one that is relatively unstructured, permitting the pupil freedom in organizing and expressing the answer in a manner that displays his personal insights and the breadth and scope of his knowledge.

Example:

a. Discuss the tax structure of the local, state, and federal branches of government in the United States.
b. Explain the purpose of the author in writing the article.

Monroe and Carter in Green (1975:111) list the variety of possible types and essay questions as follow:

1). Selective recall─basis given

- Name the presidents of Indonesia who had been in military life before their election.
- What does urban mean?

2). Evaluation recall─basis given

- Which do you consider the three most important reasons for people to take the public transportation?

- Name the two most practical reading strategy by Kinsella.

3). Comparison of two things—on a single designated basis

- Compare Eliot and Thackeray in ability in character delineation.
- Compare the armies of the North and South in the Civil War as to leadership.

4). Comparison of two things—in general

- Compare the early settlers of the Massachusetts colony with those of the Virginia colony.
- Contrast the life of Habibi in Indonesia with his life in Germany.

5). Decision—for or against

- Whom do you admire more, Abu Bakar Siddik or Ali bin Abi Thalib?

Planning Essay Tests

The major problems to be considered in the planning stage are:

1. Are the material and objectives to be measured adaptable to an essay test?
2. Do the pupils have sufficient background, both in composition and in the subject area, to write an essay test?
3. Does the test permit freedom of response sufficient to permit the pupils to bring to bear the depth and scope of their knowledge.
4. Has sufficient time been allotted the pupils to plan their responses so that they are more than mere

disjointed associations of somewhat relevant statements?

Construction of Essay Tests

The following suggestions will help the teacher to construct better essay tests and get better results from them:

a. The questions should be written with care, the language should be precise in meaning and unambiguous.
b. Time and thought should be given to the actual selection and preparation of the questions.
c. The directions for the test should be explicitly written out.
d. There should be no optional questions on the test.
e. Open-book essay tests should be avoided.
f. Pupils should be given advance notice of essay tests.
g. Pupils should be given training in taking tests.
h. Validity of the essay test is improved by restricting its use to the measurement of content and objectives to which this test form is best adapted.
i. The reliability and the extent of the sample for essay tests can be improved by increasing the number of questions and varying the types of questions.

Methods of Grading

There are two acceptable methods for grading essay tests:

1. The point-score method
    a. The Construct a grading key which includes the major aspects that the pupil should include in his response to each question.
    b. Read a single question through all the papers.

c. Total the points on each paper after all of the questions have been read.

2. The sorting method
   a. Read through all the papers as quickly as possible
   b. Reread the papers
   c. Assign letter grades on the basis of the piles into which the papers were sorted.

Problems of Grading

There are several special problems in grading essay examinations, regardless of the method used:

1. The halo effect

   The halo effect is apparent when the score given by the grader for one paper is influenced by a well organized and answered of the first question. Even though the following answers are poorer, the paper is still assigned a high score.

   In another case, a good pupil who did well at the first test, but did poor at the last test is still assigned a good score because the scorer is influenced by his previous performance. On the other hand, a poor pupil is still given a low score even though he did the last test better than the previous one.

2. High grading or generosity error

   It is related to halo effect. The grader who is affected by generosity error consistently assigns grades that are too high for all papers.

3. Low grading or penalty error

   It is unfair as generosity error. The grader is too strict. He consistently assigns grades that are low or too low with too high standard for all of the classes.

4. The influence of extraneous factors.

This factor is extremely difficult to overcome. For example, the bad quality of handwriting can cause the difficulty in reading the papers that can affect the students' scores. Other problems are ability in grammar, vocabulary, spelling, organization, neatness and composition.

D. Oral Test

The oral test is the oldest form of examination used by teachers. It was used by early teachers. Nowadays, it is used rarely to examine students' ability. When properly constructed and used, it can be both a good instructional technique and a valuable, informal means of appraising pupil progress.

Oral Questioning Strategies

According to Suchman in Green (1975), there are four types of questions that can lead pupils to clarify their thinking and improve their learning. The four questions are:

1. Verification questions.
   The pupils seek the information base, the facts, and data necessary to his quest.
2. Experimentation questions.
   The pupils verbally manipulate information gathered at the verification level.
3. Necessity questions.
   The pupils seek to sort the data and to determine what data are relevant and irrelevant to his quest.
4. Synthesis questions.
   The pupils check the validity of his hunches, theories or conclusion as explanation or solution to the problem or phenomena under scrutiny.

Types of Oral Examination

Usually, when doing oral examination, a teacher asks each pupil in turn a single question and then gives grade. Obviously such practice has little measurement value. The following classification of oral examination is suggested:

1. The orally administered examination that require an oral response. This examination can be grouped in two types:
   a. A single question is asked to individual in group situation.
   b. Numerous questions are posed to single individual.
2. The orally administered examination that require a written response.
   For example: teacher-constructed test, auditory comprehension examination, or the readiness test given to preschool.
3. The orally administered examination of the standardized type.
   For example: intelligence test.
4. The interviews in which persons are selected for particular responsibilities or positions.
   For example: an interview given to the job applicants.

Planning Oral Examination

Many teachers cannot use the oral examination properly because of poor planning. In preparing oral examination, care should be taken with the following planning steps:

1. The objectives and content areas should be listed to form a test outline. Next, the outline needs to be elaborated as that for objectives.
2. The types of oral examination should be selected on the basis of the table of specification.

Construction of Oral Examination

There are two most important principles in constructing the oral examination:

1. The questions should be written out ahead of time.
2. Acceptable answers should be written for each of the questions constructed.

The questions for oral examination should be prepared more than a simple recall response. The following is types of questions for oral examination suggested by Gallagher (1963):

a. Convergent (conventional thinking)
   Example: How does perception of students toward their teachers affect their school achievement?
b. Divergent (creative thinking)
   Example: If Institution doesn't support its students to follow exchange study, how can it promote global change in the institution?
c. Evaluative (judgment thinking)
   Example: How does fluoridation of water compare to systematic brushing in preventing tooth decay?

The types of the oral questions above are divided into the following:

1. Oral questions-oral responses
   This kind of questions should be similar to questions written for essay examination (can be restricted or extended-response questions).
2. Oral questions-Written response
   Fewer questions should be included in this examination because pupils obviously write more slowly than they speak.
3. Oral Performance examination

It is particularly well adapted to such areas as speech, dramatics, and foreign language. Both speech and dramatics stress verbal performance, and the quality of the performance cannot be measured by written examination.

UNIT V

TESTING LANGUAGE SKILLS AND COMPONENTS

A. Testing Grammar

There is an essential difference between the traditional "grammar" test for the native speaker of English and the kind of structure test appropriate for the foreign learner. Structure test for the native speaker is formal written English. On the other hand, structure tests for foreign students will have as their purpose the testing of control of the basic grammatical patterns of the spoken language.

The preparation of s structure test should always begin with the setting up include the full range of structures that were taught in the course, and each structural type should receive about the same emphasis in the test that it received in the classroom.

The following are the item types of grammar test:

1. *Completion (multiple-choice)*

   Example:

   a. Sinta (lives) (is living) (has lived) in Padang since 2010.

      A B C

b. Sinta _____ in Padang since 2010.
   A. lives C. has lived
   B. is living
c. "is Sinta still in Jakarta?" "No, _____ in Padang since 2010."
   A. she lives C. she's living
   B. she'd lived D. she's lived

2. *Sentence alternatives (multiple-choice).*

Example:

Choose the best answer.

A. Sinta is living in Padang since 2010.
B. Sinta lives in Padang since 2010.
C. Sinta has lived in Padang since 2010.

3. *Sentence interpretation (multiple-choice)*

Example:

"An old friend of Ryan's family brought him news of his uncle last night." Him refers to
A. an old friend C. the uncle
B. Ryan

4. *Scrambled sentence (multiple-choice)*

Example:

When _____ ?
A. plan C. to go
B. do D. you

5. *Completion (supply type)*

Example:

Direction: complete the sentences by writing a form of the verb given in parentheses.
Sinta _____ (live) in Padang since 2010.

6. *Conversion (supply type)*

   Example:
   Direction: Change the sentences into passive sentences
   1. She reads the novel twice a week.
   2. My lecturer gives an assignment every week.

B. Testing Vocabulary

The selection of vocabulary test words is relatively easy in achievement tests. The first decision that must be made is whether to test the students' *active or passive* vocabulary, that is the words they should be using in their speech and writing or those they will need merely to comprehend, especially in their reading.

Generally speaking, vocabulary tests on an *intermediate level* will concentrate on the *words needed in speaking* or in comprehending the oral language, while tests on an *advanced level* will deal mostly with the lexicon of *written English*—the words needed by students if they are to understand newspapers, periodicals, literature, and textbooks.

In selecting the test words, dictionary may be used, but it is more convenient to use word lists based on frequency counts of lexical items occurring in actual samples of the language.

