Chapter Two

Language Testing: An Historical Overview

2.1. Introduction

The present chapter intends to focus on an historical overview of language testing with a review of the available literature. In addition, this chapter will provide a brief survey of the types, approaches, and characteristics of the available English language tests.

2.2. Historical Overview of Language Testing

Researches in the area of language testing over the last few decades have contributed a lot in making a test more and more practical, reliable and valid. The present state of language testing is the result of a consistent theoretical advancement - though slow and even lacking in operation and application especially in ESL / EFL context - over the last one century in the shape of experimentations and innovations in language testing. The historical overview of the developments in this area is best viewed in the nine historical phases categorized by Spolsky (2000:536-552) in his review article, entitled “Language Testing in The Modern Language Journal”, where he reviews the articles published in the Modern Language Journal.
over eighty years, reflecting upon the nature, views, principles and practices of testing in these phases.

2.2.1. The Early Years:

This phase is most prominently known for the ‘aural and oral test’ that was designed and implemented by the committee of the Association of Modern Language Teachers for university admission tests in French, German, and Spanish. This test was based on data collected from 1000 schools, where the “majority of respondents agreed that a test was the only fair way to recognize and encourage the widespread direct method of teaching, which emphasized speaking the language” (Spolsky, 2000:537). Carroll (1954) considers this test as one of the earliest calls for objective psychological testing. This article, according to Spolsky (2000:537), was significant in three ways – first, because this aural-oral test concentrated on the spoken language rather than written language and ‘recognized clearly the curricular impact (washback) of a test’. Secondly, in order to meet the feasibility criteria of a test, we find ‘ease triumphing over principle’. Thirdly, objective tests became widely acceptable for uniformity.

The Committee on Resolutions and Investigations of the Association of Modern Language Teachers of the Middle States and Maryland, which became a basis for the innovations in language testing 30 years later by the
Armed Services Training Program (ASTP) and the Audio-Lingual Method after that, observed in 1917 that ‘no actual oral test is included in this (entrance examinations for French, Spanish and German) examination’ (Committee on Resolutions and Investigations, 1917:252, in Fulcher, 2003:02). However, it further claims that ‘no candidate could pass it, who had not attained abundant oral, as well as aural training’ (Committee on Resolutions and Investigations, 1917:252, in Fulcher, 2003:02).

The story in the United Kingdom is found to be different. ‘While testing speaking in the United States was frequently seen as ‘desirable but not feasible’, the University of Cambridge Local Examinations Syndicate (UCLES) was not hampered by concerns over reliability or measurement theory. A sub-test of spoken English was included in the Certificate of Proficiency in English at its introduction in 1913, (Roach, 1945, in Fulcher, 2003:05).

2.2.2. The Inter-war Years:

These years have been significant in so many ways in the area of experimentations, innovations and theoretical advancements of language testing. Tests of oral and aural skills started increasing in number. Spolsky (2000:538) lists such references as Thorndike (1921), Decker (1925),
Henmon et al. (1929), Lundeberg (1929), Tharp (1930), Clarke (1931), and Rogers & Clarke (1933).

Wood (1927:96), in his pioneering work on language tests for New York Schools, acknowledged the importance of using ‘conversational materials’ for assessing students’ speaking abilities, but this test remained ‘paper and pencil test’ because speaking tests would be ‘subjective’ and logistically difficult to put into operation.

Seeking objectivity became the key intention of the test developers in this phase. Hence they concentrated on ‘new-type’ multiple choice tests as reliable, objective measures of language ability (Fulcher, 2003:02). The first true speaking test used in North America was The College Board’s English Competence Examination, introduced in 1930 for overseas students applying to study at US colleges and universities (College Entrance Examination Board, 1939, in Fulcher 2003:02). The College Entrance Examination Board was given the task of providing this test by the American Association of Collegiate Registrars. It was used effectively only until 1935 due to financial reasons.

Some other works in this direction are Broom & Kaulfers (1927), Cheydleur (1928) and Brinsmade (1928). Seibert & Goddard (1935) developed scoring key to attain objectivity in language testing. Smith & Cambell (1942) too
developed tests and a scoring key to relieve fears about multiple-choice tests.

