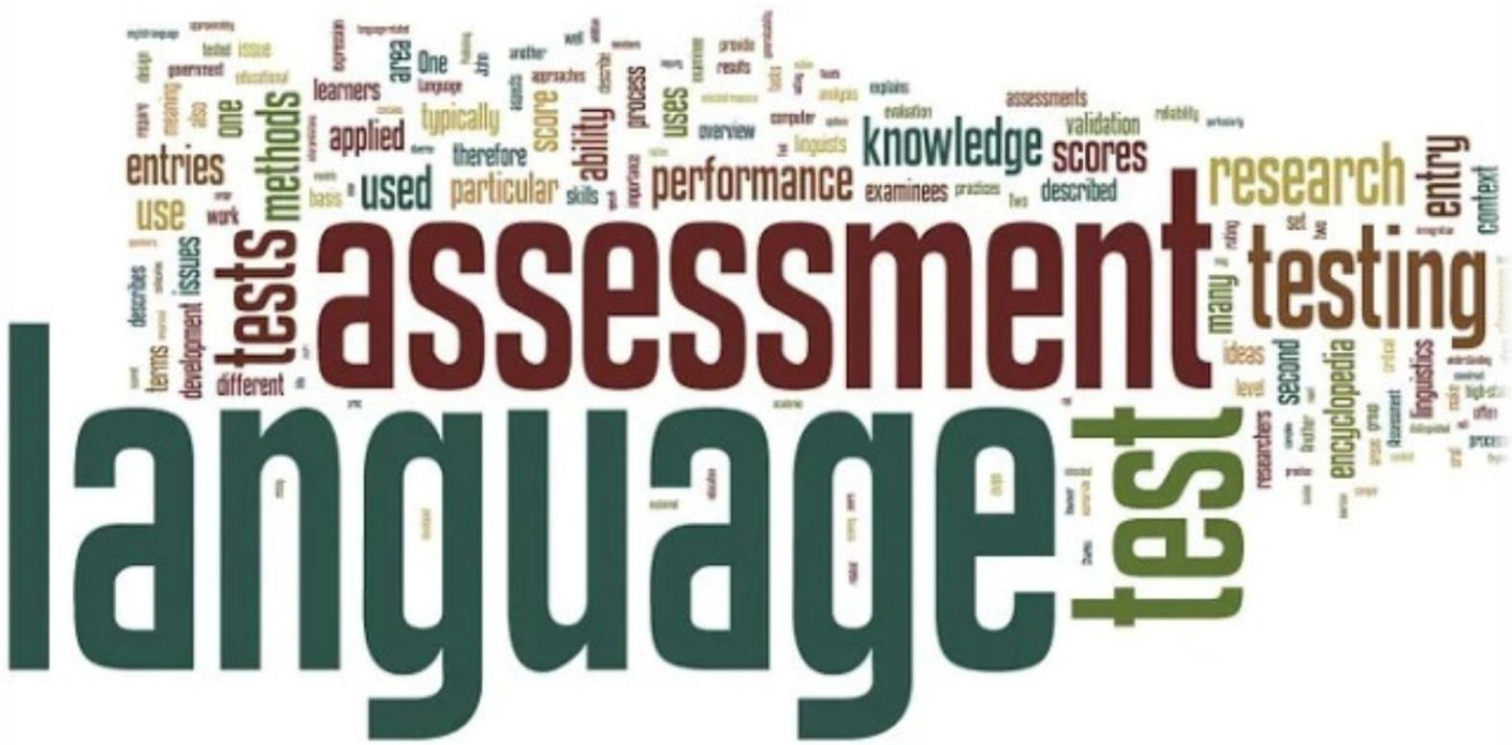


LANGUAGE ASSESSMENT



Mu'man, S.Hum., M.Pd.

ENGLISH

EDUCATION STUDY PROGRAM

LANGUAGE ASSESSMENT

Mu'man, S.Hum., M.Pd.

KATA PENGANTAR

Puji syukur ke hadirat Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya sehingga modul/materi ajar mata kuliah Language Assessment ini dapat disusun dan disajikan kepada mahasiswa program studi Pendidikan Bahasa Inggris. Mata kuliah ini dirancang untuk membekali mahasiswa dengan pemahaman teoritis dan keterampilan praktis dalam menilai kemampuan berbahasa Inggris secara valid, reliabel, dan adil. Dalam pembelajaran ini, mahasiswa akan mempelajari konsep dasar asesmen bahasa, jenis-jenis tes bahasa, prinsip-prinsip evaluasi, penyusunan instrumen tes, analisis hasil tes, hingga praktik implementasi asesmen alternatif dan autentik. Seluruh topik dikemas untuk mengembangkan kemampuan mahasiswa sebagai calon pendidik yang mampu merancang dan mengevaluasi asesmen yang sesuai dengan kebutuhan pembelajaran abad ke-21.

Penyusunan materi ini tidak lepas dari berbagai referensi ilmiah dan praktik terbaik dalam dunia pengajaran bahasa, baik secara nasional maupun internasional. Oleh karena itu, penulis menyampaikan apresiasi kepada semua pihak yang telah memberikan dukungan dalam penyusunan naskah ini. Akhir kata, semoga materi ini dapat menjadi sumber belajar yang bermanfaat dan inspiratif bagi mahasiswa dalam memahami dan menerapkan konsep asesmen dalam pembelajaran bahasa Inggris. Kritik dan saran yang membangun sangat diharapkan untuk penyempurnaan materi ini di masa mendatang.

Cimahi, 29 Februari 2024

Penulis

DAFTAR ISI

KATA PENGANTAR.....	i
DAFTAR ISI.....	ii
PENDAHULUAN	iii
BAB I INTRODUCTION TO THE COURSE	5
BAB II THE FUNDAMENTAL OF LANGUAGE ASSESSMENT	6
BAB III METHOD OF LANGUAGE ASSESSMENT.....	11
BAB IV KIND OF TESTS	16
BAB V ASSESSING VOCABULARY	24
BAB VI ASSESSING GRAMMAR.....	29
BAB VII ASSESSING PRONUNCIATION	31
BAB VIII ASSESSING LISTENING	33
BAB IX ASSESSING SPEAKING	36
BAB X ASSESSING READING	42
BAB XI ASSESSING WRITING.....	48
BAB XII IDENTIFYING RUBRICS	55
BAB XIII ITEMS ANALYSIS	57
BAB XIX VALIDITY & RELIABILITY	58
REFERENCES	62

PENDAHULUAN

A. Identitas Modul

Mata Kuliah : Language Assessment
Kode : 4222622550
Semester : 6 (Genap)
Alokasi Waktu : 16 Pertemuan

B. Capaian Pembelajaran

1. Mampu Menjelaskan secara garis besar seluk beluk kajian evaluasi pembelajaran bahasa inggris (C2, A2)
2. Mampu Menjelaskan berbagai jenis test (Kinds of Test), khususnya instrumen pengukuran yang digunakan dalam pembelajaran bahasa inggris (C2, A2)
3. Mampu Menerangkan prinsip prinsip validitas dan realibilitas (Validity & Reliability) dalam merancang instrumen pengukuran terampilan berbahasa (C2, A2)
4. Mampu merancang instrumen pengukuran kemampuan berbahasa untuk penguasaan kosa kata (Assessing Vocabulary) (C2, A2)
5. Mampu merancang instrumen pengukuran kemampuan berbahasa untuk keterampilan tata Bahasa (Assessing Grammar) (C2, A2)
6. Mampu merancang instrumen pengukuran kemampuan berbahasa untuk Pelafalan dan sikap (Assessing Pronunciations & Attitude) (C2, A2)
7. Mampu merancang instrumen pengukuran kemampuan berbahasa untuk keterampilan menyimak (Assessing Listening) (C2, A2)
8. Mampu Merancang instrumen pengukuran kemampuan berbahasa untuk keterampilan berbicara (Assessing Speaking) (C2, A2)
9. Mampu merancang instrumen pengukurankemampuan berbahasa untuk keterampilan membaca (Assessing Reading) (C2, A2)
10. Mampu merancang instrumen pengukuran kemampuan berbahasa untuk keterampilan menulis (Assessing writing) (C2, A2)
11. Mampu melakukan evaluasi rancangan instrumen pengukuran kemampuan berbahasa di dalam penelitian di ruang kelas (Items Analysis) (C2, A2)

C. Petunjuk Penggunaan Modul

1. Baca secara sistematis setiap bab
Modul ini terdiri dari beberapa bab yang disusun secara berurutan mulai dari pengenalan hingga valididitas dan realiabilitas. Mahasiswa diharapkan membaca setiap bab secara berurutan agar pemahaman dapat terbentuk secara menyeluruh dan bertahap.
2. Perhatikan Tujuan Pembelajaran
Setiap bab diawali dengan tujuan pembelajaran yang harus dicapai. Bacalah dengan saksama tujuan tersebut agar Anda dapat mengarahkan fokus saat mempelajari materi dan mengetahui kompetensi yang harus dikuasai.

3. Gunakan modul sebagai bahan diskusi

Modul ini juga dirancang sebagai bahan diskusi dalam perkuliahan. Mahasiswa dianjurkan untuk mencatat hal-hal penting, mengajukan pertanyaan, serta berdiskusi aktif dengan dosen dan teman sejawat saat sesi tatap muka atau pembelajaran daring.

BAB I

INTRODUCTION TO THE COURSE

Language Assessment is an essential course designed to equip students with both theoretical understanding and practical skills necessary for evaluating English language proficiency in various educational contexts. As future English teachers, students are expected to comprehend how assessment functions as an integral part of the teaching and learning process—providing feedback, guiding instruction, and supporting learners' progress. This course covers a wide range of topics starting with The Fundamental of Language Assessment (Chapter II), which introduces key concepts, purposes, and characteristics of effective assessments. Students will then explore different Methods of Language Assessment (Chapter III), focusing on how assessments can be applied formally and informally, traditionally or authentically.

Following that, students will learn about the Kinds of Tests (Chapter IV), including proficiency, achievement, diagnostic, and placement tests. Subsequent chapters focus on assessing specific language components and skills:

1. Assessing Vocabulary
2. Assessing Grammar
3. Assessing Pronunciation and Attitude
4. Assessing Listening
5. Assessing Speaking
6. Assessing Reading
7. Assessing Writing

In addition, students will be introduced to the construction and use of Rubrics (Chapter XII) as scoring tools that help ensure fairness and clarity in assessment. They will also learn about Item Analysis (Chapter XIII), which allows them to evaluate the quality and effectiveness of test items. Finally, the course concludes with a focus on Validity and Reliability (Chapter XIV), two essential criteria for ensuring that assessments measure what they intend to and produce consistent results over time. Then, students are expected to be able to design and evaluate language assessments that are valid, reliable, practical, and aligned with pedagogical goals and learner needs.

BAB II

THE FUNDAMENTAL OF LANGUAGE ASSESSMENT

The fundamental of language assessment refers to the essential principles and concepts that guide the process of evaluating language skills. It includes understanding the purposes of assessment, such as placement, diagnosis, progress tracking, and achievement measurement. Language assessment must be valid (measuring what it intends to measure), reliable (yielding consistent results), and practical (feasible to administer). Additionally, fairness and authenticity are important to ensure that the assessment reflects real-life language use and provides equal opportunities for all test-takers.

A strong foundation in language assessment also involves knowledge of different assessment types, such as formative and summative, as well as formal and informal methods. Teachers and educators need to design or select appropriate tools that align with the language learning objectives. These tools can include written tests, performance-based tasks, portfolios, and self-assessments. Ultimately, effective language assessment supports learning by providing meaningful feedback, guiding instruction, and motivating learners to improve their language proficiency.

1. Definition of Language Assessment

Language assessment refers to the systematic process of collecting information about a learner's language ability through various methods such as tests, observations, self-assessments, and performance tasks. It aims to measure how well learners understand and use the target language, particularly in relation to communicative competence, linguistic accuracy, fluency, and performance.