Useful as these and similar word counts are, the test maker must be alert to their several shortcomings:

1. Word counts are usually based on the written language only; therefore, many words that are extremely common in the oral language will receive low frequency ratings in the word lists.
2. The word lists classify words according to relative frequency rather than absolute difficulty, and the two are by no means always equivalent.

3. Word frequency in English does not serve as a good guide to the probable difficulty of lexical items for which there are cognate forms in the foreign students' native language.
4. Some of the word lists do not differentiate among the various meanings of a word.
5. Unless the word lists are based on very recent surveys of frequency, they are likely to contain items whose status is currently quite different from what it was at the time the data were collected.
6. Some word lists are based on a sample of written materials quite unlike those which the typical foreign learner of English is likely to have read.

Item types of testing vocabulary

a. Multiple-choice

1. Definition
   It is called the "classic" type of vocabulary item which consists of a test word followed by several possible definitions or synonyms.
   Example:
   Nap
   A. A brief sleep
   B. A happy song
   C. A sharp rock
   D. A short meeting
2. Completion
   A second item type places the problem words in context.
   Example:
   The old woman was too _____ to push open the heavy door.
   A. Feeble

B. Sincere
C. Deaf
D. Harsh

3. Paraphrase
   A third method of testing vocabulary, combining elements of two of the previously discussed devices, is to underline a word in context and provide several possible meanings.
   Example:
   George was astounded to hear her answer
   A. Greatly amused
   B. Greatly relieved
   C. Greatly surprised
   D. Greatly angered

b. Supply type

1. Paraphrase
   This is the variation of type above (paraphrase/multiple-choice) and useful highly in informal classroom testing. This type requires a structured short answer supplied by the examinee.
   Example:
   George was astounded to hear her answer.
   Direction: rewrite the sentence by substituting other words for the underlined portion.
   The possible answers include:
   - ✓ George was greatly surprised to hear her answer.
   - ✓ George was amazed to hear her answer.
   - ✓ George was astonished to hear her answer.

2. Pictures (objective)

   In the testing of children who have not yet reached the reading stage, vocabulary may be measured with pictures. There are two types of picture items have frequently been used, as follow:

   a. The examiner pronounces the name of an object and asks the child to indicate, by pointing or making a pencil mark, which one of a set of pictures shows the object named.

      Example:

      The test booklet might contain four pictures—of a book, a bird, a boat, and a box—and the examiner might ask, "Draw a circle around the boat."

   b. In the second type, the child is shown a picture of an object and is asked to name it.

The Principles of Item Writing for Testing Vocabulary

1. The definition should be expressed in simple words readily comprehensible to all examinees.
2. All the alternatives should be on approximately the same level of difficulty.
3. Whenever possible, all choices should be related to the same general area or kind of activity.
4. The choices in each item should be of approximately the same length or be paired by length.
5. Items should be kept free of extraneous spelling problems.

C. Testing Listening

Teachers need to pay close attention to listening as a mode of performance for assessment in the classroom.

Designing appropriate assessment tasks in listening begins with the specification of objectives, or criteria.

The performance of listening with specific objectives is categorized in two skills; microskills (bottom-up process of learning such as the smaller bits and chunk of language) and macroskills (top-down approach of listening task which focused on the larger elements of language).

The following lists show the micro and macroskills which provide 17 different objectives to assess in listening (adapted from Richards, 1983 in Brown (2004)).

Microskills:

1. Discriminate among the distinctive sounds of English.
2. Retain chunks of language of different lengths in short-term memory.
3. Recognize English stress patterns, words in stressed and unstressed positions, rhythmic structure, intonation contours, and their role in signaling information.
4. Recognize reduced forms of words.
5. Distinguish word boundaries, recognize a core of words, and interpret word order patterns and their significance.
6. Process speech at different rates of delivery.
7. Process speech containing pauses, errors, correction, and other performance variables.
8. Recognize grammatical word classes (nouns, verbs, etc.), systems (e.g., tense, agreement, pluralization), patterns, rules, and elliptical forms.
9. Detect sentence constituents and distinguish between major and minor constituents.
10. Recognize that a particular meaning may be expressed in different grammatical forms.
11. Recognize cohesive devices in spoken discourse.

Macroskills:

12. Recognize the communicative functions of utterances, according to situations, participants, goals.
13. Infer situations, participants, goals using real-world knowledge.
14. From events, ideas, and so on, described, predict outcomes, infer links and connections between events, deduce causes and effects, and detect such relations as main idea, supporting idea, new information, given information, generalization, and exemplification.
15. Distinguish between literal and implied meanings.
16. Use facial, kinesic, body language, and other nonverbal clues to decipher meanings.
17. Develop and use a battery o listening strategies, such as detecting key words, guessing the meaning of words from context, appealing for help, and signaling comprehension or lack thereof.

Brown (2004) in his book divides listening assessment tasks into four:

1. Intensive Listening

Listening for perception of the components (phonemes, words, intonation, discourse markers, etc) of a larger stretch of language. The focus in this section is on the micro skills of intensive listening.

There are several kinds of tasks assessed for intensive listening, they are:

a. Recognizing Phonological and Morphological Elements

The typical test gives a stimulus and ask test-takers to identify the stimulus from two or more choices, such as:

- *Phonemic pair, consonants*
  Test-takers hear : He's from Surabaya
  Test-takers read : (a). He's from Surabaya
  (b). She's from Surabaya

- *Phonemic pair, vowels*
  Test-takers hear : Is she living?
  Test-takers read : (a). Is she leaving?
  (b). Is she living?

- *Morphological pair, -ed ending*
  Test-takers hear : I loved chocolate
  Test-takers read : (a). I loved chocolate
  (b). I love chocolate
- *Stress pattern in can't*
  Test-takers hear : My brother can't come tonight
  Test-takers read: (a). My brother can't come tonight
  (b). My brother can come tonight
- *One-word stimulus*
  Test-takers hear : vine
  Test-takers read : (a) vine
  (b) wine

b. Paraphrase Recognition

In this kind of task, the words, phrases, or sentences are frequently assessed by providing a stimulus sentence asking the test-takers to choose

the correct paraphrase from a number of choices. The following are the example:

- *Sentence paraphrase*

  Test-takers hear : Hi, my name's Siti. I come from Malaysia

  Test-takers read :

  (a) Siti likes Malaysia
  (b) Siti wants to come to Malaysia
  (c) Siti is Malaysian
  (d) Keiko is happy in Malaysia

- *Dialogue paraphrase*

  Test-takers hear: Man: Hi, Mary, my name's John
  Woman: Glad to meet you, John. Are you American?
  Man: No, I'm Irish.

  Test-takers read: (a). John lives in the USA
  (b). John is American
  (c). John comes from Ireland
  (d). Mary is Irish

2. Responsive Listening

Responsive listening focuses on the assessment of a relatively short stretch of language (a greeting, question, command, comprehension check, etc) in order to make an equally short response. This typical listening task is in a question-and-answer format which provide interaction. The following are the example:

- *Appropriate response to a question*

Test-takers hear: How much time did you have to finish your project?

Test-takers read: (a) in about a month
(b) about a month
(c) about Rp.1 million
(d) yes, I did

- *Open-ended response to a question*

Test-takers hear: How much time did you need to finish your project?

Test-takers write or speak: _____

This kind of task needs students' creativity to response what they hear, and the teacher must judge the response appropriateness which takes time.

3. Selective Listening

In this kind of assessment task, the test-takers listen to a limited quantity of aural input and must discern within it some specific information.

Processing stretches of discourse such as short monologues for several minutes in order to "scan" for certain information. Assessment tasks in selective listening could ask students, for example, to listen for names, numbers, a grammatical category, directions (in a map exercise), or certain facts and events.

The techniques used in selective listening are:

a. Listening Cloze

Listening cloze tasks/ cloze dictations or partial dictations require the test-taker to listen to a story,

monologue, or conversation and simultaneously read the written text in which selected words or phrases have been deleted.

The listening cloze technique has a weakness that it may simply become reading comprehension tasks because this kind of task is most commonly associated with reading only.

The example of the task:

Test-takers hear:

Ladies and gentlemen, I now have some connecting gate information for those of you making connections to other flights out of Jakarta.

Flight_____ to Bandung will depart from gate _____ at _____ p.m. Flight to Padang will depart at _____ p.m, from gate _____. Flight _____ to Surabaya will depart at _____ p.m. from gate _____. And flight _____ to Bali will depart from gate _____ at _____ p.m.

Test-takers write the missing words or phrases in the blanks.

(the words: seven-oh-six, seventy-three, nine-thirty, ten-forty-five, nine-fifty, seventeen, four-forty, nine-thirty-five, sixty, sixteen-oh-three, nineteen, ten-fifteen).

The difference of listening cloze from standard reading cloze is that the deletions are governed by the objective of the test, not by every *n*th word; and more than one word may be deleted.

b. Information Transfer

In information transfer technique, the information must be transferred to a visual representation, such as labeling a diagram, identifying an element in a picture, completing a form, or showing routes on a map.