Different types of tests were developed by Cheydleur (1931, 1932b), Kaulfers (1933), Haden & Stalnaker (1934), Kurath & Stalnaker (1936), Stalnaker & Kurath (1935), Feder & Cochran (1936), Fife (1937), and Kaulfers (1937).

Prognosis or Aptitude testing, initiated by Kaulfers (1930) became a significant concern in this phase. Spolsky (2002:539-541) list various studies in this regard. They are Cheydleur (1932a), Richardson (1933), Symonds (1930a, 1930b), Henmon (1934), Wagner & Strabel (1935), Michel (1934, 1936), Matheus (1937), Tallent (1938), Kaulfers (1939), Spoerl (1939), Maronpot (1939) and Gabbert (1941).

The goal of instruction became important for test developments. In this regard Spolsky (2000:541) mentions mainly two studies by Conant (1934) and Frantz (1939).

This phase has also been significant in witnessing the first detailed proficiency scale for language tests in Sammartino (1938).
“Unlike in the United States, the primary purpose of an examination in the United Kingdom was to support the syllabus and encourage good teaching and learning” (Brereton 1944, in Fulcher, 2003:06).

2.2.3. The World War II Experience:

Due to the scientific advancements like planes and radios, the focus in language teaching shifted from reading to aural-oral abilities of language learners. Similar to Sammartinos (1938), Kauflers (1944) developed a performance scale for measuring aural comprehension. Spolsky (2000:543) identifies Roach (1945) as working on the same lines in England. Since the Second World War, testing second language speaking attained priority over other skills.

“When it was realized that the American service personnel were ill equipped to communicate, that their inability to speak in a second language might be ‘a serious handicap to safety and comfort’, the push for tests of second language speaking began in earnest. From the needs of the military, the practice soon spread to colleges and universities” (Fulcher, 2003: xi).

Fulcher (2003:06) further notes:

“the Second World War was a watershed in the history of testing speaking. … The Army Specialized Training Program (ASTP) was created in 1942 in order to address the communication problems of American service personnel through the delivery of language
programmes that focused on speaking in the fields of engineering, medical, and area studies / language”.

Even Angiolillo (1947:32) identifies ASTP as the first language instruction programme with the specific aim ‘to impart to the trainee a command of the colloquial spoken form of a language and to give the trainee a sound knowledge of the area in which the language is used.’

In the UK too the testing of speaking had a military focus, though slow in comparison to the US. Fulcher (2003:08) informs:

“much of the language teaching that went on in the United Kingdom, and those part of the world over which the Allies had control, was done by the Army Education Corps. Much of this was teaching English to Allied Forces in Europe.”

2.2.4. The Early Post War Period:

This period contributed couple of more tests in this phase. Bottke & Milligan (1945) and Bovee & Froelich (1946), for instance, suggested some new aptitude tests; Clapp (1947) meditated on placement tests; Coutant (1948) and Kaulfers (1948) made special observations about language teaching in California; and Miller (1948) designed a test of Spanish-American culture.

In this period, there had been very reluctant language testing research as such. As language teaching was not a distinct discipline, language testing followed the general principles of testing available in humanities or social sciences. Classroom tests were conducted by the teachers themselves which
were the results of Grammar Translation or Reading-oriented methods that were in use at that time.

2.2.5. The 1950s:

This decade witnessed an addition of some more tests, mainly concerning aspects of oral communication. They are Raymond (1951), Furness (1955), Eley (1956), Borg & Goodman (1956), Buechel (1957), Harding (1958), Mueller (1959), Valley (1959), Ayer (1960), and Sadnavitch & Popham (1961).

This decade was dominated by the Foreign Service Institute (FSI) testing unit in the Army services, though political and bureaucratic problems held up the implementation work from 1952 to 1956. The FSI was again given the responsibility in 1956 to provide evidence of foreign language proficiency for all Foreign Service Personnel. In 1958 the FSI testing unit further developed the 1952 scales by adding a checklist of five factors for raters, each to be measured on the six-point scale. These five factors were accent, comprehension, fluency, grammar and vocabulary. This can be called to be the first step towards developing multiple trait rating.
2.2.6. The 1960s:

Spolsky (2000:544) reports that ‘interest in aptitude reemerged in the postwar period and took on a new urgency …’. This resulted in such studies as Salomon (1954), Eterno (1961), Blinksenstaff (1963), Mendeloff (1963), Guerra, Abrahamson, and Newmark (1964), Di Giulio (1967), McCarthur (1965), and Shane (1971).