2. Purposes of Language Assessment

Understanding the purpose of an assessment is crucial, as it determines the method, design, and interpretation of the results. Language assessments can serve several specific purposes:

1) Diagnostic Assessment

Conducted at the beginning of a course or unit to identify learners' strengths, weaknesses, prior knowledge, and learning needs. It helps teachers tailor instruction effectively.

2) Formative Assessment

Ongoing assessment carried out during the learning process. Its main aim is to provide feedback that guides both teaching and learning. Examples include quizzes, reflections, classroom observations, and draft reviews.

3) Summative Assessment

Administered at the end of a course or instructional unit to evaluate overall learning outcomes. Final exams, end-of-semester projects, and oral presentations are common types.

4) Placement Assessment

Used to assign students to appropriate language levels or classes based on their current proficiency. For example, it helps distinguish whether a student belongs in an elementary or advanced class.

5) Proficiency Assessment

Measures general language ability independent of any curriculum. These are often standardized tests like TOEFL, IELTS, or TOEIC, which assess whether someone can use the language effectively in real-world settings.

6) Achievement Assessment

Designed to assess what learners have specifically learned in a course or instructional program. It is aligned with the curriculum and learning objectives.

3. Principles of Effective Assessment

Effective language assessment should meet key quality criteria to ensure it serves its intended purpose. The following principles are widely accepted:

1) Validity

The degree to which a test measures what it claims to measure. For instance, a speaking test should actually assess speaking ability, not reading or grammar knowledge.

2) Reliability

The consistency of assessment results over time, across different raters, or within the same test. A reliable test yields similar results in repeated administrations under consistent conditions.

3) Practicality

Refers to how feasible an assessment is in terms of time, cost, and administration. A highly valid test may be impractical if it is too expensive or time-consuming.

4) Authenticity

The extent to which assessment tasks reflect real-life language use. Authentic tasks are meaningful, engaging, and closer to real-world communication, such as writing emails or conducting interviews.

5) Washback (Backwash)

The impact that an assessment has on teaching and learning. Positive washback encourages good teaching practices and meaningful learning, while negative washback may promote teaching to the test.

6) Transparency

Learners should clearly understand what is being assessed, the criteria involved, and how they will be graded. Transparency enhances motivation and reduces anxiety.

4. Types of Assessment Approaches

Language assessment can follow two broad approaches:

1) Traditional Assessment

Typically includes standardized formats such as multiple-choice, true/false, or fill-in-the-blank items. Focuses more on discrete language points like grammar or vocabulary.

Pros: Easy to score, objective, and efficient.

Cons: May not reflect real communicative ability.

2) Alternative/Authentic Assessment

Emphasizes real-life tasks and learner involvement. Types include portfolios, self-assessments, peer-assessments, journals, and performance tasks such as oral presentations or group projects.

Pros: More learner-centered, encourages reflection, shows holistic language ability.

Cons: Time-consuming and more subjective in scoring.

5. Components of Language Being Assessed

Language is a complex system, and assessment may target different areas of competence:

1) Linguistic Competence

The ability to use grammar, vocabulary, and pronunciation correctly.

2) Sociolinguistic Competence

The ability to use language appropriately depending on the social and cultural context (e.g., formal vs. informal registers).

3) Discourse Competence

The ability to organize sentences into coherent spoken or written texts. Includes the use of connectors, transitions, and logical flow.

4) Strategic Competence

The ability to manage communication breakdowns using strategies like paraphrasing, asking for clarification, or using gestures when vocabulary is lacking. These components are often integrated into communicative tasks rather than assessed in isolation.

6. The Role of Teachers in Language Assessment

Teachers are central figures in language assessment. Their responsibilities go beyond assigning grades:

1) Test Designers:

Teachers must create valid, clear, and fair assessment tasks aligned with learning objectives.

2) Facilitators of Assessment:

They organize the administration of assessments, ensuring all learners are treated fairly and ethically.

3) Evaluators:

Teachers must score assessments objectively using rubrics or answer keys and consider individual student progress.

4) Feedback Providers:

Offering constructive, actionable feedback is vital for student growth and motivation.

5) Instructional Decision-Makers:

Assessment results help teachers plan future lessons, differentiate instruction, and identify learners who need support or enrichment.

7. Current Trends in Language Assessment

Language assessment practices continue to evolve to reflect pedagogical and technological advancements:

1) Learner-Centered Assessment

Encourages learners to be involved in their own assessment through self-assessment, peer feedback, and goal setting.

2) Technology-Enhanced Assessment

Digital tools like online quizzes, interactive tasks, automatic scoring systems, and virtual classrooms are increasingly used to facilitate assessment.

3) Performance-Based Assessment

Focuses on how well students can use the language in real-world situations, rather than how much they know about the language.

4) Sustainable Assessment

Develops learners' capacity to monitor and evaluate their own learning beyond the classroom, promoting life-long learning skills.

BAB III

METHOD OF LANGUAGE ASSESSMENT

Reasons for testing i.e. the objectives of test-types

1. What basic questions should a teacher/tester ask himself/herself before setting a test?
2. What are the relationships between learning, teaching & testing?
3. What are the main reasons for testing?
4. What are the essential differences between a classroom test and an external examination?
5. Briefly describe the following types of test in terms of their objectives: a) aptitude b) placement c) diagnostic d) achievement e) proficiency.
6. How far is it possible to predict student success in language learning?
7. Can one measure a student's progress in learning a L2? What are the difficulties?
8. How does one specify test objectives? How far does the way objectives are defined influence subsequent assessment?

Explanation.

- 1) What basic questions should a teacher/tester ask himself/herself before setting a test?
What to test. How to do it. Whether to test at all. Why the assessment is being made.
What it should contain. The consequences for teaching, learning and administration.
The quality of the proposed test material.

Explanations

What to test. How to do it. Whether to test at all. Why the assessment is being made.
What it should contain. The consequences for teaching, learning and administration.
The quality of the proposed test material.

The characteristics of a good test are:

- a. Validity - it should measure what it is intended to measure and nothing else.
- b. Reliability - (unless valid it cannot be reliable): if administered a second time a reliable test would result in the same order of merit when neither learning nor teaching has intervened.
- c. Discrimination: Decide first whether the primary purpose is to discriminate between testees. School exams are generally designed to discriminate as widely as possible among the testees.

- d. Backwash: Effects of the test on learning & teaching. Does it have a good influence on the learning & teaching that takes place before the test.
- 2) The characteristics of a good test are:

Validity - it should measure what it is intended to measure and nothing else.

Reliability - (unless valid it cannot be reliable): if administered a second time a reliable test would result in the same order of merit when neither learning nor teaching has intervened.

 - a. Content validity: purposes for assessment: aims of teacher / learner. Content specification = list: ensures that test reflects all areas to be assessed in suitable proportion.
 - b. Balanced sample (nature of items included). Backwash/Washback: influence on learning & teaching before the test. Teaching through testing.
 - c. Discrimination: Decide first whether the primary purpose is to discriminate between testees. School exams are generally designed to discriminate as widely as possible among the testees.
 - d. Backwash: Effects of the test on learning & teaching. Does it have a good influence on the learning & teaching that takes place before the test.
3. What are the main reasons for testing?

Achievement/Attainment tests: usually more formal, designed to show mastery of a particular syllabus (e.g. end-of-year tests, school-leaving exams, public tests) though similar (re-syllabus) to progress tests. Rarely constructed by classroom teacher for a particular class. Designed primarily to measure individual progress rather than as a means of motivating or reinforcing language.

 - a. **Progress Tests**: Most classroom tests take this form. Assess progress students make in mastering material taught in the classroom. Often given to motivate students. They also enable students to assess the degree of success of teaching and learning and to identify areas of weakness & difficulty. Progress tests can also be diagnostic to some degree.
 - b. **Diagnostic Tests** can include Progress, Achievement and Proficiency tests, enabling teachers to identify specific weaknesses/difficulties so that an appropriate remedial programme can be planned. Diagnostic Tests are primarily designed to assess students' knowledge & skills in particular areas before a course of study is begun. Reference back to class-work. Motivation. Remedial work.

- c. **Placement Tests** sort new students into teaching groups so that they are approx. the same level as others when they start. Present standing. General ability rather than specific points of learning. Variety of tests necessary. Reference forward to future learning. Results of Placement Tests are needed quickly. Administrative load.
 - d. **Proficiency Tests** measure students' achievements in relation to a specific task which they are later required to perform (e.g. follow a university course in the English medium; do a particular job). Reference forward to particular application of language acquired: future performance rather than past achievement. They rarely take into account the syllabus that students have followed. Definition of operational needs. Practical situations. Authentic strategies for coping. Common standard e.g. driving test regardless of previous learning. Application of common standard whether the syllabus is known or unknown.
 - e. **Aptitude Tests**: measure students probable performance. Reference forward but can be distinguished from proficiency tests. Aptitude tests assess proficiency in language for language use (e.g. will S experience difficulty in identifying sounds or the grammatical structure of a new language?) while Proficiency tests measure adequacy of control in L2 for studying other things through the medium of that language.
4. What are the essential differences between a classroom test and an external examination?
- Most external exams are designed to discriminate as widely as possible among testees.
5. Briefly describe the following types of test in terms of their objectives: a) aptitude b) placement c) diagnostic d) achievement e) proficiency.
- a. Aptitude has no past and concerns the future: re language performance itself e.g. Modern Language Aptitude Test University of York
 - b. Placement Tests: sort new students into teaching groups so that they are approx. the same level as others when they start.
 - c. Diagnostic concerns the past. It may or may not refer to a known syllabus (e.g. Kernel entry test and tests on each 3 units)
 - d. Achievement has a known syllabus and concerns the future (e.g. "O" & A level or University degree exams)

- e. Proficiency relates to the future: re- Use of Language to undertake a non-language task (e.g. Cambridge Proficiency, TOEFL)
6. How far is it possible to predict student success in language learning?
- Aptitude tests assess proficiency in language for language use i.e. language performance itself. They do not assess the use of language to undertake a non-linguistic task.
- a. Problem one: definition of student success in language learning: should we consider language learning for other goals.
 - b. Problem two: before we can predict success we have to be able to measure it. How do we measure proficiency?
 - c. Problem three: if aptitude tests are conceived as measuring the amount of linguistic skill needed for language learning, supposing we are able to measure linguistic skill, surely this is not the only factor which accounts for successful performance.

Is it possible to separate linguistic skill (even if we think we are testing it!) from factors such as intelligence, motivation, the whole teaching situation? There are too many variables. Need for more research into what is fundamentally involved in LL. How do we separate APTITUDE from IQ? There is a correlation between HIGH IQ and SUCCESS IN LL.

7. Can one measure a student's progress in learning a L2? What are the difficulties?
- Defining the syllabus: Clear statement of aims & methods & specification of the content of learning. The design and content of the progress test will seek to show that students have attained those abilities the course seeks to develop. Difficulty in knowing whether to attribute progress to the course, previous knowledge, outside influences, time in host environment or classroom, etc. The quality of the test and the assumptions on which it is based.
8. How does one specify test objectives? How far does the way objectives are defined influence subsequent assessment?
- Definition of objectives: "The best basis for setting valid tests is to ask questions at every stage, but especially at the beginning of test development process, so that specification is as clear a statement as possible of why assessment is being made, what it will contain and the consequences for teaching, learning and administration.

- Linguistic Competence ("Levels"): Tests of grammar & usage (Morphology & Syntax), vocabulary (Lexis & Semantics: collocations), LC (Phonology: discrimination, recognition, pronunciation, stress & intonation) & writing skills (graphology)

BAB IV

KIND OF TESTS

The needs of assessing the outcome of learning have led to the development and elaboration of different test formats. Testing language has traditionally taken the form of testing knowledge about language, usually the testing of knowledge of vocabulary and grammar. Stern (1983, p. 340) notes that „if the ultimate objective of language teaching is effective language learning, then our main concern must be the learning outcome“. In the same line of thought, Wigglesworth (2008, p. 111) further adds that “In the assessment of languages, tasks are designed to measure learners’ productive language skills through performances which allow candidates to demonstrate the kinds of language skills that may be required in a real world context.” This is because a “specific purpose language test is one in which test content and methods are derived from an analysis of a specific purposes targetlanguage use situation, so that test tasks and content are authentically representative of tasks in the target situation” (Douglas, 2000, p. 19).