The techniques used for this information transfer task are:

- *Information transfer: multiple-picture – cued selection*

Test-takers hear:
Choose the correct picture.

There are two thick books on the table. One pencil is on the book on the right.

Test-takers see:

a. A picture of a table with one thick book and one thin book on it.
b. A picture of a table with one thick book on it.
c. A picture of a table with one thick book and one pencil on it.
d. A picture of a table with two thick books and a pencil on the book on the right.

- *Information transfer: single-picture – cued verbal multiple-choice*

Test-takers see: a photograph of a woman in a laboratory setting, with no glasses on, squinting through a microscope

with her right eye, and with her left eye closed.

Test-takers hear: (a)she's speaking into a microphone
(b) she's putting on her glasses
(c) she has both eyes open
(d) she's using a microscope

- *Information transfer: chart-filling*

Test takers-hear:

Now you will hear information about Diana's daily schedule. The information will be given twice. The first time just listen carefully. The second time, there will be a pause after each sentence. Fill in Diana's blank daily schedule with the correct information. The example has already been filled in.

Diana' schedule

No	Day	Time	Activities
1.	Monday	05.00	Wake up
		05.10	Pray Subuh
		05.30	……..
		…….	Have breakfast
2.			

c. Dictation

Dictation is a widely researched genre of assessing listening comprehension. In a dictation, test-takers hear a passage, typically of 50 to 100 words, recited three times: first, at normal speed; then, with long pauses between phrases or natural word group, during which time test-takers write down what they have just heard; and finally, at normal speed once more so they can check their work and proofread.

There are some difficulties in dictation. First, the difficulty of a dictation task can be easily manipulated by the length of the word groups, the length of the pauses, the speed at which the text is read, and the complexity of the discourse, grammar, and vocabulary used in the passage. Second is the difficulty in scoring. Scoring the dictation should be depended on the context and purpose by deciding on scoring criteria for several possible kinds of errors:

- Spelling error only, but the word appears to have been heard correctly.
- Spelling and/or obvious misrepresentation of a word, illegible word.
- Grammatical error.
- Skipped words or phrases.
- Permutation of words.
- Additional words not in the original.
- Replacement of a word with an appropriate synonym

Beside the disadvantages above, dictation also has benefits. First, it is practical to administer. Second, there is a moderate degree of reliability in a well-established scoring system. Third, there is a

strong correspondence to other language abilities speaks well for the inclusion of dictation among the possibilities for assessing extensive (or quasi-extensive) listening comprehension.

The example of the techniques used in dictation:

First reading (natural speed, no pauses, test-takers listen for gist):
Most mothers have a good piece of advice: Never go into a supermarket hungry! If you go shopping for food before lunchtime, you'll probably buy more than you plan to. Unfortunately, however, just this advice isn't enough for consumers these days. Modern shoppers need an education in how —and how not — to buy things at the grocery store.

Second reading (slowed speed, pause at each//break, test-takers write):
Most mothers // have a good piece of advice: Never go // into a supermarket // hungry! If you go shopping // for food // before lunchtime, you'll probably buy // more than you plan to. Unfortunately, however, just this advice // isn't enough // for consumers // these days. Modern shoppers // need an education // in how —and how not — // to buy things // at the grocery store.

Third reading (natural speed, test-takers check their work.

d. Communicative Stimulus-Response Tasks

Communicative stimulus-response tasks are another-and more authentic-example of extensive

listening which is found in a popular genre of assessment task. In this kind of listening task, test-taker is presented with a stimulus monologue or conversation and then is asked to respond to a set of comprehension questions.

The following is a typical example of such a task. (Taken from Brown, 2004):

Test-takers hear:

Direction: Now you will hear a conversation between Lynn and her doctor. You will hear the conversation two times. After you hear the conversation the second time, choose the correct answer for questions 1-5 below.

Doctor: Good morning, Lynn. What's the problem?
Lynn : Well, you see, I have a terrible headache, my nose is running, and I'm really dizzy.
Doctor: Okay. Anything else?
Lynn : I've been coughing, I think I have a fever, and my stomach aches.
Doctor: I see. When did this start?
Lynn : Well, let's see, I went to the lake last weekend, and after I returned home I started sneezing.
Doctor: Hmmm. You must have the flu. You should get lots of rest, drink hot beverages, and stay warm. Do you follow me?
Lynn : Well, uh, yeah, but…shouldn't I take some medicine?
Doctor: sleep and rest are as good as medicine when you have the flu.
Lynn : okay, thanks, Dr. Brown.

Test-takers read:

1. What is Lynn's problem?
   a. She feels horrible
   b. She ran too fast at the lake
   c. She's been drinking too many hot beverages

2. When did Lynn's problem start?
   a. When she saw her doctor
   b. Before she went to the lake
   c. After she came home from the lake

3. The doctor said that Lynn _____
   a. Flew to the lake last weekend
   b. Must not get the flu
   c. Probably has the flu

4. The doctor told Lynn _____
   a. To rest
   b. To follow him
   c. To take some medicine

5. According to Dr. Brown, sleep and rest are _____ medicine when you have the flu.
   a. More effective than
   b. As effective as
   c. Less effective than

4. Extensive Listening

Listening to develop a top-down, global understanding of spoken language. Listening for the gist, for the main idea, and making inferences are all part of extensive listening.

a. Note-taking

Note-taking is an authentic listening task that is very appropriate to test students' listening ability. The students are asked to listen to a lecture and take note any important point from the lecturer's explanation.

b. Editing

Editing is another authentic task provides both a written and a spoken stimulus, and requires the test-takers to listen for discrepancies.

The following is the way the task proceeds:

Test-takers read:

The written stimulus material such as a news report, an email from a friend, notes from a lecture, or an editorial in a newspaper.

Test-takers hear:

A spoken version of the stimulus that deviates, in a finite number of facts or opinions, from the original written form.

Test-takers mark:

The written stimulus by circling any words, phrases, facts, or opinions that show a discrepancy between the two versions.

Example: The description of political scandal that can be gotten from a newspaper, then compared to the one from a radio broadcast.

In scoring, the smaller number of specific differences are identified, the more reliable the score is.

c. Interpretive tasks

An interpretive task extends the stimulus material to a longer stretch of discourse and forces the test-taker to infer a response. Potential stimuli include song lyrics, recited poetry, radio/television news report, and an oral account of an experience.

Test-takers are then directed to interpret the stimulus by answering a few questions (in open-ended form), such as:

- Why was the singer feeling sad?
- What events might have led up to the reciting of this poem?
- What do you think the political activists might do next, and why?

It is difficult to get high reliability in scoring this kind of task because there may be more than one correct interpretation.

d. Retelling

In a related task, test-takers listen to a story or news event and simply retell it, or summarize it, either orally or in writing.

In making a good retelling, test-takers should identify first the gist, main idea, purpose, supporting points, and/or conclusion to show full comprehension.

D. Testing Speaking

Speaking is a productive skill that can be directly and empirically observed. Those observations are invariably colored by the accuracy and effectiveness of a test-taker's

listening skill. In productive performance, the oral or written stimulus must be specific enough to elicit output within an expected range of performance such that scoring or rating procedures apply appropriately.

Brown (2004:142) also writes a list of speaking micro and macroskills. The microskills refer to producing the smaller chunks of language such as phonemes, morphemes, words, collocations, and phrasal units. The macroskills imply the speakers's focus on the larger elements: fluency, discourse, function, style, cohesion, nonverbal communication and strategic options.

There are 16 different objectives to assess in speaking as described in micro and macroskills.

Microskills

1. Produce differences among English phonemes and allophonic variants.
2. Produce chunks of language of different lengths.
3. Produce English stress patterns, words in stressed and unstressed positions, rhythmic structure, and intonation contours.
4. Produce reduced forms of words and phrases.
5. Use an adequate number of lexical units (words) to accomplish pragmatic purposes.
6. Produce fluent speech at different rates of delivery.
7. Monitor one's own oral production and use various strategic devices ─ pauses, fillers, self-corrections, backtracking ─ to enhance the clarity of the message.
8. Use grammatical word classes (nouns, verbs, etc), systems (e.g., tense, agreement, pluralization), word order, patterns, rules, and elliptical forms.
9. Produce speech in natural constituents: in appropriate phrases, pause groups, breath groups, and sentence constituents.

10. Express a particular meaning in different grammatical forms.
11. Use cohesive devices in spoken discourse.

Macroskills

12. Appropriately accomplish communicative functions according to situations, participants, and goals.
13. Use appropriate styles, registers, implicature, redundancies, pragmatic conventions, conversation rules, floor-keeping and –yielding, interrupting, and other sociolinguistic features in face-to-face conversations.
14. Convey links and connections between events and communicate such relations as focal and peripheral ideas, events, and feelings, new information, and given information, generalization and exemplification.
15. Convey facial features, kinesics, body language, and other nonverbal cues along with verbal language.
16. Develop and use a battery of speaking strategies, such as emphasizing key words, rephrasing, providing a context for interpreting the meaning of words, appealing for help, and accurately assessing how well your interlocutor is understanding you.

