Evaluation, measurement, and testing are inter-related concepts. Evaluation is the superordinate term to both measurement and testing. It is a systematic gathering of information for the purpose of making decisions. It allows one to make judgement about the desirability or value of a measure, and determines the congruence between performance and objectives. Measurement encompasses testing when decision making is done through the use of a specific sample of behavior. It quantifies the characteristics of persons according to explicit procedures and rules. A test is a method of collecting numerical data on an attribute, or a formal method of collecting numerical data. It is a presentation of a set of questions to be answered. A psychological or educational test is a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual by a collection of letters, or by naturalistic observation.

Tests are categorized into two major groups that help teachers to match the correct type of test with any type of decision: Norm-referenced tests and criterion- referenced tests. They differ in their intended purposes, content selection, and the scoring process which defines how the test results must be interpreted. In norm-referenced tests (NRTs) which help compare one student's performance with the performances of a large group of students, any student who then takes the test once it is published has his or her scores compared to those of the norm group and the student learns how S/he scored relative to the students who took the test when it was normed. NRTs highlight achievement differences between and among students to produce a dependable rank order of students across a continuum of achievement from high achievers to low achievers. Criterion-referenced tests (CRTs), on the other hand, are focused on the abilities and knowledge of the test takers. CRTs may be used as one piece of information to decide how well the student is learning the desired curriculum, and how well the school is teaching that curriculum. Students rarely know in any detail what content to expect on an NRT. On a CRT students can predict the question formats on the test and the language points that will be tested. NRT is relatively long and contains various types of question content, and students may find it difficult to know exactly what will be tested. In contrast, CRTs usually consist of numerous, shorter subtests, and each subtest represents a different instructional objective, and often, each objective will have its own subtest.

A variety of different types of decisions are made in almost any language program, and teachers must be clear about his or her purpose for making a given decision, and then match the correct type of test to that purpose. In administering language programs, four kinds of decisions are: Proficiency, placement, achievement, and diagnostic. The purpose of a placement test is to help sort new students into teaching groups of the same level, and the subject matter of any reading and listening texts, speaking and writing tasks is usually based on common human experience. A diagnostic test helps teachers and the students know how well they have learnt particular course elements and are typically done at the end of course book units or recent classwork. The test content and question types should be familiar to students and you should expect a high degree of success as they know what’s in the test. The results also show which areas need revising with the class or individuals. Achievement tests are related to the aim of syllabus, and measure what has been learnt over a longer period of time than diagnostic tests. These tests show the standard reached, and results can be compared to other students. Proficiency tests have a future orientation, and are often used by further education bodies, employers and immigration authorities to determine acceptable standards for applicants. These four test types overlap; there are elements of proficiency in the three other types, but all have to be valid, reliable and practical. Language testing, has evolved and expanded in a number of ways in the past decades.

Testing trends are labeled into the pre-scientific period, the psychometric or structuralist period, and the integrative or sociolinguistic period. The pre-scientific period assumes that one can and must rely completely on the subjective judgment of an experienced teacher, who can identify after a few minutes of conversation, or after reading a student’s essay, what mark to give him or her in order to specify the related language ability. The pre-scientific period is characterized by translation tests developed exclusively by the classroom teachers. In psychometric-structuralist period of language testing, language tests became increasingly scientific, reliable, and precise. The testers and psychologists, being responsible for the development of modern theories and techniques of educational measurement, were trying to provide objective measures, using various statistical techniques to assure reliability and certain kind of validity. Language learning was chiefly concerned with the systematic acquisition of a set of habits. In this period, for the first time language test development followed scientific principles. In psychometric-structuralist period, the claim was that communicative competence is so global that it requires the integration of all linguistic abilities. Such global nature cannot be captured in additive tests of grammar, reading, vocabulary, and other discrete points of language.

Some issues may arise in language testing. There are various approaches to language teaching and there should be some ways to integrate teaching and testing. Another issue in language testing is the channel for communication, which can be written or oral. Specific tests are designed for assessing each channel. The other issue is the mode of communication. The two modes of communication are receptive and productive modes. Some test items are receptive, some are productive, and some may include both receptive and productive modes.

An item is the smallest and the basic unit of language testing. In item format analysis,testers focus on the degree to which each item is properly written so that it measures all and only the desired content. Such analyses often involve making judgments about the adequacy of item formats. Teachers will of course want their item formats to match the purpose and content of the item. Teachers should make sure that each question has only one correct answer. Each item should be written at approximately the level of proficiency of the test takers. The use of negatives and double negatives may be confusing and should be used if the purpose of the item is to test negatives. Teachers should also avoid including extra information that is irrelevant to the concept or skill being tested.All teachers should also be on the alert for bias that may have crept into their test items. The statement should be carefully worded to avoid any ambiguities that might cause the students to miss it for the wrong reasons (true/ false). Teachers should also avoid absoluteness clues. Absoluteness clues allow students to answer correctly without knowing the correct response (true/ false). In designing each test item, teachers should consider the given points.