Thus, the issue of authenticity is central to the assessment of language for specific functions. This is another way of saying that testing is a socially situated activity although the social aspects have been relatively under-explored (Wigglesworth, 2008). Yet, language tests differ with respect to how they are designed, and what they are for, in other words, in respect to test method and test purpose. In terms of method, we can broadly distinguish traditional paper-and-pencil language tests from performance tests.

Paper-and-pencil language tests are typically used for the assessment either of separate components of language knowledge (grammar, vocabulary etc.), or of a receptive understanding (listening and reading comprehension). In performance-based tests, the language skills are assessed in an act of communication. Performance tests¹ are most commonly tests of speaking and writing, for instance, to ask a language learner to introduce himself or herself formally or informally and to write a composition, a paragraph or an essay, on the way he or she spent her summer holidays. These examples are elicited in the context of simulations of real-world tasks in realistic contexts. In terms of purpose, several types of language tests have devised to measure the learning outcomes accordingly. However, each test has its specific purpose, properties and criterion to be measured². The test types that will be dealt with in this part have been laid-out not in terms of importance, they are all of equal importance, but on the basis of alphabetical order. Yet, dictation, the traditional testing device

which focuses much more on discrete language items, will have its fair of attention in terms of its pro's and con's.

1. Achievement Test

An achievement test, also referred to as attainment or summative test, are devised to measure how much of a language someone has learned with reference to a particular course of study or programme of instruction, e.g. end-of-year tests designed to show mastery of a language. An achievement test might be a listening comprehension test based on a particular set of situational dialogues in a textbook. The test has a two-fold objective:

- 1) To help the teachers judge the success of their teaching.
- 2) To identify the weaknesses of their learners.

In more practical and pedagogical terms, Brown (1994, p. 259) defines an achievement test as „tests that are limited to particular material covered in a curriculum within a particular time frame“. In other words, they are designed primarily to measure individual progress rather than as a means of motivating or reinforcing language. Ideally, achievement tests are rarely constructed by classroom teacher for a particular class.

2. Cloze Test

A cloze test, also alternately referred to as cloze procedure, consists of a set of techniques for measuring, for example, reading comprehension. In a cloze test words are removed from a reading passage at regular intervals, leaving blanks. For example every fifth word may be removed. The reader must then read the passage and try to guess the missing words. For example, a cloze passage looks like this:

*A passage used in cloze test is a of written material in
..... words have been regularly The learners must then to
reconstruct the passage filling the missing*

(Adapted from Richards et al., 1989, p. 41)

Here, the test-taker or the reader has to guess the following missing words: a, passage, which, removed, try, by and words. The cloze test can also be used to judge the difficulty of reading materials. If the cloze procedure is being used for language testing, the test-taker is given a score according to how well the words guessed match the original words, or whether or not they make sense. Two types of scoring procedure are used:

- 1) The reader must guess the exact word which was used in the original (as in the example) above. This is called exact word method.
- 2) The reader can guess any word that is appropriate or acceptable in the context.
- 3) This is called the acceptable word method.

Another illustrative example of close test looks something like the following:

'A week has seven'. The only word which will fit in this blank is „days“. But sometimes one can choose between two or more words, as in: 'We write with a.....'. In this blank one can write „pen“ or „pencil“ or even „chalk“, „computer“ or „typewriter“.

However, two substantial criticisms have been made to the cloze-test types (Broughton et al., 1980). The first of these criticisms is that such tests rarely afford the person being tested any opportunity to produce language spontaneously. The second is that they are fundamentally trying to test that knowledge of the language system that underlies any actual instance of its use –linguistic competence in the Chomskyan sense- they are not concerned with the ability to master the language system for particular purposes with particular people in particular situations.

3. Diagnostic Test

As its name denotes, a diagnostic test is primarily designed to diagnose some particular linguistic aspects. Diagnostic tests in pronunciation, for example, might have the purpose of determining which particular phonological features of the English language are more likely to pose problems and difficulties for a group of learners. One of the well-known diagnostic tests in English is Prator's (1972) Diagnostic Passage. It consists of a short written passage that the learner reads orally; the teacher then examines a tape recording of that reading against a very detailed checklist of pronunciation errors. Basically, diagnostic language tests have a threefold objective:

- 1) To provide learners with a way to start learning with their own personal learning programme or what would be called in the literature of testing learning paths.
- 2) To provide learners with a way to test their knowledge of a language.
- 3) To provide learners with better information about their strengths and weaknesses.

Ideally, diagnostic tests are designed to assess students' linguistic knowledge (knowledge of and about the language) and language skills (listening, speaking, reading and writing) before a course is begun. However, the term formative is sometimes used to

designate a diagnostic test. One of the main advantages of a diagnostic test is that it offers useful pedagogical solutions for mixed-ability classes. In this very specific context, Broughton et al. (1980) contend that:

There will certainly be a large block in the middle of the ability range who can be separated off as a group for some parts of the lesson, or for some lessons, and will form a more homogenous teaching group. If this strategy is adopted, the poor ones and the better ones must receive their due time and attention.

(Broughton et al. 1980, p. 189)

4. Discrete-Point Test

The discrete-point test, also called discrete-item test, is a language test which measures knowledge of individual language items, such as a grammar test which has different sections on tenses, adverbs and prepositions. Discrete-point tests are based on the theory that language consists of different parts such as speech sounds, grammar and vocabulary, and different skills such as listening, speaking, reading and writing, and these are made up of elements that can be tested separately. Test consisting of multiple-choice questions are usually regarded as discrete-point tests. Discrete-point tests are all too often contrasted with what are called integrative tests. An integrative test is one which requires a learner to use several skills at the same time. An essay-writing is an integrative test because it leans heavily on the knowledge of grammar, vocabulary, and rules of discourse; a dictation is also an integrative test as it requires knowledge of grammar, vocabulary and listening comprehension skills.

In this vein, Harmer notes the following distinction between discrete-point testing and integrative testing, “Whereas discrete point-testing only tests on thing at a time such as asking students to choose the correct tense of a verb, integrative test items expect students to use a variety of language at any one given time – as they will have to do when writing a composition or doing a conversational oral test” (Harmer, 2001, p. 323). In the same line of thought and Broughton et al. ,more than some thirty years ago, noted that “Since language is seen as a number of systems, there will be items to test knowledge of both the production and reception of the sound segment system, of the stress system, the intonation system, and morphemic system, the grammatical system, the lexical system and so on” (Broughton et al., 1980, pp. 149-150).

5. Language Aptitude Test

Before one ventures into defining what a language aptitude test is, it would be wiser to start first by defining what a language aptitude is. Language aptitude, as a hybrid concept

part linguistic and part psychological, refers to the genuine ability one is endowed with to learn a language. It is thought to be a combination of several abilities:

- Phonological ability, i.e. the ability to detect phonetic differences (e.g. of stress, intonation, vowel quality) in a new language.
- Syntactic ability, i.e., the ability to recognize the different grammatical functions of words in sentences.
- Psychological ability, i.e. rote-learning abilities and the ability to make inferences and inductive learning.

Additionally, Crystal (1989, p. 371) suggests other variables conducive to successful language learning such as „empathy and adaptability, assertiveness and independence with good drive and powers of application“. A high language-aptitude person can learn more quickly and easily than a low language-aptitude individual. The evidence in such assertion is axiomatic in a language aptitude test.

A language aptitude test tends to measure a learner aptitude for language learning, be it second or foreign, i.e. students performance in a language. Thus, it is used to identify those learners who are most likely to succeed. Language aptitude tests usually consist of several different test items which measures such abilities as:

- Sound-coding ability, i.e. the ability to identify and remember new sounds in a new language.
- Grammar-coding ability, i.e. the ability to identify the grammatical functions of different parts of sentences.
- Inductive-learning ability, i.e. the ability to work out meanings without explanation in the new language.
- Memorization, i.e. the ability to remember and to recall words, patterns, rules in the new language.

Two well-known standardized language aptitude tests have been used in the United States, the Modern Language Aptitude Test (Carroll and Sapon, 1958) and the Primsleur Language Aptitude Battery (Primsleur, 1966). Both of these are English tests and require students to perform such tasks as learning numbers, listening, detecting spelling clues and grammatical patterns and memorizing (Brown, 1994).

6. Placement Test

A placement test, as its name implies, is originally designed to place learners at an appropriate level in a programme or course. The term “placement test” as Richards et al. (1989) note does not refer to what a test contains or how it is constructed, but to the purpose for which it used. Various types or testing procedures such as dictation, interview or a grammar test (discrete or integrative) can be used for placement purposes. The English Placement test (EPT), which is a well-known test in America, is an illustrative example of this test-type. The EPT is designed to assess the level of reading and writing skills of entering undergraduate students so that they can be placed in appropriate courses. Those undergraduate students who do not demonstrate college or university-level skills will be directed to remedial courses or programmes to help them attain these skills.

7. Proficiency Test

A proficiency test is devised to measure how much of a language someone has learned. It is not linked to any particular course of instruction, but measures the learner’s general level of language mastery. Most English language proficiency tests base their testing items on high frequency-count vocabulary and general basic grammar. Some proficiency tests have been standardized for worldwide use, such as the well-known American tests, the TOEFL, and the English Language Proficiency Test (ELPT)³ which are used to measure the English language proficiency of foreign students intending further study at English-speaking institutions, namely the USA. However, the Cambridge Certificate of Proficiency in English or CPE, as it is generally referred to, is the most advanced remains the only British top-value and highprestige standardized⁴ language test. It is the most advanced general English exam provided by the University of Cambridge. The Certificate is recognized by universities and employees throughout the world. The English level of those who pass the CPE is supposed to similar to that of a fairly educated native speaker of English. Clearly, as Valette posits, „the aim of a proficiency test is to determine whether this language ability corresponds to specific language requirements“ (Valette, 1977, p. 6).

Actually, there are four other types of Cambridge proficiency tests, the Cambridge Key English Test (KET), the Cambridge Preliminary English Test (PET), The Cambridge First Certificate of English (FCE) and the Cambridge Certificate in Advanced English (CAE). The material contained in proficiency tests can be used for teaching as well as for testing. In essence, a proficiency test measures what the student has learned in relation

to a specific purpose, e.g. does the student know enough English to follow a course offered in English?

8. Progress Test

A progress test is an achievement-like test. It is closely related to a particular set of teaching materials or a particular course of instruction. Progress tests are usually administered at the end of a unit, a course, or term. A progress test may be viewed as similar to an achievement test but much narrower and much more specific in scope (Richards et al., 1989). They help examiners in general and language teachers in particular to assess the degree of success of their programmes and teaching and therefore to identify their shortcomings and weaknesses respectively. Progress tests can also be diagnostic to some degree, in the sense that they help identify areas of difficulties encountered by learners in general.

9. TOEFL

The Test of English as a Foreign Language, or TOEFL for short, is a large-scale language assessment. It is, “arguably the most well-known and widely used large-scale language assessment in the world” (Kunnan, 2008, p. 140). It was first developed in 1963 in the United States to help in the assessment of the language competence of non-native speakers. As a test type, it is a standardized test of English proficiency administered by the Educational Testing Service, Princeton. It is widely used to measure the English-language proficiency of foreign students wishing to enter American colleges and universities. According to Taylor and Angelis (cited in Kunnan, 2008) the first TOEFL was administered in 1964 at 57 test centres to 920 test candidates. Recently, the TOEFL has widely been recognized as a model test and have-take-test for our students, graduate and postgraduate, as well as our teachers and researchers in universities and higher education institutions wishing to read for higher degrees and develop further their research potential in North American universities⁵. Kunnan (2008, p. 141) notes that, “Over the years, the TOEFL became mandatory for non-American and non-Canadian native speakers of English applicants to undergraduate and graduate programs in U.S. and Canadian English-medium universities”.