Furthermore, Brown (2004) divides the assessment of speaking into five:

1. Imitative Speaking

At one end of a continuum of types of speaking performance is the ability to simply parrot back (imitate) a word or phrase or possibly a sentence. The imitative speaking tasks are as follow:

a. Word repetition task

In a simple repetition task, test-takers repeat the stimulus, whether it is a pair of words, a sentence, or perhaps a question.

Study the following example:

Test-takers hear: Repeat after me:

Beat [pause] bit [pause]
Bat [pause] vat [pause]

I bought a boat yesterday
The glow of the candle is growing

When did they go on vacation?
Do you like coffee?

Test-takers repeat the stimulus.

The above task can be scored by using the following scoring scale.

2	Acceptable pronounciation
1	Comprehensible, partially correct pronounciation
0	Silence, seriously incorrect pronunciation

One example of this kind of test is Phonepass test. **The PhonePass** test is an example of a popular test that uses imitative production task. It elicits computer-assisted oral production over a telephone. Test-takers read aloud, repeat sentences, say words, and answer questions.
The test has five sections.

PhonePass test specifications

Part A:
Test-takers read aloud selected sentences from among those printed on the test sheet. Examples:

1. Traffic is a huge problem in Southern California.
2. The endless city has no coherent mass transit system.
3. Sharing rides was going to be the solution to rush-hour traffic.
4. Most people still want to drive their own cars, though.

Part B:
Test-takers repeat sentences dictated over the phone. Examples: "Leave town the next train."

Part C:
Test-takers answer questions with a single word or a short phrase of two or three words. Example: "Would you get water from a bottle or a newspaper?"

Part D:
Test-takers hear three word groups in random order and must link them in a correctly ordered sentence. Example: was reading/my mother/a magazine.

Part E:
Test-takers have 30 seconds to talk about their opinion about some topic that is dictated over the phone. Topics center on family, preferences, and choices.

Scores for the PhonePass test are calculated by a computerized scoring template and reported back to the test-takers within minutes. There are six scores given:

- An overall score between 20 and 80
- Five subscores on the same scale that rate pronunciation, reading fluency, repeat accuracy, repeat fluency, and listening vocabulary.

2. Intensive Speaking

A second type of speaking frequently employed in assessment contexts is the production of short stretches (not more than a sentence) of oral language designed to demonstrate competence in a narrow band of grammatical, phrasal, lexical, or phonological relationships. Part C and D of the PhonePass test fulfill the criteria of intensive tasks.

The tasks for intensive speaking test are:

a. Directed Response Tasks
Example:
Directed response
Test-takers hear:
Tell me he went to the school
Tell me that you like pop music
Tell me that you aren't interested in Basketball
Tell her to visit me next week
Remind me to check the report soon

b. Read-Aloud Tasks

This task includes reading beyond the sentence level up to a paragraph or two. This

technique has advantages in the administering and scoring because both aspects are relatively easy to be done.

c. Sentence/Dialogue Completion Tasks and Oral Questionnaires

Test-takers are acquired to read dialogue in which one speaker's lines have been omitted. First, the test-takers are given time to read through the dialogue and think the possible answers to fill it. Then as the tape, teacher, or test administrator produces one part orally, the test-takers respond.
Here is the example:

Dialogue completion task

Test-takers read (and then hear):
In a department store:

Salesperson:	May I help you?
Customer :	_____
Salesperson:	Okay, what size do you wear?
Customer :	_____
Salesperson:	Hmmm. How about this green sweater here?
Customer:	_____
Salesperson:	Oh, well, if you don't like green, what color would you like?
Customer:	_____
Salesperson:	How about this one?
Customer:	_____
Salesperson:	Great!

Customer: _____
Salesperson: It's on sale today for $39.95
Customer: _____
Salesperson: Sure. We take Visa, MasterCard, and American Express.
Customer: _____

Test-takers respond with appropriate lines

d. Picture-Cued Tasks

Picture-cued stimulus requires a description from the test-taker. The pictures may be very simple, designed to elicit a word or a phrase.

Example:

- Picture-cued elicitation of minimal pairs

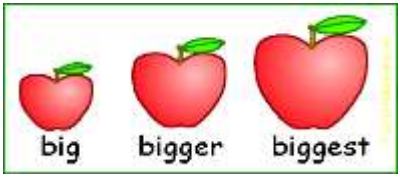
Test-takers see:

(source: www.odopod.com)

Test-takers hear: [test administrator points to each picture in succession]
What's this?

- Picture-cued elicitation of comparatives

Test-takers see:



Test-takers hear: use a comparative form to compare these objects.

- Picture-cued elicitation of future tense

Test-takers see: a picture of a family going on their vacation to Hawaii. Each member imagines the interesting activities they will do at their destination.
Test-takers hear: This family is at an airport going on their vacation.

1. [point to the picture in general] Where are they going for their vacation?
2. [point to the father] What will he do in Hawaii?
3. [point to the girl] What is she going to do there?

- Picture-cued elicitation of nouns, negative responses, numbers, and location

Test-takers see: (some pictures in a classroom)

Test-takers hear:

1. [point to the table] What's this?
2. [point to the blackboard] what is this?
3. [point to the teacher] is he a teacher?

- Picture-cued elicitation of responses and description

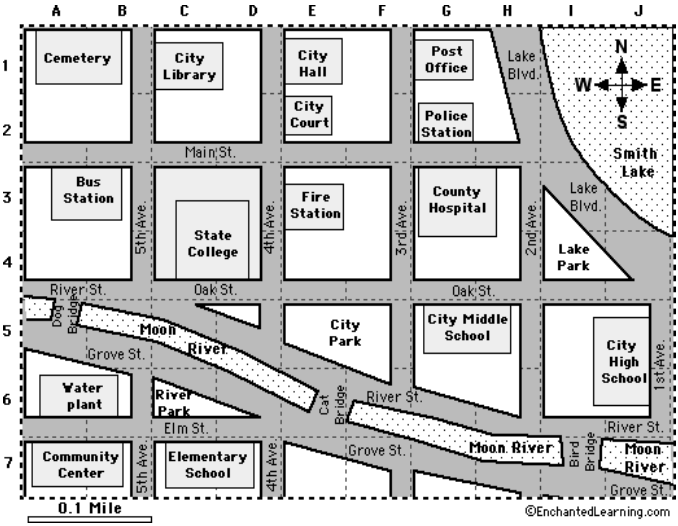
Test-takers see: (some pictures of paintings with its description)

Test-takers hear:

1. [point to the painting on the right] when was this one painted?
2. [point to both] which painting is older?
3. Which painting would you buy? Why?

- Map-cued elicitation of giving directions

Test-takers see:



Test-takers hear:
You are at Main Street [point to the spot]. People ask you for directions to get to five different places. Listen to their questions, then give directions.

1. Please give me directions to the City Park.
2. Please give me directions Community Center.
3. Please tell me how to get to the Bus Station.

Scoring responses on picture-cued intensive speaking tasks varies, depending on the expected performance criteria. The tasks that asked just for one-word or simple-sentence responses can be evaluated simply as *correct* or *incorrect.*

The following rubric can also be used with modifications:

2	Comprehensible; acceptable target form
1	Comprehensible; partially correct target form
0	Silence; or seriously incorrect target form

e. Translation (of Limited Stretches of Discourse)

The test takers are given a native language word, phrase, or sentence and are asked to translate it. As an assessment procedure, the advantages of translation lie in its control of the output of the test-takers, which of course means that scoring is more easily specified.

3. Responsive Speaking

Responsive assessment tasks include interaction and test comprehension but at the somewhat limited level of very short conversations, standard greetings and small talk, simple requests and comments, and the like.

The tasks of responsive speaking are:

a. Question and Answer

Question-and-answer tasks can consist of one or two questions from an interviewer, or prepare a battery of questions and prompts in an oral interview.

The questions and answers form are:

- *Questions eliciting open-ended responses*
  Test takers hear:

1. What do you think about the weather today?
2. What do you like about the English language?
3. Why did you choose your academic major?

Test-takers respond with a few sentences at most.

- *Elicitation of questions from the test-takers*
  Test-takers hear
  - Do you have any questions for me?
  - Ask me about my family or job or interests.

  Test-takers respond with questions

b. Giving Instruction and Directions

The technique is simple: the administrator poses the problem, and the test-takers responds.

Example:

Eliciting instructions or directions
Test-takers hear:
1. Describe how to make an omelet?
2. How do I go to the Post Office from here?
3. What's a good recipe for making a cake?

Test-takers respond with appropriate instructions/directions.

c. Paraphrasing

The test-takers hear a limited number of sentences (perhaps two to five) and produce a paraphrase of the sentence.

For example:

- Paraphrasing a story

Test-takers hear: Paraphrase the following little story in your own words.