Due to the success of the FSI rating, the method was adopted and even adapted by the Defence Language Institute (DLI), the Central Intelligence Agency (CIA), and the Peace Corps. Quinones (Fulcher 2003:10). It is further informed that in 1968, partly as a result of experiences of language needs during the Vietnam War, these diverse agencies came together to produce a standardized version, today known as Interagency Language Roundtable (ILR).

From the early 1950s through the late 1960s, when contrastive analysis was a thriving discipline and structuralism and behaviorism combined to make language teaching scientific, testing focused on specific language elements such as phonological, grammatical and lexical contrasts between the two languages. Tests during this decade stressed mastery of discrete linguistic skills. That is why this approach came to be known as the ‘discrete point’ approach to language testing as against the ‘integrative’ approach. Such
discrete item tests were constructed by breaking down the language into its component parts relating to the four skills of listening, speaking, reading and writing and various hierarchical units of language in phonology, graphology, morphology, lexis and syntax.

2.2.7. The 1970s:


Liskin-Gasparro (1984b:22) provides information regarding the substantial use of the FSI by the Peace Corps. The new testing system for academics and teachers outside the control of the government agencies was introduced. In the 1970s, as a consequence, it was adopted by many universities and states for the purpose of bilingual teacher certification. Barnwell (1987:36) identified three reasons for the popularity of the FSI approach in the educational context. These are:

i. as a direct test of speaking ability, having high validity,
ii. high inter rater reliability, and
iii. the growing interest in the notional/functional approach to teaching English.
Due to these reasons the relationship between teaching and testing was also realized. Carroll’s (1967) experimentation with German, Russian, French, and Spanish students in the USA became the focus of considerable attention by 1979, when it was replicated by the Educational Testing Services (ETS) (Liskin-Gasparro 1984b:27).

Besides this in 1979 the President’s Commission on Foreign Language and International Studies recommended the setting up of a National Criteria and Assessment Program to develop language tests and to assess language learning in the US.

This decade has also been able to raise various diversified issues such as teachers’ competence to test (Taylor, 1972), differences between men and women response in tests (Jacobson & Imhoof, 1974), similarities in first and second language learning (Politzer, 1974; Oller 1976a), computerized testing (Boyle, Smith, and Eckert, 1976), repetition tasks (Bell, Doyle, and Talbott 1977), enforced latency period for tests (Meredith, 1978), aural comprehension test (Madsen 1979), and repetition and dictation (Natalicio, 1979).
2.2.8. The 1980s:

The recommendations of the President’s Commission of 1979 gave the ACTFL (The American Council on the Teaching of Foreign Languages) the role of producing National Criteria, and the ACTFL Provisional Proficiency Guidelines appeared in 1982 (ACTFL, 1982), the complete Guidelines in 1986 (ACTFL, 1986), and the revised Guidelines in 1999 (ACTFL, 1999). These guidelines gave birth to the Proficiency Movement.

The 1980s argued the need for a ‘common yardstick’ to develop listening and reading tests (Woodford, 1980). It focused on direct proficiency testing, communicative skills in classroom tests, and the computerized testing (Clark, 1983). It witnessed more samples of cloze tests (Shohamy, 1982; Heilenman, 1983; Bensoussan and Ramzar, 1984), and continued the debate over prognostic testing (Curtin, Avner, and Smith 1983).

This decade has been remarkable for the publication of the ACTFL (the American Council on the Teaching of Foreign Languages) Proficiency Guidelines in 1982 which provided a common yardstick, but it had to undergo certain attacks, from Savignon (1985), Lantolf and Frawley (1985), Schulz (1986) and Kramsch (1986). This Guideline actually talked of a shift in testing target from the ‘grammar-based achievement testing’ to the ‘communicative-based and oral proficiency’. Lowe (1986) came forward to
defend the Guidelines, but suggested that there is a need for further studies and research in this direction. The debate took a positive shape with the contributions of such researchers as Hieke (1985) and Byrnes (1987).