One of the most important realizations in the TOEFL enterprise was the launching of a more innovative test, the iBTTOEFL, internet-based TOEFL, in 2005. This iBTTOEFL is regarded as a significant development over the previous TOEFL forms and the TOEFL CBT, Computer-Based Test, launched in 1996. The novel features of the iBTTOEFL are a speaking section consisting of independent and integrated skills tasks, a listening section

with longer lectures and conversations with note-taking, a reading section made up of questions that ask test-takers to categorize information and fill in a chart or complete a summary and a writing section that has both an independent and integrated task.

BAB V

ASSESSING VOCABULARY

A. Definition of Vocabulary

There are many ways of assessing vocabulary. Most often, people equate vocabulary tests with test formats that involve measuring test takers' ability to link the forms of words with their meanings. Examples of this type of lexical assessment are form recall tests in which test takers are provided with the meanings of words (e.g., land with water all around it _____) and have to write their written forms (e.g., island) and form recognition tests in which the meaning of a word is provided (e.g., land with water all around it) and testtakers must select the correct word from among several different options (e.g., a) game b) island c) mouth d) movie e) song f) yard). However, because vocabulary knowledge is multidimensional, assessing the ability to link the written forms of words with their meanings is not the only way to measure vocabulary knowledge. We can also evaluate whether students can correctly spell words, pronounce words, and recognize or produce their derivations, associations, secondary meanings, and collocations. The ability to successfully answer one type of lexical assessment (e.g., spelling) does not ensure that another type of vocabulary test will also be correct (e.g., collocation). This means that it is very important that teachers, learners, and researchers understand the different ways that we can assess vocabulary, and the degree to which each test format reveals vocabulary learning.

B. Vocabulary Level Tests

The Vocabulary Levels Test (VLT; Nation, 1983, Schmitt et al., 2001; Webb et al., 2017) is a matching test that assesses learners' receptive vocabulary knowledge of different word frequency levels. It is perhaps the best known test of L2 vocabulary knowledge (Read, 2000) because of its great practical value to teachers and learners. The VLT was initially developed by Nation (1983) and then revised by Schmitt et al. (2001) and Webb et al. (2017) to reveal students' lexical knowledge and indicate which words they should learn. The rationale behind assessing knowledge of word frequency levels is that the value of a word is indicated by its word frequency. For example, Webb (2021) shows that the most frequent 1000, 2000, and 3000 word families account for approximately 75% - 91%, 4% - 11%, and 2% - 4%, respectively, across a range of different discourse types. The large difference in frequency of occurrence of words at different frequency levels reveals the importance of learning words according to their frequencies.

In the most recent version of the VLT (Webb et al., 2017), knowledge of each of the 1000, 2000, 3000, 4000, and 5000 most frequent word families is assessed. Thirty items are used to measure knowledge of each of the five levels. The items in a level are presented in 10 sets of 3. In each set of items, the meanings of three target words are presented together with six words from that frequency level. Test takers are required to match the words with their meanings. The following example is one set from the 1000-word frequency level.

	choice	computer	garden	photograph	price	week
cost						
picture						
place where things grow outside						

The VLT score at each word level needs to be carefully interpreted, rather than aggregating scores across word levels as a total score. This is because the VLT was designed to indicate which word frequency levels students should be learning. For example, if VLT scores indicate that learners have mastered the 1000-word frequency level, but no levels beyond the 1000 word-level, then students should focus on learning the 2000 level. If learners have mastered the 1000 and 2000 word levels but not the 3000 level, then the focus of learning should be the 3000 level, and so on. The cut-off score for mastery of each word level varies. Webb et al. (2017) recommended scores of 29/30 or higher as the cutting point for the 1000, 2000, and 3,000 levels and 24/30 or higher for the 4,000 and 5,000 levels. The rationale for having different cut-off scores for different levels is that because the highest frequency levels account for the greatest proportion of English, there is a need to ensure that learners have a higher degree of knowledge of these levels. For lower word frequency levels, a less demanding cut-off scores may be appropriate because there is less variation in the word frequencies across these levels.

Laufer and Nation's (1999) Productive Vocabulary Levels Test (PVLТ) was developed based on the original version of the VLT (Nation, 1983) to assess learners' productive knowledge of different word levels. In the PVLТ (Laufer & Nation, 1999), knowledge of the 2,000, 3,000, 5,000, 10,000, and university word list is measured. The PVLТ uses a gapped word completion task, requiring learners to fill in the missing word. The following example for the target item opportunity was taken from the 2000 level of the PVLТ.

1. I'm glad we had this opp_____ to talk.

Both the receptive and productive VLT indicate which word level learners have mastered, enabling researchers to identify areas for further investigation and assisting teachers in targeting specific vocabulary levels in their classes. The strength of the VLT is that it provides a diagnostic tool to determine the vocabulary that students should learn, as well as a useful indication of the lexical development of students. However, it is important to be aware that receptive vocabulary knowledge will be greater than productive vocabulary knowledge (e.g., Webb, 2008). Although a learner may demonstrate mastery of the most frequent 2,000 word families on Webb et al.'s (2017) VLT, that does not mean that the learner would master the 2000 word level of Laufer and Nation's (1999) PVLT. Therefore, measuring both receptive and productive knowledge may be useful to provide a more accurate assessment of the vocabulary knowledge of learners.

C. Vocabulary Size Test

The Vocabulary Size Test (VST; Nation & Beglar, 2007) is a multiple-choice meaning recognition test that assesses the written receptive vocabulary breadth of learners. A target word included in a short, non-defining sentence, and four alternative definitions of the word are presented, and test takers (learners) are asked to choose the best meaning for each target word. The following example was taken from the VST.

Circle the letter a-d with the closest meaning to the key word in the question.

1. SEE: They saw it.

a. cut b. waited for c. looked at d. Started

The VST is made up of 140 items and was designed to measure knowledge of the 14,000 most frequent word families found in the British National Corpus. The vocabulary size of a learner is calculated by multiplying the total score from the VST by 100 which is the number of words each item represents (Nation & Beglar, 2007). The strength of the VST is that the test covers knowledge all of the word families that beginner, intermediate and advanced L2 learners are most likely to have encountered. This makes it a particularly useful test to differentiate between the lexical knowledge of advanced learners.

D. Word Associates Test

The Word Associates Test (WAT; Read, 2000, pp. 180-186) was designed to measure knowledge of two aspects of vocabulary knowledge. In the WAT eight options are presented with a target word, and test-takers are asked to select four words that either have related

meanings (i.e., semantic associates) or may co-occur together in a phrase (i.e., collocates). The following example is for the target word sudden.

Sudden

beautiful	quick	surprising	thirsty	change	doctor	noise	school
-----------	-------	------------	---------	--------	--------	-------	--------

In the first box, quick and surprising are associates, sharing a semantic relationship with sudden, while in the second box change and noise are collocates, sharing a syntactic relationship with sudden. By assessing two types of vocabulary knowledge, the WAT measures a deeper knowledge of target words than breadth tests such as the VLT and VST.

E. Word Part Levels Test

The Word Part Levels Test (WPLT; Sasao & Webb, 2017) is a valid and reliable measure of three aspects (form, meaning, and use) of affix knowledge. Across three difficulty levels (beginner, intermediate, advanced), the WPLT assesses knowledge of 118 affixes. Items which evaluate knowledge of affix form involve choosing an affix form among four options (an affix form and three distractors). Items which evaluate knowledge of affix meaning involve selecting the correct meaning from four options when presented with the target affix (e.g., re-) and two example words with the affix underlined (e.g., replay; rebuild). Items which evaluate knowledge of affix use involve selecting the part of speech of words that include a target affix item (e.g., en-). Two example words with the affix underlined (e.g., ensure; enable) are provide along with four part of speech options (noun, verb, adjective, adverb). In the following items taken from the WPLT, item 1 assesses affix form, item 2 assesses affix meaning, and item 3 assesses affix use. Because the WPLT measures three aspects of affix knowledge and has three difficulty levels, it is a useful tool for teachers to use to determine their students' knowledge of affixes and select affixes to teach.

1. (1) sal- (2) cau- (3) lin- (4) dis
2. mono- (monotone; monoculture)
 - (1) person/thing
 - (2) not
 - (3) one
 - (4) into another state/place
3. -ment (development; management)
 - (1) Noun

(2) Verb

(3) Adjective

(4) Adverb

BAB VI ASSESSING GRAMMAR

In term of assessment, grammar is central to language description and test taker performance. Communicative competence has four components: grammatical, socio-olinguistic, discourse, and strategic competence. Grammatical knowled ge consist of three elements:

1. GRAMMATICAL FORMS OR THE STRUCTURES OF A LANGUAGE

Form is both morphology or how words are formed, and syntax, how words are strung together, both of them are concerned with the linguistic accuracy of language

2. THE GRAMMATICAL MEANING OF T HOSE FORMS

Grammatical meaning consists of both the literal and intended message that is conveyed by the form. It is concerned with the meaningful ess of the language used

3. THEIR PRAGMATIC MEANING OR USE IN A GIVEN CONTEXT.

The pragmatic or implied meaning results from the appropriate language choices a learner makes in a given communicative event.

Table 5.6 Examples of task types

Selected-response tasks	Limited-production tasks	Extended-production tasks
<ul style="list-style-type: none"> • Multiple-choice activities • True/false activities • Matching activities • Discrimination activities • Lexical list activities • Grammaticality judgment activities • Noticing activities 	<ul style="list-style-type: none"> • Gap-filling activities • Cloze activities • Short-answer activities • Dictation activities • Information-transfer activities • Some information-gap activities • Dialogue (or discourse) 	<ul style="list-style-type: none"> • Summaries, essays • Dialogues, interviews • Role-plays, simulations • Stories, reports • Some information-gap activities • Problem-solving activities • Decision-making

A. Designing Assessment Tasks: Selected Response

1. Multiple-choice tasks

The most common selected responses task presents a blank or underlined wor ds and a sentence and the test taker must choose the orrect respons es from the options that are ven. (Grammatical form, Grammatical meaning, Grammatical for and meaning)

2. Discrimination tasks

The tasks that asks the test-takers to attend to input that can be either language or non-language and to respond in the form of a choice between or among contrasts or opposites, such as true/false. (grammatical form and meaning)

3. Discrimination tasks

Test-takers are asked to indicate (underline circle) that they have identified a specific feature in the language sample. (grammatical meaning)

B. Designing Assessment Tasks: Limited Production

1. Gap-filling tasks

The language is presented in the form of sentence, dialogue, or passage in which a number of words are deleted

2. Short-answer tasks

The input is presented in the form of a question following reading passage or oral/visual stimulus.

3. Dialogue-completion tasks

The input is presented in the form of a short conversation or dialogue in which a part of the exchange or the entire exchange is left blank and the expected response meant to be grammatically correct.

C. Designing Assessment Tasks: Extended Production

1. Information Gap Tasks

2. Presents the input in terms of incomplete information.

3. Role-play or Simulation Tasks

The input presents test-takers with a language or non-language prompt that asks them to take on a role or stimulate a situation to solve a problem, make decision, or perform some transaction collaboratively.

BAB VII

ASSESSING PRONUNCIATION

Assessment is an important part of learning. Without assessment, development of one's abilities or learning outcomes will not be well known. To find out learning outcomes, tools are needed, instrument, or a quality assessment rubric. Quality of tools, instruments, or the assessment rubric has a positive impact on the quality of the assessment results. Associated with assessment, in particular the assessment of speaking or pronunciation skills, the quality of tools, instruments, or the assessment rubric determines the quality of the ability assessment results speaking or pronunciation.

A. What Is Pronunciation?

Pronunciation is the ability to produce speech orally and systematically to express a certain purpose. A person speaking or pronunciation has purpose, namely to convey something to others. Therefore, In order to be skilled, speaking is done in a systematic, coherent, and patterned manner. Speaking is a person's ability to express ideas, thoughts, feelings, opinions to others verbally.