Test-takers respond with two or three sentences

- Paraphrasing a phone message

Test-takers hear:

Please tell Josie that I'm tied up in traffic so I'm going to be about a half hour late for the nine o'clock meeting.

Test-takers respond with two or three sentences.

The example of this test is Test of Spoken English (TSE). The TSE is a 20-minute audiotaped test of oral language ability within an academic or professional environment.

The tasks are designed to elicit oral production in various discourse categories rather than in selected phonological, grammatical, or lexical targets. For example:

- ✓ Describe something physical
- ✓ Narrate from presented material
- ✓ Summarize information of the speaker's own choice
- ✓ Give directions based on visual material
- ✓ Give an opinion, etc.

4. Interactive Speaking

Interactive tasks are what some would describe as interpersonal speech events.

The tasks are:

a. Interview

A test administrator and a test-taker sit down in a direct face-to face exchange and proceed through a protocol of questions and directives.

The interview is scored on one or more parameters such as: accuracy in pronunciation and/or grammar, vocabulary usage, fluency, sociolinguistic/pragmatic appropriateness, task accomplishment, and comprehension.

Michael Canale in Brown (2004) suggests four stages of interview, as follow:

1). Warm-up

Preliminary small talk for about one or two minutes. Samples of questions are:

- How are you?
- What's your name?
- What country are you from?

2). Level check

This stage check the test-taker's readiness to speak, confidence, etc. The questions are:

- Tell me about your family
- What is your academic major?
- What are your hobbies?
- What will you be doing ten years from now?

3). Probe

Probe questions challenge test-takers to go to the heights of their ability. The questions are:

- What is your opinion about that issue?
- If you were president of your country, what would you like to change about your country?

- What career advice would you give to your younger friends?
- What are your goals for taking the English program?

4). Wind-down

It is simply a short period of time during which the interviewer encourages the test-taker to relax with some easy questions.

The samples of questions are:

- Did you feel okay about this interview?
- Do you have any questions you want to ask me?
- It was interesting to talk with you. Best wishes.

b. Role Play

Role plays open some windows of opportunity for test-takers to use discourse that might otherwise be difficult to elicit. It frees the students to be somewhat creative in their linguistic output.

Example:

- Pretend that you're a tourist asking me for directions.
- You're buying a new bag from me in *Matahari* Department store, and you want to get a lower price."

c. Discussion and Conversation

As formal assessment devices, discussions and conversations with and among students are difficult to specify and score. But as informal techniques to assess learners, they offer a level of authenticity and spontaneity that other assessment techniques may not offer.

d. Games

There are a variety of games that directly involve language production.

Example:

Assessment games

1. "Tinkertoy" game: A Tinkertoy (or Lego block) structure is built behind a screen. One or two learners are allowed to view the structure. In successive stages of construction, the learners tell "runners" (who can't observe the structure) how to re-create the structure. The runners then tell "builders" behind another screen how to build the structure. The builders may question or confirm as they proceed, but only through the two degrees of separation. Object: re-create the structure as accurately as possible.
2. Crossword puzzles are created in which the names of all members of a class are clued by obscure information about them. Each class member must ask questions of others to determine who matches the clues in the puzzle.
3. Information gap grids are created such that class members must conduct mini-interviews of other classmates to fill in boxes, e.g., "born in July," "plays the guitar," has a two-year-old brother," etc.
4. City maps are distributed to class members. Predetermined map directions are given to one student who, with a city map in front of him or her, describes the route to a partner, who must then trace the route and get to the correct final destination.

5. Extensive Speaking

Extensive speaking tasks involves complex, relatively lengthy stretches of discourse. They are frequently variations on monologues, usually with minimal verbal interaction. This kind of assessment includes more transactional speech events.

The tasks are:

a. Oral Presentation

In academic and professional arenas, the oral presentation can be the presentation of a report, a paper, a marketing plan, a sales idea, a design of a new product, or a method. The following is the example of a checklist for a prepared oral presentation at the intermediate or advanced level of English.

Oral presentation checklist

Evaluation of oral presentation

Assign a number to each box according to your assessment of the various aspects of the speaker's presentation.

3 Excellent
2 Good
1 Fair
0 Poor

Content:

☐ The purpose or objective of the presentation was accomplished.
☐ The introduction was lively and got my attention.
☐ The main idea or point was clearly stated

toward the beginning.

☐ The supporting points were
- Clearly expressed
- Supported well by facts, argument

☐ The conclusion restated the main idea or purpose

Delivery:

☐ The speaker used gestures and body language well.

☐ The speaker maintained eye contact with the audience.

☐ The speaker's language was natural and fluent.

☐ The speaker's volume of speech was appropriate.

☐ The speaker's rate of speech was appropriate

☐ The speaker's pronunciation was clear and comprehensible.

☐ The speaker's grammar was correct and didn't prevent understanding.

☐ The speaker used visual aids, handouts, etc effectively.

☐ The speaker showed enthusiasm and interest.

☐ [if appropriate] The speaker responded to audience questions well.

b. Picture-Cued Story-Telling

At this level, we consider a picture or a series of pictures as a stimulus for a longer story or description.

Example:

Picture-cued story-telling task (Brown, 2004)

Test-takers see the following nine-picture sequence:

Test-takers hear or read:
Tell the story that these pictures describe.

Test-takers use the pictures as a sequence of cues to tell a story.

c. Retelling a story, News Event

Test-takers hear or read a story or news event that they are asked to retell.

d. Translation (of Extended Prose)

In this case, longer texts are presented for the test-takers to read in the native language and then translate into English. The texts can be in various forms, such as: dialogue, directions for assembly of a product, a synopsis of a story or play or movie, directions on how to find something on a map, and others.

E. Testing Reading

In foreign language learning, reading is likewise a skill that teachers simply expect learners to acquire. There are two primary concepts that must be cleared to become efficient readers, they are: 1. They need to be able to master fundamental **bottom-up** strategies for processing separate letters, words, and phrases, as well as **top-down**, conceptually driven strategies for comprehension, 2. A part of top-down approach, second language readers must develop appropriate content and formal schemata ─ background information and cultural experience ─ to carry out those interpretation effectively.

Brown (2004:186) divides the types (genres) of reading as follow:

1. Academic reading

This includes:

- ✓ General interest articles (in magazines, newspapers,etc).
- ✓ Technical reports, professional journal articles
- ✓ Reference material

- ✓ Textbooks, theses
- ✓ Essays, papers
- ✓ Test directions
- ✓ Editorials and opinion writing

2. Job-related reading

This includes:

- ✓ Messages
- ✓ Letters/emails
- ✓ Memos
- ✓ Reports
- ✓ Schedules, labels, signs, announcements
- ✓ Forms, applications, questionnaires
- ✓ Financial documents
- ✓ Directories
- ✓ Manual, directions

3. Personal reading

This includes:

- ✓ Newspapers and magazines
- ✓ Letters, emails, greeting cards, invitations
- ✓ Messages, notes, lists
- ✓ Schedules
- ✓ Recipes, menus, maps, calendars
- ✓ Advertisements
- ✓ Novels, short stories, jokes, drama, poetry
- ✓ Financial documents
- ✓ Forms, questionnaires, medical reports, immigration documents
- ✓ Comic strips, cartoons

Like listening and speaking, reading also has micro and macro-skills which represent the spectrum of possibilities for objectives in the assessment of reading comprehension.

Brown (2004:187) mentions the micro and macroskills for reading comprehension:

Microskills

1. Discriminate among the distinctive graphemes and orthographic patterns of English.
2. Retain chunks of language of different lengths in short-term memory.
3. Process writing at an efficient rate of speed to suit the purpose.
4. Recognize a core of words, and interpret word order patterns and their significance.
5. Recognize grammatical word classes (nouns, verbs, tense, agreement, pluralization), poattern, rules.
6. Recognize that a particular meaning may be expressed in different grammatical forms.
7. Recognize cohesive devices in written discourse and their role in signaling the relationship between and among clauses.

Macroskills

8. Recognize the rhetorical forms of written discourse and their significance for interpretation.
9. Recognize the communicative functions of written texts, according to form and purpose.
10. Infer context that is not explicit by using background knowledge.
11. From described events, ideas, etc, infer links and connection between events, deduce causes and effects, and detect such relations as main idea,

supporting idea, new information, given information, generalization, and exemplification.
12. Distinguish between literal and implied meanings.
13. Detect culturally specific references and interpret them in a context of the appropriate cultural schemata.
14. Develop and use a battery of reading strategies, such as scanning and skimming, detecting discourse markers, guessing the meaning of words from context, and activating schemata for the interpretation of texts.