B. Important Aspects in Assessing Speaking Skills Include

1. fluency,
2. accuracy, and
3. oral communicative strategies.
4. The speaking accuracy in question includes (a) grammar (grammar), (b) vocabulary (vocabulary), and (c) pronunciation (pronunciation).

C. Why do we need to learn skills for pronunciation?

With pronunciation skills we can convey various kinds of information (facts, events, ideas, ideas, responses, and so on). We can express our wills and desires, and express various kinds of feelings.

D. What to Assess?

1. Vowels
2. Consonants
3. Diphthongs
4. Triphthongs
5. Word stress
6. Syllabic

7. American T
8. S/es
9. D/ed
10. Intonation

BAB VIII

ASSESSING LISTENING

Assessing listening involves evaluating a person's ability to understand spoken language. This can be done through various tasks that assess different aspects of listening, from basic sound discrimination to comprehending complex spoken discourse. Effective listening assessments often combine multiple methods and task types to capture a holistic view of a learner's listening proficiency.

A. The Importance of Listening

Listening plays an important role in developing language skills because listening is the main factor in communicating, because we listen first and then the results of what we listen to are what we will reveal. Therefore, it can be said that a good speaker is also a good listener.

B. Basic types of listening

1. Intensive

Testing the smallest parts of language, such as: phonemes, words, intonation, discourse markers, etc.

2. Responsive

Responds exactly to what has been heard. In this case, the test-taker is expected to be able to master the "WH question" in order to be able to answer the question correctly.

3. Selective

Test takers are able to choose the most appropriate answer based on the specific information that has been heard. In selective listening, students are usually asked to listen to names, numbers, directions, or certain facts and events.

4. Extensive

In extensive listening, a broader and more general understanding is needed. Usually includes the essence or main idea of what is heard.

C. Micro Skill of Listening

Distinguish between specific sounds.

Masterpieces of language that differ in length or in short

Introduce the forms of emphasis in English, the word position is stressed or not stressed, structure, rhythmic, intonation in its role in marking information.

Introduces the reduction of word forms



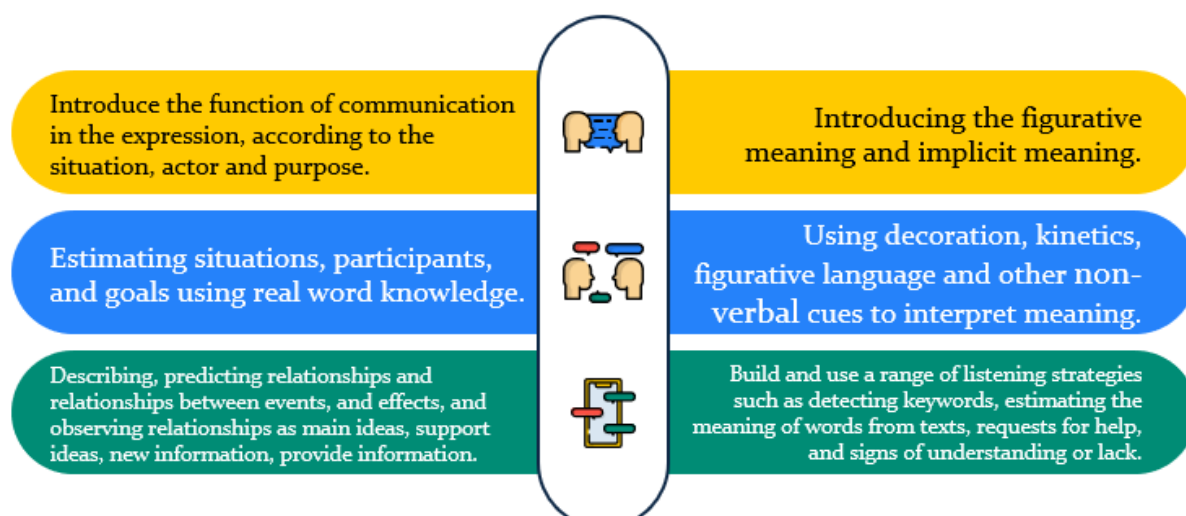
Distinguish foreign words, introduce core words.

Average speaking process at different delivery. (g) the speaking process contains breaks, errors, corrections from other delivery variables.

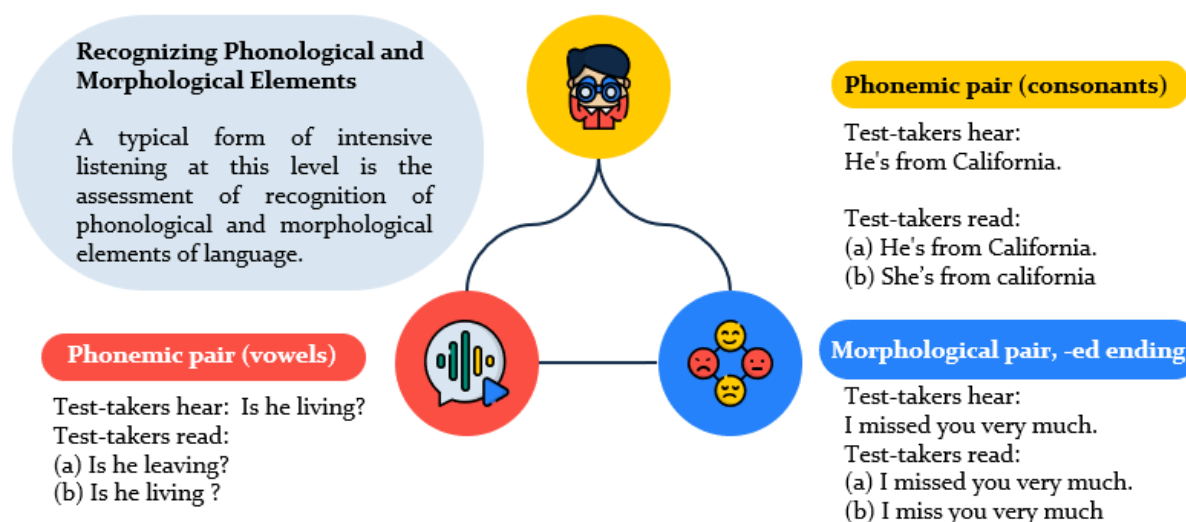
Introduce groups of words grammatically (nouns, verbs, and so on).

Introduce cohesive sentences in speaking discourse

D. Macro Skill of Listening



E. Designing Assessment Tasks : Intensive Listening

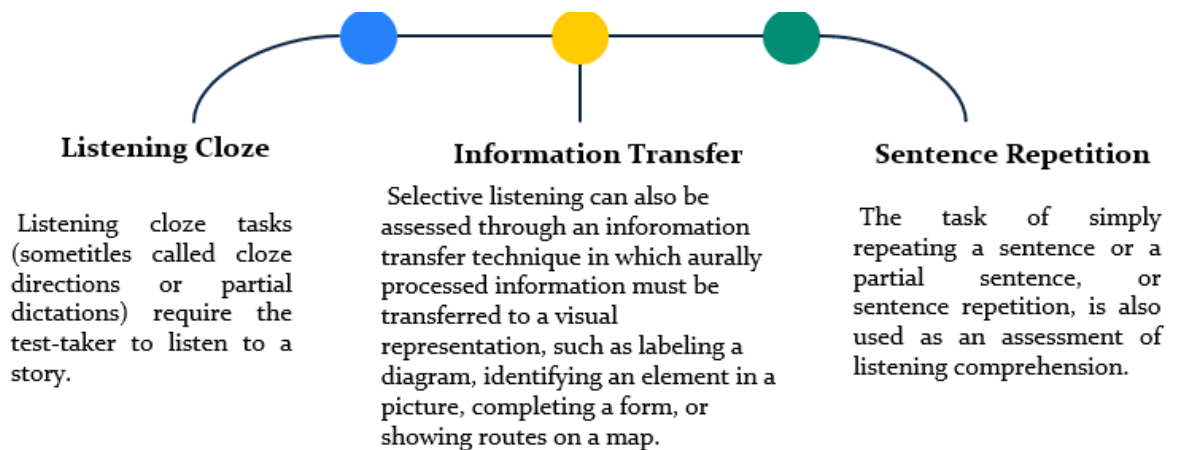


F. Designing Assessment Tasks : Responsive Listening

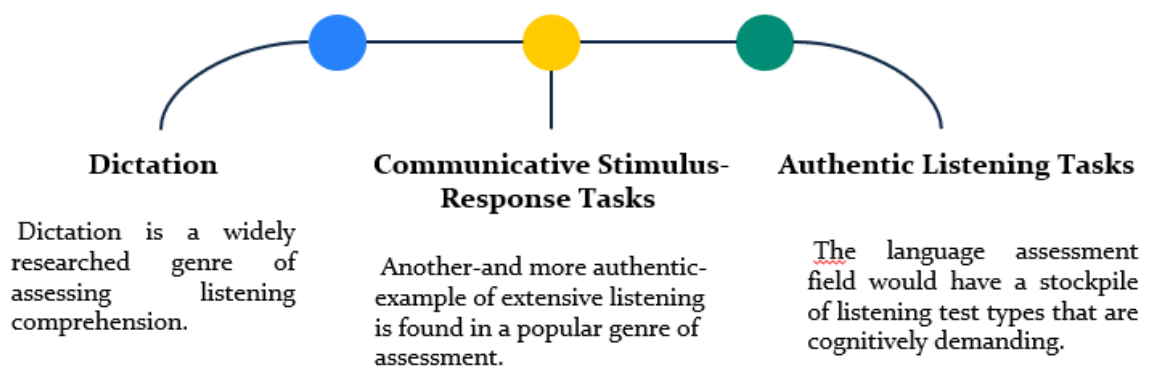
A questions and answer format can provide some interactivity in these lower end listening tasks. Responsive listening involves listening to a small amount of a language such a command, questions, or greeting. After listening, the listener is expected to develop an appropriate short response.

G. Designing Assessment Tasks : Selective Listening

A third type of listening performance is selective listening, in-which the test-taker listens to a limited quantity of aural input and must discern within it some specific information.



H. Designing Assessment Tasks : Extensive Listening



BAB IX

ASSESSING SPEAKING

From a pragmatic view of language performance, listening and speaking are almost always closely interrelated. Speaking is a productive skill that can be directly and empirically observed, those observations are invariably colored by the accuracy of effectiveness of a test-taker's listening skill, which necessarily compromises the reliability and validity of an oral production test. The interaction of speaking and listening challenges the designer of an oral production test to tease apart, as much as possible, the factor accounted aural intake.

Another challenge is the design of elicitation techniques. Because speaking is product of creative construction of linguistic strings, the speaker make choices of lexicon, structure, and discourses. If your goal is to have test-taker's demonstrate certain spoken grammatical categories. In receptive performance, the elicitation stimulus can be structured to anticipate predetermined responses and only those responses. In productive performance, the oral or written stimulus must be specific enough to elicit output within an excepted range performance such that scoring or rating procedures apply appropriately. For example, in a pictures-series task, the objective of which is to elicit output within a story in a sequence of events, test-takers could opt for a variety of plausible ways to tell the story, all of which might be equally accurate.

A. Imitative Speaking

At one end of a continuum of types of speaking performance is the ability to simple parrot back imitate a word or phrase or possibly a sentence. This is a purely phonetic of level of oral production, a number of prosodic, lexical, and grammatically properties of language may be included in the criterion performance. Test takers are assessed if they have the ability to imitate a word while pronouncing. However conveying the meaning is not the purpose. The role of listening here is the short term storage.

Phonepass test

A popular test that use imitative production task is phonepass, a widely used, commercially available speaking test in many countries. Among a number speaking task on the test, repetition of sentences occupies a prominent role. It is remarkable that research on the phonepass test has supported the construct validity of its repetition tasks not just for test-takers phonological ability but also for discourse and overall oral production ability (Townshend et al., 1998; Bernstein et al., 2000; Cascallar & Bernstein, 2000)., The phonepass findings could signal an increase in the future use of repetition and read aloud procedures for the assessment of oral

production. Because a test-takers output is completely controlled, scoring using speech recognition technology become achievable and practical.