The following is some principal strategies for reading comprehension. (Brown, 2004:188):

1. Identify your purpose in reading a text.
2. Apply spelling rules and conventions for bottom-up decoding.
3. Use lexical analysis (prefixes, roots, suffixes, etc) to determine meaning.
4. Guess at meaning (of words, idioms, etc) when you aren't certain.
5. Skim the text for the gist and for main ideas.
6. Scan the text for specific information (names, dates, key words).
7. Use silent reading techniques for rapid processing.
8. Use marginal notes, outlines, charts, or semantic maps for understanding and retaining information.
9. Distinguish between literal and implied meaning.
10. Capitalize on discourse markers to process relationships.

Brown (2004) in his book divides the assessment of reading into four categories, they are:

1. Perceptive reading

Perceptive reading tasks involve attending to the components of larger stretches of discourse: letters, words, punctuation, and other graphemic symbols. Bottom-up processing is implied here, and focused mostly in form.

The tasks are:

a. Reading aloud

The test-takers see separate letters, words, and/or short sentences and read them aloud, one by one, in the presence of an administrator.

b. Written response

In this case, the test-takers are given a task to reproduce the probe in writing. If an error occurs, the test-makers should determine its source. For example, what might be assumed to be a writing error may actually be a reading error, or vise versa.

c. Multiple-choice

Here are some possibilities:

- ✓ Minimal pair distinction

  Test-takers read: Circle "S" for same or "D" for different.

1.	Led	let	S	D
2.	Bit	bit	S	D
3.	Seat	sit	S	D
4.	Too	to	S	D
5.	Meet	meat	S	D

- ✓ Grapheme recognition task

Test-takers read: circle the "odd" item, the one that doesn't belong.

1.	Piece	peace	piece
2.	Book	book	boot
3.	Let	led	let

d. Picture-cued items

Test-takers are shown a picture along with a written text and are given one of a number of possible tasks to perform.

The tasks are as follow:

✓ Picture-cued word identification

Test-takers hear: Point to the word that you read here.

Cat clock chair

✓ Picture-cued sentence identification

Test-takers hear: point to the part of the picture that you read about here.

Test-takers see the picture and read each sentence written on a separate card.

The man is reading a newspaper

The cat is under the chair

✓ Picture-cued true/false sentence identification

Test-takers read:

1. The pencils are under the table T F
2. The cat is under the chair T F
3. The picture is one the wall T F

- ✓ Picture-cued matching word identification
  Test-takers read:

1. Clock
2. Glass
3. Spoon
4. Dog
5. Chair

- ✓ Multiple-choice picture-cued word identification
  Test-takers read: square

  Test-takers see, and choose the correct item

2. Selective Reading

A combination of bottom-up and top-down processing may be used.

a. Multiple-choice (for form-focused criteria)

The most straightforward multiple-choice items may have little context, but might serve as a vocabulary or grammar check.

Example:

1). The cat is _____ the room.

a. under
b. between
c. around
d. in

2). He's not married. He's _____

a. single
b. young
c. a husband
d. first

3). Manager: Do you like to work by yourself?
Employee: Yes, I like to work _____
a. independently
b. definitely
c. impatiently
d. dependently

4). I've lived in England (20) _____ three years. I (21) _____ live in Indonesia. I (22) _____ speak any English. I used to (23) _____ homesick, but now I enjoy (24)_____ here. I have never (25) _____ back home (26) _____ I came to England, but I might (27) _____ to visit my family soon.

20. a. since
b. for
c. during

21. a. used to
b. use to
c. was

22. a. couldn't
b. could
c. can

23. a. been

24. a. live
b. to live
c. living

25. a. be
b. been
c. was

26. a. when
b. while
c. since

27. a. go

b. be	b. will go
c. being	c. going

b. Matching tasks

The most frequently appearing criterion in matching procedures is vocabulary. The format is as follow:

Direction: Write in the letter of the definition on the right that matches the word on the left.

_____ 1. Exhausted	a. unhappy
_____ 2. Disappointed	b. understanding of others
_____ 3. Enthusiastic	c. tired
_____ 4. Empathetic	d. excited
	e. enjoy

c. Editing tasks

Editing for grammatical or rhetorical errors is a widely used test method for assessing linguistic competence in reading. Here is a typical set of examples of editing:

Direction: choose the letter of the underlined word that is not correct.

There are two <u>way</u> of <u>making</u> a gas
A **B**
<u>condense</u>: cooling it or <u>putting</u> it under
C **D**
pressure.

d. Picture-cued tasks

The methods that are commonly used are:

1). Test-takers read a sentence or passage and choose one of four pictures that is being described.

2). Test takers read a series of sentences or definitions, each describing a labeled part of a picture or a diagram.

e. Gap-filling tasks

An extension of simple gap-filling tasks is to create sentence completion items where test-takers read part of a sentence and then complete it by writing a phrase.
For example:

Melani : Doctor, what should I do if I get sick?
Doctor : It is best to stay home and _____
You should drink as much _____
You should also _____
This task requires both reading and writing performance and has low validity of reading test.

3. Interactive reading

Top-down processing is typical of such tasks, although some instances of bottom-up performance may be necessary

a. Cloze tasks

Cloze procedure is very popular for this task. Cloze procedure itself can be divided into: cloze procedure, fixed-ratio deletion (every n^{th} word); Cloze-procedure, rational deletion; c-test procedure, and cloze-elide procedure.

1). Cloze procedure, fixed-ratio deletion (every fifth word).

Today I received some (1)_____from Malaysia. There were (2)_____beautiful stamps on the (3)_____. I took them off (4)_____gave them to my (5)_____. He collects stamps from (6)_____over the world.

2). Cloze procedure, rational deletion (preposition and conjuction)

Today I received some letters from Malaysia. There were some beautiful stamps (1) _____ the envelopes. I took them off (2) _____ gave them to my brother. He collects stamps from all over the world.

3). C-test procedure

Today I rece---- some letters fr-- Malaysia. There were some beautiful sta--- on the envelopes. I to-- them off and ga-- them to my brother. He col----- stamps from all over the world.

4). Cloze elide procedure

Today I received some letters from Malaysia. There were some beautiful stamps on the envelopes. I took them off and then gave them to my brother. He just collects stamps from all over the world.

Cloze elide procedure has two disadvantages:

1). neither the words to insert nor the frequency of insertion appears to have any rationale.

2). Fast and efficient readers are not adept at detecting the intrusive words.

b. Impromptu Reading Plus Comprehension Questions

This is a typical reading comprehension task where the test-takers are given a passage and answer some questions.

For example:

Text 1 (question 1-10)

Sometimes mail arrives at the post office, and it is impossible to deliver the mail. Perhaps there is an inadequate or illegible address and no return address. The post office cannot just throw this mail away, so this becomes "dead mail." This dead mail is sent to one of the U.S. Postal Service's dead-mail offices in Atlanta, New York, Philadelphia, St. Paul, or San Francisco. Seventy-five million pieces of mail can end up in the dead-mail office in one year.

The staff of the dead-mail offices has a variety of ways to deal with all of these pieces of dead mail. First of all, they look for clues that can help them deliver the mail; they open packages in the hope that something inside will show where the package came from or is going to. Dead mail will also be listed on a computer so that people can call in and check to see if a missing item is there.

However, all of this mail cannot simply be stored forever; there is just too much of it. When a lot of dead mail has piled up, the dead-mail offices hold public auctions. Every three months, the public is invited in and bins containing items found in dead-mail packages are sold to the highest bidder.

1. The best title for the passage is
   (A). the U.S Postal Service
   (B). staff responsibilities at the U.S Postal Service
   (C). why mail is undeliverable
   (D). dead-mail offices

2. According to the passage, how many dead-mail offices does the U.S Postal Service have?
   (A). 3
   (B). 5
   (C). 15
   (D). 75

3. The word "illegible" in line 2 is closest in meaning to which of the following
   (A). incorrect
   (B). unreadable
   (C). missing
   (D). incomplete

4. Which of the following is NOT mentioned as a way that post office staff members deal with dead mail?
   (A). they search for clues
   (B). they open dead mail
   (C). they throw dead mail away
   (D). they list dead mail on a computer

5. The word "auctions" in line 16 is closest in meaning to
   (A). meetings
   (B). deliveries
   (C). sales
   (D). demonstrations

c. Question-Answer Tasks

This is an alternative task that a teacher can give to his students. For example:

Open-ended reading comprehension questions

1. What do you think the main idea of this passage is?
2. What would you infer from the passage?

d. Editing (longer texts)

This technique is applied to longer passages of 200 to 300 words. The advantages are: authenticity is increased, the task stimulates proofreading one's own essay, the test designer can draw up specifications for a number of grammatical and rhetorical categories that match the content of the course.

e. Scanning

Assessment of scanning is carried out by presenting test-takers with a text and requiring rapid identification of relevant bits of information.

f. Ordering tasks

Sentence-ordering task

Put the following sentences in the correct order:

A. It was called "The last Waltz"
B. The street was in total darkness
C. Because it was one he and Richard had learnt at school
D. Peter looked outside
E. He recognized the tune
F. And it seemed deserted
G. He thought he heard someone whistling

g. Information transfer: Reading charts, maps, graphs, and diagrams

Converting such nonverbal input into comprehensible intake requires not only an understanding of the graphic and verbal conventions of the medium but also a linguistic ability to interpret that information to someone else.