B. Intensive Speaking

A second type of speaking frequently employed in assessment context is the production of short stretches of oral language designed to demonstrate competence in a narrow band of grammatical, phrasal, lexical, or phonological relationship. The intensive level, test-takers are prompted to produce short stretches of discourse (no more than one sentence) through which they demonstrate linguistic ability at a specified level of language.

Direct Response Tasks

The test administrator elicits a particular grammatical form or a transformation of sentence. Such tasks are clearly mechanical and not communicative, but they do require minimal processing of meaning in order to produce the correct grammatical output.

Read Aloud Tasks

Intensive reading-aloud tasks include reading beyond the sentence level up to a paragraph or two. This technique is easily administered by selecting passage that incorporate test specs and by recording test-takers output; the scoring relatively easy because all of test-takers oral production is controlled. In this task teacher, teacher listening to the recording would then rate student on a number of phonological factors (vowels, diphthongs, consonants, stress, and intonation) by completing diagnostic checklist on which all error or questionable item were noted. Some variations on the task of simply reading a short passage : reading a scripted dialogue. Reading sentences containing minimal pairs, and reading information from a table or chart.

Sentence / Dialogue Completion Tasks And Oral Questioners

Test-takers are first given time to read through the dialogue to get its gist and to think about appropriate lines to fill in. Then as the tape, teacher or test administrator produces one part orally, the test-takers respond. The advantage of this technique are its reliance on literacy and ability to transfer from written to spoken English.

Picture cued tasks

One of the more popular ways to elicit oral language performance at both intensive and extensive levels is a picture-cued stimulus that requires a description from the test-takers. Pictures may be very simple, designed to elicit a word or a phrase; somewhat more elaborated and “busy” ; or composed of a series that tells story or incident. Evaluating interview could be

used : Grammar vocabulary, comprehensible, fluency, pronunciation, and task (accomplishing the objective of the elicited task)

Translation

Translation is a part of our tradition in language teaching that we tend to discount or disdain, if only because of our current pedagogical stance plays down its importance. Translation method of teaching are certainly passé in an era of direct approaches to creating communicative classrooms. We should remember that in where English is not the native or prevailing language, translation is a meaningful communicative device in context where the English user is called on to be an interpreter

C. Responsive Speaking

Responsive assessment tasks include interaction and test comprehension but at the somewhat limited level of very short conversations, standard greeting and small talk, simple requests and the like. It helps English language instructors assess students' ability to engage in discussion with one or more interlocutors. More creativity on the students' part is required.

Question and Answer

Question and answer task can consist of one or two questions from an interviewer, or they can make up a portion of a whole battery of questions and prompts in an oral interview. This form of assessment requires one person, an interlocutor, to ask the student questions. Students' answers can be 1 sentence responses, or, can entail a more detailed answer depending upon the question asked.

Giving Instructions and Directions

This type of assessment entails a student asking the teacher questions. Essentially, the student assumes control over the conversation. This type of assessment requires students to be able to give directions given a particular situation. Topics should be relevant and of immediate interest to students.

Paraphrasing

This type of assessment requires students to listen to a 4 to 6 sentence paragraph. Upon finishing the paragraph, instructors prompt students to give a 1 to 2 sentence summary of what they just heard. The advantage is the elicit short stretches of output and perhaps tap into test-takers to practice the conversational art of conciseness by reducing the output/ input ratio.

D. Interactive Speaking

Interactive speaking is extremely similar to responsive speaking, however varying in that an interactive speaking session is much more involved, may include multiple interlocutors, and is commonly found in the English speaking world. Interactive speaking assessments are important because they allow instructors to evaluate students' ability in producing fluid, detailed, and in-depth discussions with one or more interlocutors.

Interview

Oral production assessment' is mentioned the first things that comes to mind is oral 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. Interviews provide students with chances to use what they have learned in an authentic situation; they give students the chance to engage in meaningful dialogue.

Role Play

Role playing is a popular pedagogical activity in communicative language-teaching classes. Within constraints set forth by guidelines, it frees students to be somewhat creative in their linguistic output. Role play opens some windows of opportunity for test-takers to use discourse that might otherwise be difficult to elicit. Role playing provides students with a chance to have fun with the English language. Moreover, it provides students with mock situations that mimic real-world situations.

Discussion and Conversation

and even more difficult to score. Assessing the performance of participants through scores or checklist course, discussion is an integrative task, and so it is also advisable to give some cognizance to comprehension performance in evaluating learners. Discussion may be especially appropriate task through which elicit and observe such abilities:

- Topic nomination, maintenance, and termination
- Attention getting, interrupting, and paraphrasing
- Comprehension signals
- Intonation patterns for pragmatic effect

Games

Games are usually the easiest way to get students engaged in learning English. One type of interactive game assessment is information gap grids. Information Gap Grids In this game, students interview each other, in English, to determine the answers to various questions.

E. Extensive Speaking

Extensive speaking tasks involve complex, relatively lengthy stretches of discourse. They are frequently variations on monologues, usually with minimal verbal interaction. Extensive speaking is extremely important; students are left by themselves to produce clear and intelligible speech. There usually is some type of audience present; there is no dialogue between presenter and audience members. Only the person presenting speaks during this time.

Oral Presentations

Oral presentations would not be uncommon to be called on to present a report, a paper, a marketing plan, a sales idea, a design of new product, or method in academic or professional arenas. Oral presentations give students a chance to use what they have learned in English class by culminating everything into one strong and concise presentation.

Picture Cued Story telling

Picture cued story telling is one of the most techniques for eliciting oral production is through visual picture, photographs, diagrams, and charts. The purpose of picture-cued story telling is to provide students with examples of how chronology is used in discussions. Also, it is used to illustrate situations.

Retelling Story

This is type of task in which test-takers hear or read a story or news event that they are asked to retell. In this form of extensive speaking assessment, students re-tell a story they heard. The story can be modified from its original form. We have already looked at this elicitation device for intensive tasks, but at this level we consider a picture or a series of picture as a stimulus for a longer story or description. If we are eliciting specific grammatical or discourse features, you might add to the directions like “tell the picture that these pictures describe. Using the past tense of verb” .

Translation

Translation of words, phrases, or short sentences was mentioned under the category of intensive speaking. The longer texts are presented for the test-takers to read in native language and translate into English. Those texts could come in many forms such as dialogue, directions for assembly product, a synopsis of a story, etc. The advantage of this task are it may control the vocabulary, grammatical and discourse features.

Picture-Cued Tasks

One of the more popular ways to elicit oral language performance at both intensive and extensive levels is a picture-cued stimulus that requires a description from the test-takers. Pictures may be very simple, designed to elicit a word or a phrase. Picture-Cued Tasks A

picture-cued stimulus requires a description from the test-taker. It may elicit a word, a phrase, a story, or incident. This form of assessment entails an interviewer asking students what is in the picture or what is happening in the picture. Picture-cued test will make test-takers open their imagination about a picture or things, improve their vocabulary, and test-takers can describe about a things by using oral language performance.

BAB X

ASSESSING READING

Reading is the most essential skill for success in all educational context, remains a skill of paramount importance as we create assessment of general language ability. Assessing reading involves evaluating a learner's ability to understand, interpret, and respond to written texts using various skills. It goes beyond simply checking if a student can pronounce or recognize words—it focuses on how well they comprehend and interact with a text. Here's a breakdown of what “assessing reading in detail” usually includes:

1. Genre of Reading



2. Micro Skills for Reading

- 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), system (tense, agreement, pluralization), patterns, rules, and elliptical forms.
- 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.

3. Macro Skills for Reading

- 1) Recognize the rhetorical forms of written discourse and their significance for interpretation

- 2) Recognize the communicative functions of written texts, according to form and purpose.
- 3) Infer context that is not explicit by using background knowledge.
- 4) From described events, idea, etc., 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.
- 5) Detect culturally specific references and interpret them in a context of the appropriate cultural schemata.
- 6) Develop and use a battery of reading strategies such as scanning and skimming, detecting discourse markers.

4. Designing assessment task: Perspective Reading

At the beginning level of reading a second language lies a set of tasks that are fundamental and basic: recognition of alphabetic symbols, capitalized and lowercase letters, punctuations, word and grapheme-phoneme correspondences

1) Literacy Tasks

Implying that learner is in the early stages of becoming “literate”.

2) Reading Aloud

- The test taker sees separate letters, words and/or short sentences and reads them aloud, one by one, in the presence of an administrator.
- Any recognizable oral approximation of the target response is considered correct.

3) Written Response

- The same stimuli is presented, and the test taker’s task is to reproduce the probe in writing.
- Evaluation of the test taker’s response must be carefully treated.

4) Multiple Choice: Grapheme Recognition Task

Test taker’s read: Circle the “odd” item, the one that doesn’t “belong”.

Piece	Peace	Piece
Book	Book	Boot

5) Minimal Pair





A minimal pair is two words that vary by only a single sound, usually meaning sounds that may confuse English learners.

Test takers read: Circle “S” for same or “D” for different.

- Led Let S D
- Bit Bit S D
- Seat Set S D
- Too To S D

6) Picture-Cued Items

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

<p>A.</p> 	<p>B.</p> 	<p>Test Takers read:</p> <p>1. washing the dishes _____</p> <p>2. chatting with a friend _____</p> <p>3. studying the lesson _____</p> <p>4. washing the clothes _____</p>
<p>C.</p> 	<p>D.</p> 	

5. Designing Assessment Task: Selective Reading

- Focus on formal aspects of language (lexical, grammatical, and a few discourse features).
- It includes what many incorrectly think of as testing “vocabulary and grammar”.

1. Multiple Choice: Vocabulary/grammar tasks

1. He's not married. He's _____.

- A. young
- B. single
- C. first
- D. a husband

2. If there's no doorbell, please _____ on the door.

- A. kneel
- B. type
- C. knock
- D. shout

2. Contextualized Multiple-choice Vocabulary/grammar tasks

1. Oscar: Do you like champagne?
Lucy: No. I can't _____.
A. stand
B. prefer
C. hate
2. Manager: Do you like to work by yourself?
Employee: Yes, I like to work _____.
A. independently
B. definitely
C. impatiently

3. Multiple-choice Cloze Vocabulary/grammar tasks

I've lived in the United States (21) _____ three years. I (22) _____ live in Costa Rica. I (23) _____ speak any English. I used to (24) _____ homesick, but now I enjoy (25) _____ here. I never (26) _____ back home (27) _____ I came to the United States, but I might (28) _____ to visit my family soon.

21. A. since 23. A. couldn't 25. A. live 27. A. when
B. for B. could B. to live B. while
C. during C. can C. living C. since

22. A. used to 24. A. been 26. A. be 28. A. go
B. use to B. be B. been B. will go
C. was C. being C. was C. going

4. Matching Task; Vocabulary Matching Task

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 |

5. Selected Response Fill-in Vocabulary

1. At the end of the long race, the runners were totally _____.
2. My parents were _____ with my bad performance on the final exam.
3. Everyone in the office was _____ about the new salary raises.
4. The _____ listening of the counselor made Christina feel well understood.

Choose from among the following:

disappointed
empathetic
exhausted
enthusiastic

6. Multiple-choice Grammar Editing Task

Choose the letter of the underlined word that is not correct.

1. The abrasively action of the wind wears away softer layers of rock.
A B C D
2. There are two way of making a gas condense: cooling it or putting it under pressure.
A B C D
3. Researchers have discovered that the application of bright light can sometimes be uses to overcome jet lag.
A B C D

7. Picture-cued Tasks

Carlo has a bar of chocolate. He gives half a bar of chocolate to his brother. See the following four pictures. Choose the picture that shows the relative amount of chocolate left to Carlo.



8. Sentence Completion

Oscar: Doctor, what should I do if I get sick?

Doctor: It is best to stay home and _____.

If you have a fever, _____.

You should drink as much _____.

The worst thing you can do is _____.

You should also _____.