Reading a map is necessary to tell someone where to turn, how far to go, and what place to be reached.

4. Extensive Reading

Extensive reading applies to texts of more than a page, up to and including professional articles, essays, technical reports, short stories, and books. Some tasks described in interactive reading can also apply here, they are: impromptu reading plus comprehension questions, short-answer tasks, editing, scanning, ordering, information transfer, and interpretation.

The tasks are:

a. Skimming tasks

Like the previous category, skimming can apply to a longer text. Readers read a text fast, try to get a sense of the topic, the purpose of the text, the organization, the view, etc. Then, the readers can answer several questions such as:

- What is the main idea of this text?
- What kind of writing is this?
- What do you think you will learn from the text?

b. Summarizing and responding

Summarizing is one of most common tasks given in extensive reading.

For example:

Direction: *Write a summary of the text. Your summary should be about one paragraph length, and should include your understanding of the main idea and supporting ideas.*

Evaluating summary is difficult. The following is the criteria for the evaluation of a summary. (Imao, 2001, p.184)

1. Express accurately the main idea and supporting ideas.
2. Is written in the student's own words; occasional vocabulary from the original text is acceptable.
3. Is logically organized.
4. Display facility in the use of language to clearly express ideas in the text.

Summarizing requires a synopsis or overview of the text, while **responding** asks the reader to provide his or her own opinion on the text as a whole or on some statement or issue within it.

In responding task, the students are asked to write an essay based on an issue from the text, and support their opinion with information from the article and from their own experience.

c. Note-taking and outlining

These are included in informal assessment of reading because it is difficult to control the conditions and time frame for both techniques. A teacher, perhaps in one-on-one conferences with students can use students' notes/outlines as

indicators of the presence or absence of effective reading strategies, and thereby point the learners in positive directions.

F. Testing Writing

There are three genres of writing, they are:

1. Academic writing
   - Papers and general subject reports
   - Essays, compositions
   - Academically focused journals
   - Short-answer test responses
   - Technical reports
   - Theses, dissertations
2. Job-related writing
   - Messages
   - Letters/emails
   - Memos
   - Reports
   - Schedules, labels, signs
   - Advertisements, announcements, manuals
3. Personal writing
   - Letters, emails, greeting cards, invitations
   - Messages, notes
   - Calendar entries, shopping lists, reminders
   - Financial documents
   - Forms, questionnaires, medical reports, immigration documents
   - Diaries, personal journals
   - Fiction

Types of writing performance

1. Imitative Writing

This category includes the ability to spell correctly and to perceive phoneme-grapheme correspondences in the English spelling system.

2. Intensive (controlled)

   Meaning and context are some importance in determining correctness and appropriateness, but mostly focused on form.

3. Responsive Writing

   The learners are required to perform at a limited discourse level, connecting sentences into a paragraph and creating a logically connected sequence of two or three paragraphs.

4. Extensive Writing

   It implies successful management of all the processes and strategies of writing for all purposes, up to the length of an essay, a term paper, a major research paper, and even a thesis.

Brown (2004:221) mentions the micro and macroskills of writing:

Microskills

1. Produce graphemes and orthographic patterns of English.
2. Produce writing at an efficient rate of speed to suit the purpose.
3. Produce an acceptable core of words and use appropriate word order patterns.
4. Use acceptable grammatical systems (e.g. tense, agreement, pluralization), patterns, patterns, and rules.
5. Express a particular meaning in different grammatical forms.
6. Use cohesive devices in written discourse.

Macroskills

7. Use the rhetorical forms and conventions of written discourse.
8. Appropriately accomplish the communicative functions of written texts according to form and purpose.
9. Convey links and connections between events, and communicative such relations as main ideas, supporting idea, new information, given information, generalization, and exemplification.
10. Distinguish between literal and implied meanings when writing.
11. Correctly convey culturally specific references in the contexts of the written text.
12. Develop and use a battery of writing strategies, such as accurately assessing the audience's interpretation, using prewriting devices, writing with fluency in the first drafts, using paraphrases and synonyms, soliciting peer and instructor feedback, and using feedback for revising and editing.

Testing writing can be divided into four big categories (Brown, 2004):

1. Imitative Writing

   Many beginning level English learners, from young children to older adults, need basic training in and assessment of imitative writing the rudiments of forming letters, words, and simple sentences.

   The tasks are:

   a. Tasks in (hand) writing letters, words, and punctuation

      For example:

1). Copying

The test-takers read: Copy the following words in the spaces marks

Bit	bet	bat	but
___	___	___	___

2). Listening cloze selection tasks
3). Picture-cued tasks
4). Form completion tasks
5). Converting numbers and abbreviations to words.

Test takers hear: fill in the blanks with words

Test takers see:

9:00	_____	4:45	_____
5/3	_____	156 main st.	_____

b. Spelling tasks and detecting phoneme-grapheme correspondences

For example:
1). Spelling tests
2). Picture-cued tasks
3). Multiple-choice techniques

Test-takers read:

Choose the word with the correct spelling to fit the sentence, then write the word in the space provided.

The doorbell rang, but when I went to the door, no one was _____

a. Their
b. There
c. they're
d. thair

4). Matching phonetic symbols.

2. Intensive (controlled) writing
   The tasks are:
   a. Dictation and Dicto-Comp
      Here, a paragraph is read at normal speed, usually two or three times; then the teacher asks students to rewrite the paragraph from the best of their recollection.

   b. Grammatical Transformational tasks
      The tasks can be as follow:
      - Change the tenses in a paragraph
      - Change full forms of verbs to reduced forms (contractions)
      - Change questions into statement, etc.

   c. Picture-cued tasks
      It can be divided into:
      1). Short sentences
      2). Picture description
      3). Picture sequence description

   The tasks in controlled writing can be scored by using the following criteria:

2	grammatically and lexically correct
1	either grammar or vocabulary is incorrect, but not both
0	both grammar and vocabulary are incorrect

   d. Vocabulary Assessment tasks
      For example:
      Test taker read:
      Write two sentences, A and B. In each sentence, use the two words given.

A. interpret, experiment _____
B. interpret, language _____

e. Ordering tasks
Reordering words in a sentence
Test-takers read:
Put the words below into the correct order to make a sentence

1). Cold/winter/is/weather/the/in/the
2). Doing/what/they/are

f. Short answer and sentence completion tasks
Limited response writing tasks
Test takers see:
Ratna : Who's that?
Doni : _____ Gina.
Ratna : where's she from?
Doni : _____ Bandung.

3. Responsive and Extensive Writing (open-ended writing)

In this type of writing test, the testee is given a freedom to express his ideas freely.
The tasks are:

a. Paraphrasing

In paraphrasing, the testee says something in his own words, avoid plagiarizing, and offer some variety in expression.

b. Guided questions and answers

The test makers ask several questions in order based on a story/text given to the testees. Then,

based on the answers, the testee can construct a paragraph or an essay.

c. Paragraph construction tasks
   1). Topic sentence writing
   2). Topic development within a paragraph
   3). Development of main and supporting ideas across paragraphs.

UNIT VI

ASSIGNING GRADES AND COURSE MARKS

A. Scoring tests

According to Green (1975: 159), scoring refers to the process of checking the tests to determine the number of correct and incorrect responses and assigning numerical scores. These scores are called *raw scores*, and they indicate the number of items that pupils have answered correctly. In addition, there are also *derived scores*, such as percentiles and standard scores, which are statistically calculated from the raw scores.

Different types of tests can be scored in the different manner. It is explained further as follow:

1. Objective Tests

Objective tests can be scored most quickly and accurately of all the test types. If an answer sheet is used for the test, a scoring key can be made by punching out the correct responses on a cardboard sheet, which fits over the answer sheets in such a way that the pupils' errors can be marked through the holes in the cardboard.

2. Essay Tests

Essay tests can be scored by using two kinds of scoring methods: point-score method and sorting method. Scorers should carefully prepare the keys in order to improve the reliability. There is no doubt that careless, cursory reading by the evaluator and lack of clearly defined grading criteria have been major factors contributing to the unreliability of the essay test.

3. Performance tests

Since there are several types of performance tests, the methods of scoring these tests vary. Objective performance tests, such as the identification test, can be scored in the same manner as the conventional objective test. Usually the work-sample test can also be set up so that it can be scored numerically, when a specified number of points is allotted to each work-sample station, pupils' scores can be compiled and handled in the conventional manner.

When pupil products are evaluated, however, the scoring procedure is different in that check lists and rating scales are used to increase the reliability of the evaluation.

B. Assigning Grades

This section is focused on principles and methods of grading.

Problems of Grading

There are several problems in grading that are needed to be solved:

1. Course marks frequently do not reflect the actual course achievement of individual pupils.
2. Teachers lack objective, clearly defined criteria for assigning grades.
3. The halo effect frequently influences teachers to grade those whom they like higher than their achievement warrants.
4. Occasionally personality conflicts between a pupil and his teacher cause the pupil to be unfairly penalized when he is assigned a grade.
5. There is a tendency for male teachers to assign higher grades to female pupils than to male pupils for comparable achievement.
6. Course marks are often given on the basis of insufficient data concerning the achievement of pupils.
7. The grade may reflect a cultural bias detrimental to minority group pupils.
8. The basis for assigning grades may not be clear to pupils and parents.

General Principles of Grading

It might be well to mention some general principles that teachers should follow when assigning marks to pupils:

1. The course mark should accurately reflect the level of achievement of each pupil in the class.

   A mark should tell the pupil how successful he has been in achieving the goals or objectives of the class and in mastering the content area studied.
2. Class marks should not be used for disciplinary purposes.

   To correct the disciplinary infractions, teachers should not lower the marks of pupils because it is not effective to solve the problem.
3. Pupils should be acquainted with the grading method that the teacher is using.

   They should know which factors will be weighed in grading and their relative importance.
4. Pupils should be permitted to know their grade progress throughout the term.

   Tests should be returned to pupils and discussed. Then, they can make improvement from the errors they made. Grades may also be posted periodically.
5. If the grades are to have comparability from one class to another within a school system, it is essential that there be staff consensus on the grading policy.

   There should be agreement on the method of reporting marks, the method of assigning marks, and the relative weight of final examinations.
6. If the grading system is norm referenced, the grade distribution will approximate the normal-curve percentages.
7. If the grading system is criterion referenced, the standards should be set at a level that most pupils can reasonably be expected to achieve.

Conventional Methods of Grading

It is necessary that a teacher be familiar with the numerous methods of grading, and it is probably desirable that he uses more than one of the methods in his classes. The methods are:

1. Individual Standard

   Grading based on an individual standard is sometimes referred to as ability grading. An effort is made through a standardized testing program to determine each pupil's relative capacity for learning, and a standard of achievement appropriate to that potential is set for each pupil in the class. Thus, for a comparable grade, less achievement is expected of a dull pupil than of a bright pupil.

2. Fixed Standard

   With fixed-standard grade assignment, grades are assigned on the basis of subject matter standards which are set by the teacher and which represent, in his judgment, appropriate levels of competence.

   Furthermore, the most frequently used method for assigning grades is to convert the raw scores into percent scores and assign letter-grade equivalents.

   The advantage of the fixed-standard is it gives a generally understood meaning to the competence levels that the different letter grades given in a course imply. However, this method has two weaknesses: 1. the standards are variable in that they are actually set by individual teachers. Any teachers can change the standards in his class without the immediate consequence of more than minor repercussions. 2. When the teacher holds too rigidly to a fixed standard, he ignores the difference among pupils and among classes.

3. Group or Norm-Referenced Standard

   Individual pupil's achievement is compared to his group or class rather than judged against a rigid set of subject matter standards. With this method, a curve is used as the basis for assigning grades.

   This method has disadvantage that it tends to overemphasize competition, which arises when pupils vie with one another for the top ranks within the group.

C. Conventional Methods of Assigning Course Marks

General Principles

A course mark represents a composite picture of the grades that have been given on examinations and of the class work during the semester. The assigned marks should be a valid representation of the pupil's achievement. It should be relatively easy if the examinations are good and the class works are graded carefully.

The following suggestions should be useful for teachers in assigning marks:

1. Certainly teachers should not assign marks without collecting sufficient data about the pupils' achievement to permit a valid judgment to be made.
2. It is important that several achievement factors be weighed in a course mark and that the pupil be aware of the factors to be included as well as their relative weight. Factors that should normally be weighed in a course mark include (a) examinations, (b) daily written assignments, (c) review or teaching tests, (d) written research reports, and (e) oral presentations.
3. The final examination should have a significant weight. In general, the final examination should be assigned a weight of from one-fifth to one-third of the

total mark, but the weight should vary according to the course and the objectives of instruction.

Use of Median and Mode

Teachers are most familiar with the arithmetic mean as the basis for assigning the course mark. However, in some cases it can cause a poor representation of the pupil's achievement. For example, a pupil earns the following seven scores—100, 95, 95, 90, 85, 80, and 0. He would get an average score of only 78. Actually, his actual work is of A or B caliber, but his mark based on the mean will probably be C. That's why some teachers avoid this difficulty by eliminating one or two of the pupil's lowest scores from the list to be averaged.

Perhaps a better solution is to rely upon the median rather than the mean as a basis for assigning marks. The median is unaffected by extreme scores because it is the midpoint of the distribution of scores. In the example above the pupil's median grade would be 90, which more truly represent his actual achievement than does the mean of 78.

If there are numerous grades to be averaged, the mode (most frequent score received by the pupil) is also useful, although it is less reliable than the median as a measure of central tendency.

Point-Score Methods

The point-score technique of assigning course marks is useful in classes from the upper elementary grades through the university. There are alternative methods for calculating marks when using the point-score technique:

1. Cumulative-point score

   It may be used with a curve. This method is useful when some of the factors to be included in a final mark are not readily assigned either a letter or

numerical grade. For example, the daily class work can quickly be evaluated as "unsatisfactory" (symbol: ─), "satisfactory" (√), and "superior" (+).

After each pupil's cumulative-point score has been totaled, the mean and standard deviation should be calculated and the final marks assigned on the basis of the calculated curve.

2. Grade-point average

It is easier to use than the cumulative-point score method and is a better system to follow in most courses. With this system pupils are given letter grades on all their assignments and examinations, and the letter grades are translated into the following point scale before being averaged: A=4 points, B=3 points, C=2 points, D=1 point, and F=0 point.

D. New Methods of Grading

For many years, educators at the public elementary and secondary schools and also higher education are finding that they face the conflicting social mandates that they educate pupils who in the past would not have competed successfully within then group, and also that the graduate and certify the level of competence of those pupils. Included in this group of pupils are: (a) the culturally different, (b) the language handicapped, (c) the educationally handicapped, (d) the slow learner, and (e) the non competitive.

Two decades ago, the acceptable answer was to screen out these pupils through academic failure and dropout, but today these methods are increasingly unacceptable. As a result, there is growing acceptance of individualized approaches to instruction and assessment with heavy reliance

upon behavioral objectives, performance-based instruction, criterion-referenced measurement and grading, and continuous-assessment procedures.

The following are the new methods of grading:

1. Criterion-Referenced Grading

The criterion or criteria of success are carefully spelled out in terms of specified minimum levels of acceptable competence, levels that can generally be demonstrated in specific behavior or performance. The emphasis is on successful performance, not failure, and diagnostic measurement is used by the teacher at the beginning to analyze pupil status and establish a base line of competence.

During the instruction, formative measurement helps in pacing the individual pupil's progress, in determining mastery of specified competencies or concepts, and in recycling pupils through identified problem areas. At the end of the instructional period, summative evaluation is used to demonstrate the exit-level competence of each pupil. At this point there are several options open to the teacher (Green, 1975):

a. He may certify on the pupil's record only the specific competencies that have been mastered, assigning no grade for unmastered competencies requiring future instruction.
b. He may use a pass-fail grade, passing all who demonstrate the minimum acceptable level of competence.
c. He may give letter grades. Assigning A grades to all who achieve the specified minimum competence.

d. He may use conventional methods of assigning course marks.

2. Continuous-Assessment Methods

Particularly at the lower elementary level where basic skills are stressed, continuous monitoring or assessment of achievement is carried out. In this instance the record of each subskill, for example, word attack skills in phonics, is evaluated with frequent (daily) tests, and the record is kept graphically with a continuous charting of errors and successes.

This procedure provides both teachers and students constant feedback concerning progress, a type of feedback which may be more meaningful than a letter grade.

REFERENCES

Athanasou James A and Iasonas Lamprianou. 2002. A Teacher's Guide to Assessment. Sydney: Social Science Press.

Bachman, Lyle. F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Brown, H. Douglas. 2004. *Language Assessment:* Principles and Classroom Practices. Longman: Pearson Education, Inc.

Fulcher, Glenn and Fred Davidson. 2007. *Language Testing and Assessment. An advance resource book.* Great Britain: MPG Books Ltd, Bodmin.

Green, John A. 1975. *Teacher-Made Tests* (2nd edition). New York: Harper and Row, Publishers.

Harris, David P. 1980. *Testing English as a Second Language*. New Delhi: McGraw Hill Hook.

Henning, Grant. 1987. *A Guide to Language Testing*: Development, Evaluation, Research. New York: Newbury House Publishers.

Hughes, Arthur. 1989. *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Thornbury, Scott. 1999. *How to Teach Grammar*. Edinburgh: Pearson Education Limited.

Weir, Cyrill J. 1990. *Communicative Language Testing*. New York: Prentice Hall.

Weir, Cyrill. 1993. *Understanding and Developing Language Test*. New York: Prentice Hall.