GRADE ELEMENT	EXCELLENT (44 – 50)	GOOD (40 – 44)	ADEQUATE (30 – 39)	POOR (20 – 29)	VERY POOR (BELOW 20)
COMPREHENSION	Student answered more than 80% of comprehension questions correctly. Student fully demonstrates comprehension of main idea from the reading.	Student answered at least 75% of comprehension questions correctly. Student comprehends main idea of the reading.	Student answered at least half of the comprehension questions correctly.	Student answered at least 35% of the comprehension questions correctly. Students have difficulty understanding the main idea of the reading.	Student did not complete assignment and/or did not answer more than half of questions correctly.
FORM	Student answers in complete sentences and includes variety of sentence structure.	Student answers most questions in complete sentences and includes variety of sentence structure.	Student answers half of the questions in complete sentences and includes an occasional compound sentence, but most of the sentences are simple sentences.	Student answers half of the questions in complete sentences and includes simple sentences only.	Student did not answer any of the comprehension question in complete sentences.
	Has few or no errors in grammar, spelling, capitalisation or punctuation. Any errors do not interfere with understanding.	Has some errors in grammar, spelling, capitalisation or punctuation. Mistakes do not diminish understanding.	Has several errors in grammar, spelling, capitalisation or punctuation. Errors may interfere with understanding.	Has numerous errors in grammar, spelling, capitalisation or punctuation. Errors interfere with understanding.	Has serious errors in grammar, spelling, capitalisation or punctuation. Errors diminish understanding.
GRAMMAR					
VOCABULARY	Words can be read correctly and without prompting. Independently uses context clues to decode unfamiliar words.	Can read/decode words with little or no prompting.	Some words are recognised but prompting is more frequent.	Direct help may be needed to elicit the correct word. Searches for context clues within sentences unsuccessfully.	Unable to identify unfamiliar words. Refused to read words and no attempt to read word using context clues or other prompting.
DETAILS	Is able to identify details without help. Some details are included.	Is able to identify details with little help. Some details are included.	May have to ask more direct questions to get the answer. Some details are included.	Can recall detail with help. More direct and straightforward questions were asked. Little to no details are included in answer.	Unable to identify specific details. The answer might not make sense. Answers are incorrect.

In conclusion, assessing reading is how we can range the assignments from simply asking questions during a lesson to class presentations after a unit of study. It helps us to understand the strength and needs of each of our students. Also, assessing reading are meant to provide feedback on the skills, processes and knowledge resources that represent reading abilities.

BAB XI

ASSESSING WRITING

Assessing writing refers to the process of evaluating a learner's written performance based on specific criteria such as content, organization, grammar, vocabulary, and mechanics. It is an essential part of language teaching and learning because writing is a productive skill that demonstrates a learner's ability to express ideas clearly and effectively in written form. Writing assessments can be either formative, where feedback is provided to support learning, or summative, where scores are used to determine achievement or proficiency levels.

There are several approaches to assessing writing, including holistic, analytic, and primary trait scoring. Holistic assessment involves giving a single overall score based on a general impression of the writing. Analytic assessment breaks down writing into different components (e.g., content, organization, grammar, vocabulary, and mechanics), and each is scored separately to provide more detailed feedback. Primary trait scoring focuses on evaluating a specific feature relevant to a particular writing task, such as the argument in a persuasive essay or narrative structure in a story.

Effective writing assessment should be reliable, valid, and fair. Reliability ensures that scores are consistent across different raters and times. Validity ensures that the assessment truly measures writing ability rather than unrelated skills like typing speed or handwriting. Fairness involves creating assessment tasks that are appropriate for the students' language levels and free of bias. Teachers should also provide clear rubrics and feedback to guide learners in improving their writing skills over time.

A. Genre of Assessing Writing

1. Academic writing

- papers and general subject reports
- essays, compositions
- academically focused journals
- short-answer test responses
- technical reports (e.g., lab reports)
- theses, dissertations

2. Job related writing

- messages (e.g., phone messages)

- letters/emails
 - memos (e.g., interoffice)
 - reports (e.g., job evaluations, project reports)
 - schedules, labels, signs
 - advertisements, announcements
 - manuals
3. Personal writing
- letters, emails, greeting cards, invitations
 - messages, notes
 - calendar entries, shopping lists, reminders
 - financial documents (e.g., checks, tax forms, loan applications)
 - forms, questionnaires, medical reports, immigration documents
 - diaries, personal journals
 - fiction (e.g., short stories, poetry)

B. Type of Writing Performances

1) Imitative

This category includes the ability to spell correctly and to perceive phoneme-grapheme correspondences in the English spelling system. It is a level at which learners are trying to master the mechanics of writing. At this stage, form is the primary if not exclusive focus, while context and meaning are of secondary concern.

2) Intensive

skills in producing appropriate vocabulary within a context, collocations and idioms, and correct grammatical features up to the length of a sentence. Meaning and context are of some importance in determining correctness and appropriateness, but most assessment tasks are more concerned with a focus on form, and are rather strictly controlled by the test design.

3) Responsive

assessment tasks require learners to perform at a limited discourse level, connecting sentences into a paragraph and creating a logically connected sequence of two or three paragraphs. Tasks respond to pedagogical directives, lists of criteria, outlines, and other guidelines.

4) Extensive

Extensive writing 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 project report, or even a thesis. Writers focus on achieving a purpose, organizing and developing ideas logically, using details to support or illustrate ideas, demonstrating syntactic and lexical variety, and in many cases, engaging in the process of multiple drafts to achieve a final product. Focus on grammatical form is limited to occasional editing or proofreading of a draft.

Extensive writing 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 project report, or even a thesis. Writers focus on achieving a purpose, organizing and developing ideas logically, using details to support or illustrate ideas, demonstrating syntactic and lexical variety, and in many cases, engaging in the process of multiple drafts to achieve a final product. Focus on grammatical form is limited to occasional editing or proofreading of a draft.

C. Micro – Macro Skills of Writing

- Micro Skills

- 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, and rules.
- 5) Express a particular meaning in different grammatical forms.
- 6) Use cohesive devices in written discourse.

- Macro Skills

- 1) Use the rhetorical forms and conventions of written discourse.
- 2) Appropriately accomplish the communicative functions of written texts according to form and purpose.
- 3) Convey links and connections between events, and communicate such relations as main idea, supporting idea, new information, given information, generalization, and exemplification.
- 4) Distinguish between literal and implied meanings when writing.

- 5) Correctly convey culturally specific references in the context of the written text.
- 6) 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.

D. Designing Assessment : imitative writing

Many beginning-level English learners, need basic training in and assessment of imitative writing: the rudiments of forming letters, words, and simple sentences. We examine this level of writing first.

1. Tasks in [Hand] Writing Letters, Words, and Punctuation

- Listening cloze selection tasks.

These tasks combine dictation with a written script that has a relatively frequent deletion ratio (every fourth or fifth word, perhaps).

- Picture-cued tasks.

Familiar pictures are displayed, and test-takers are told to write the word that the picture represents. Assuming no ambiguity in identifying the picture (cat, hat, chair, table, etc.), no reliance is made on aural comprehension for successful completion of the task.

- Form completion tasks.

A variation on pictures is the use of a simple form (registration, application, etc.) that asks for name, address, phone number, and other data. Assuming, of course, that prior classroom instruction has focused on filling out such forms, this task becomes an appropriate assessment of simple tasks such as writing one's name and address.

- Converting numbers and abbreviations to words.

Some tests have a section on which numbers are written—for example, hours of the day, dates, or schedules and test-takers are directed to write out the numbers. This task can serve as a reasonably reliable method to stimulate handwritten English. It lacks authenticity, however, in that people rarely write out such numbers (except in writing checks), and it is more of a reading task (recognizing numbers) than a writing task.

2. Spelling Tasks and Detecting Phoneme-- Grapheme Correspondences

Spelling tests.

- In a traditional, old-fashioned spelling test, the teacher dictates a simple list of words, one word at a time, followed by the word in a sentence, repeated again, with a pause for test-takers to write the word. Scoring emphasizes correct spelling. You can help to control for listening errors by choosing words that the students have encountered before-- words that they have spoken or heard in their class.
- Picture-cued tasks.
Pictures are displayed with the objective of focusing on familiar words whose spelling may be unpredictable. Items are chosen according to the objectives of the assessment, but this format is an opportunity to present some challenging words and word pairs: boot/book, read/reed, bit/bite, etc.

E. Vocabulary Assessment Task

Most vocabulary study is carried out through reading. The major techniques used to assess vocabulary are (a) defining and (b) using a word in a sentence. The latter is the more authentic, but even that task is constrained by a contrived situation in which the test-taker, usually in a matter of seconds, has to come up with an appropriate sentence, which may or may not indicate that the test-taker "knows" the word. Vocabulary assessment is clearly form-focused in the above tasks, but the procedures are creatively linked by means of the target word, its collocations, and its morphological variants. At the responsive and extensive levels, where learners are called upon to create coherent paragraphs, performance obviously becomes more authentic, and lexical choice is one of several possible components of the evaluation of extensive writing.

Test-takers read:

- Write two sentences, A and B. In each sentence, use the two words given.
 - interpret, experiment _____
 - interpret, language _____
- Write three words that can fit in the blank.

To interpret a(n) _____

 - _____
 - _____
 - _____
- Write the correct ending for the word in each of the following sentences:

Someone who interprets is an interpret _____

Something that can be interpreted is interpret _____

Someone who interprets gives an interpret _____

F. Short-Answer and Sentence Completion Tasks

Some types of short-answer tasks were discussed in Chapter 8 because of the heavy participation of reading performance in their completion. Such items range from very simple and predictable to somewhat more elaborate responses. Look at the range of possibilities.

The reading-writing connection is apparent in the first three item types but has less of an effect in the last three, where reading is necessary in order to understand the directions but is not crucial in creating sentences. Scoring on a 2-1-0 scale (as described above) may be the metopropriate way to avoid selfing about the appropriateness of a response.

Test-takers see:

- Alicia: Who's that?
 Tony: _____ Gina.
 Alicia: Where's she from?
 Tony: _____ Italy.
- Jennifer: _____?
 Kathy: I'm studying English.
- Restate the following sentences in your own words, using the underlined word. You may need to change the meaning of the sentence a little.
 - I never miss a day of school. always
 - I'm pretty healthy most of the time. seldom
 - I play tennis twice a week. sometimes
- You are in the kitchen helping your roommate cook. You need to ask questions about quantities. Ask a question using *how much* (#4a) and a question using *how many* (#4b), using nouns like *sugar, pounds, flour, onions, eggs, cups*.
 - _____
 - _____
- Look at the schedule of Roberto's week. Write two sentences describing what Roberto does, using the words *before* (#5a) and *after* (#5b).
 - _____
 - _____
- Write three sentences describing your preferences: #6a: a big, expensive car or a small, cheap car; #6b: a house in the country or an apartment in the city; #6c: money or good health.
 - _____
 - _____
 - _____

G. Paraphrasing

One of the more difficult concepts for second language learners to grasp is paraphrasing. The initial step in teaching paraphrasing is to ensure that learners understand the importance of paraphrasing: to say something in one's own words, to avoid plagiarizing, to offer some variety in expression. With those possible motivations and purposes in mind, the test designer needs to elicit a paraphrase of a sentence or paragraph, usually not more. Scoring of the test-taker's response is a judgment call in which the criterion of conveying the same or similar message is primary, with secondary evaluations of discourse, grammar, and vocabulary. Other components of analytic or holistic scales (see discussion below, page 242) might be considered as criteria for an evaluation. Paraphrasing is more often a part of informal and formative assessment than offormal, summative assessment, and therefore student responses should be viewed as opportunities for teachers and students to gain positive washback on the art o paraphrasing.

BAB XII

IDENTIFYING RUBRICS

Rubrics are essential tools in language assessment because they provide clear, structured criteria for evaluating student performance. A rubric typically consists of several components: assessment criteria (what is being measured), performance levels (such as excellent, good, fair, poor), and descriptors (detailed explanations of each level). In language learning, rubrics help both teachers and students understand expectations and areas for improvement. Rubrics can be used for all language skills: listening, speaking, reading, and writing.

A. Types of Rubrics in Language Assessment:

1. **Holistic Rubric:** This rubric assigns a single score based on the overall impression of a learner's performance. It is quick and efficient but lacks detailed feedback. For example, in speaking assessments, a holistic rubric might judge overall fluency and coherence with a single rating.
2. **Analytic Rubric:** This rubric breaks down performance into separate categories (e.g., grammar, vocabulary, pronunciation, content, coherence). Each category is scored independently, allowing for more specific feedback. This is especially useful in writing or speaking assessments where multiple skills are demonstrated.
3. **Primary Trait Rubric:** This focuses on one key feature of performance related to the task. For example, in a persuasive essay, the primary trait might be the strength of the argument. This type of rubric is useful for task-specific evaluations.

B. Elements of Rubrics

Typically designed as a grid-type structure, a grading rubric includes criteria, levels of performance, scores, and descriptors which become unique assessment tools for any given assignment. The table below illustrates a simple grading rubric with each of the four elements for a history research paper.

Sample rubric demonstrating the key elements of a rubric

Criteria	Excellent (3 points)	Good (2 points)	Poor (1 point)
Number of sources	Ten to twelve	Five to nine	One to four
Historical accuracy	No apparent inaccuracies	Few inaccuracies	Lots of historical inaccuracies
Organization	Can easily tell from which sources information was drawn	Can tell with difficulty from where information came	Cannot tell from which source information came
Bibliography	All relevant bibliographic information is included	Bibliography contains most relevant information	Bibliography contains very little information

C. Criteria

Criteria identify the trait, feature or dimension which is to be measured and include a definition and example to clarify the meaning of each trait being assessed. Each assignment or performance will determine the number of criteria to be scored. Criteria are derived from assignments, checklists, grading sheets or colleagues.

D. Examples of Criteria for a term paper rubric

1. Introduction
2. Thesis
3. Arguments/analysis
4. Grammar and punctuation
5. Spelling
6. Internal citations
7. Conclusion
8. References

BAB XIII

ITEMS ANALYSIS

Item analysis is a process used to evaluate the effectiveness of individual test items (questions) within an assessment. It helps determine whether the items are functioning as intended, contributing to the overall assessment's validity and reliability. This involves examining statistics like item difficulty and discrimination to identify items that may be too easy, too difficult, or not effectively distinguishing between high and low-performing students.

Key aspects of item analysis:

1. **Item Difficulty:**
This refers to how easy or difficult a test item is for the test-takers. It's typically measured by the proportion of students who answer the item correctly (p-value). A higher p-value indicates an easier item, while a lower p-value indicates a more difficult item.
2. **Item Discrimination:**
This measures how well an item differentiates between students who do well on the overall test and those who don't. A good discriminating item will be answered correctly more often by high-performing students and incorrectly more often by low-performing students.
3. **Distractor Analysis:**
For multiple-choice questions, item analysis also examines the effectiveness of the answer choices (distractors). It looks at how many students chose each distractor, and whether the incorrect choices are attracting students who are likely to be struggling with the material.
4. **Test Score Reliability:**
Item analysis can provide information about the reliability of the test as a whole by identifying poorly performing items that might be negatively impacting the overall score.

Why is item analysis important?

1. **Improving Test Quality:**
By identifying poorly performing items, item analysis helps educators improve the quality of their assessments for future use. This can involve revising unclear wording, adjusting difficulty levels, or replacing items altogether.
2. **Enhancing Validity and Reliability:**
A well-analyzed test is more likely to be a valid and reliable measure of student knowledge and skills.
3. **Informing Instruction:**
Item analysis can provide valuable insights into student learning, revealing areas where students are struggling and where instruction may need to be adjusted.
4. **Making Informed Decisions:**
In summative assessments, item analysis can help ensure that decisions about student grades or placement are based on accurate and reliable data.

BAB XIX

VALIDITY & RELIABILITY

Validity refers to the extent to which a test or assessment accurately measures what it is intended to measure. In language assessment, a valid test truly evaluates a learner's language ability—such as reading, writing, speaking, or listening—without being influenced by unrelated factors like test-taking skills or background knowledge not relevant to the language skill. There are different types of validity:

1. **Content validity:** whether the test covers the intended material.
2. **Construct validity:** whether the test truly measures the theoretical language construct (e.g., grammar, fluency).
3. **Face validity:** whether the test appears appropriate and relevant to both test takers and teachers.
4. **Criterion-related validity:** whether the test results correlate with other reliable measures of the same ability.

Reliability refers to the consistency and stability of test results over time or across different raters. A reliable assessment produces similar results under consistent conditions. For example, if a student takes the same test twice under the same conditions, or if two trained teachers score the same student's speaking performance, the results should be similar. Types of reliability include:

1. **Test-retest reliability:** consistency over time.
2. **Inter-rater reliability:** consistency between different scorers.
3. **Internal consistency:** how well the items on a test measure the same construct.

In summary, validity ensures that a test measures the right thing, while reliability ensures that it does so consistently. Both are essential for high-quality language assessment. A test can be reliable but not valid, but a test cannot be valid if it is not reliable.

In line with this definition, Reliability and validity are concepts used to evaluate the quality of research. They indicate how well a method, technique or test measures something. Reliability is about the consistency of a measure, and validity is about the accuracy of a measure. It's important to consider reliability and validity when you are creating your research

design, planning your methods, and writing up your results, especially in quantitative research. Reliability and validity are closely related, but they mean different things. A measurement can be reliable without being valid. However, if a measurement is valid, it is usually also reliable.

Validity refers to how accurately a method measures what it is intended to measure. If research has high validity, that means it produces results that correspond to real properties, characteristics, and variations in the physical or social world, while Reliability refers to how consistently a method measures something. If the same result can be consistently achieved by using the same methods under the same circumstances, the measurement is considered reliable.

Types of Reliability and Validity

Reliability

Type of reliability	Measures the consistency of...
Test-retest	The same test over time .
Interrater	The same test conducted by different people .
Parallel forms	Different versions of a test which are designed to be equivalent.
Internal consistency	The individual items of a test.

1. Test-retest reliability measures the consistency of results when you repeat the same test on the same sample at a different point in time. You use it when you are measuring something that you expect to stay constant in your sample.

A test of color blindness for trainee pilot applicants should have high test-retest reliability, because color blindness is a trait that does not change over time.

2. Interrater reliability (also called interobserver reliability) measures the degree of agreement between different people observing or assessing the same thing. You use it when data is collected by researchers assigning ratings, scores or categories to one or more variables.

In an observational study where a team of researchers collect data on classroom behavior, interrater reliability is important: all the researchers should agree on how to categorize or rate different types of behavior.

3. Parallel forms reliability measures the correlation between two equivalent versions of a test. You use it when you have two different assessment tools or sets of questions designed to measure the same thing.

4. Internal consistency assesses the correlation between multiple items in a test that are intended to measure the same construct.

You can calculate internal consistency without repeating the test or involving other researchers, so it's a good way of assessing reliability when you only have one data set.

Validity

Validity tells you how accurately a method measures something. If a method measures what it claims to measure, and the results closely correspond to real-world values, then it can be considered valid. There are four main types of validity:

- **Construct validity:** Does the test measure the concept that it's intended to measure?
- **Content validity:** Is the test fully representative of what it aims to measure?
- **Face validity:** Does the content of the test appear to be suitable to its aims?
- **Criterion validity:** Do the results accurately measure the concrete outcome they are designed to measure?

1. A construct validity refers to a concept or characteristic that can't be directly observed, but can be measured by observing other indicators that are associated with it.

There is no objective, observable entity called "depression" that we can measure directly. But based on existing psychological research and theory, we can measure depression based on a collection of symptoms and indicators, such as low self-confidence and low energy levels.

2. Content validity assesses whether a test is representative of all aspects of the construct. A mathematics teacher develops an end-of-semester algebra test for her class. The test should cover every form of algebra that was taught in the class. If some types of algebra are left out, then the results may not be an accurate indication of students' understanding of the subject. Similarly, if she includes questions that are not related to algebra, the results are no longer a valid measure of algebra knowledge.
3. Face validity considers how suitable the content of a test seems to be on the surface. It's similar to content validity, but face validity is a more informal and subjective assessment.

You create a survey to measure the regularity of people's dietary habits. You review the survey items, which ask questions about every meal of the day and snacks eaten in between for every day of the week. On its surface, the survey seems like a good representation of what you want to test, so you consider it to have high face validity.

4. Criterion validity evaluates how well a test can predict a concrete outcome, or how well the results of your test approximate the results of another test.

A university professor creates a new test to measure applicants' English writing ability. To assess how well the test really does measure students' writing ability, she finds an existing test that is considered a valid measurement of English writing ability, and compares the results when the same group of students take both tests. If the outcomes are very similar, the new test has high criterion validity.

REFERENCES

- Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Thousand Oaks, CA: Corwin Press, Inc.
- Bachman, L.F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
- Bachman, L.F. and Palmer, A.S. 1996. *Language Testing in Practice*. New York: Oxford University Press.
- Brown, H.D. 2004. *Language Assessment: Principles and Classroom Practices*. White Plains, NY: Pearson Education.
- Brown, H.D. and Abeywickrama, P. 2010. *Language Assessment: Principles and Classroom Practices* (2nd edition). White Plains, NY: Pearson Education.
- Burke, K. 2009. *How to Assess Authentic Learning* (5th edition). Thousand Oaks, CA: Corwin.
- Heaton, J.B. 1988. *Writing English Language Tests* (New Edition). London: Longman.
- Hill, B.C., Ruptic, C., and Norwick, L. 1998. *Classroom Based Assessment*. Norwood, MA: Christopher-Gordon.
- Kemdikbud. 2013. *Peraturan Menteri Pendidikan dan Kebudayaan R.I. No. 81A tentang Implementasi Kurikulum. Lampiran IV: Pedoman Umum Pembelajaran*. Jakarta: Kemdikbud.
- Koizumi, R., & In'nami, Y. (2020). Structural equation modeling of vocabulary size and depth using conventional and Bayesian methods. *Frontiers in Psychology*, 11, 512909.
- Laufer, B., & Nation, I. S. P. (1999). A vocabulary size test of controlled productive ability. *Language Testing*, 16(1), 33-51. [https://www.lexutor.ca/tests/levels/productive/\(PVLt\)](https://www.lexutor.ca/tests/levels/productive/(PVLt))
- Meara, P., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, 28(1), 19-30.
- Nation, I. S. P. (2022). *Learning vocabulary in another language*. (3rd ed.). Cambridge University Press.
- Nguyen, T. M. H., & Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*, 21(3), 298-320.
- O'Malley, J.M. and Pierce, L.V. 1996. *Authentic Assessment for English Language Learners: Practical Approaches for Teachers*. White Plains, NY: Addison Wesley.
- Richards, J.C. and Rodgers, T.S. 2001. *Approaches and Methods in Language Teaching* (2nd edition). Cambridge: Cambridge University Press.
- Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

- Sasao, Y., & Webb, S. (2017). The word part levels test. *Language Teaching Research*, 21(1), 12-30. <http://doi.org/10.1177/1362168815586083>
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke: Palgrave Macmillan.
- Stevens, D. D., & Levi, A. J. (2005). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning*. Sterling, VA: Stylus.
- The Teaching, Learning, and Technology Group (n.d.). Rubrics: Definition, tools, examples, references. <http://www.tltgroup.org/resources/flashlight/rubrics.htm>
- Webb, S. (2021). Research investigating lexical coverage and lexical profiling: What we know, what we don't know, and what needs to be examined. *Reading in a Foreign Language*, 33(2), 287-302.
- Webb, S., Sasao, Y., & Balance, O. (2017). The updated vocabulary levels test: Developing and validating two new forms of the VLT. *ITL- International Journal of Applied Linguistics*, 168(1), 34-70.
- Weir, C.J. 1990. *Communicative Language Testing*. New York: Prentice Hall.
- Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, 26(4), 696-725.