Besides these guidelines, testing situation in this decade was involved with some other issues too. These issues involved such aspects as scoring criteria (Currall and Kirk, 1986), relationship between intelligence and language ability (Boyle, 1987; Cooper, 1987), the relationship between students’ overall grade and language ability (Spurling, 1987), and the goal of developing a ‘common metric’ (Lambert, 1987). Premised on ACTFL Proficiency Guidelines for foreign language reading, Lee & Musumeci (1988) constructed a test based on text-type, reading skill, and task-based performance.

From the late 1960s to the present time, the dissatisfaction with structuralism and behaviorism led to linguistic research on communicative competence and the context of language use. According to Spolsky (1978) and Jones (1977), there has been a large discrepancy between language teaching and testing. While the aim of teaching was to develop practical communication, language testing stressed on the mastery of linguistic elements. However, recently when the goal of teaching has become communicative, language tests have started testing communication and other pragmatic skills (Morrow
1977, Canale and Swain 1980, Carroll 1980). Testers have come to realize that the whole of the communicative event is considerably greater than the sum of its linguistic elements (Clark 1983). As against the ‘discrete’ point approach, ‘integrative’ approach emerged with its emphasis on communication, authenticity and context. Oller (1976, 1978) criticized the ‘discrete’ approach and said that language competence is a unified set of interacting abilities which can not be separated and tested and claimed that it required integration rather than testing of discrete items of grammar, reading and vocabulary. Following this line of thinking, two types of integrative tests emerged. ‘Cloze’ tests are reading passages mutilated by the deletion of every nth word. The subject is required to fill up the blanks with appropriate words, which tests the subjects’ competence in a language. The ‘dictation’ tests are short passages read to the students by the teacher: the passage is read at a normal speed and the students listen. In the second reading it is broken into phrases and the students write what they listen. Lastly, the work of the students is checked by themselves while the teacher reads the passage at a normal speed. Dictation and cloze tests are appropriate integrative tests (Savignon 1982, Oller 1979, Streiff 1975). At present, when the aim of teaching is to develop communicative competence, the tests are largely performance based which requires the learners to use language naturally.
There is a call for direct testing which tests the learners in a variety of language functions.

**2.2.9. The 1990s:**

This decade witnessed further studies concerning some recent issues pertaining to language testing. It becomes evident from various works like Heining-Boynton (1990) who prepared an inventory for evaluating language teaching programmes in elementary schools; Meredith (1990) on the use of Oral Proficiency Interview (OPI); Anderson (1991) on the individuality of proficiency tests; Dunkel (1991) on using computer to test listening comprehension; Abraham & Chapelle (1992) on Cloze tests; Shohamy, Gordon, and Kraemer (1992) on the usefulness of writing scales and intensive training for raters; Stansfield & Kenyon (1992) on developing a tape-recorded version of the interview; Stansfield, Scott, & Kenyon (1992) on the development of a test for translators.

The historical overview of language testing is also done by researchers in terms of three stages in the recent history of language testing: the Pre-scientific, the Psychometric-Structuralist, and the Psycholinguistic-Sociolinguistic, as specified by Spolsky (1975). These phases also reflect the intentions and theoretical basis for language testing before the emergence of Communicative testing.
i) **The Pre-Scientific Era:** Although Spolsky (1975) traces this era from the Chinese civil service examinations more than two thousand years ago; he considers its present form from the 18th century Cambridge Tripos onwards. It was characterized by the use of essays, open-ended examinations, oral examining. Testing in this era did not rely on linguistic theory.

ii) **The Psychometric-Structuralist Era:** This era refers to the joint venture of the psychometrics and the structural linguistics. While psychometrics suggested ways on the objective and reliable methods of testing, the structuralists identified elements of language to be tested. Language testing in this era was dominated by criteria for the establishment of educational measuring instruments developed within the traditions of psychometrics (Alderson & Hughes, 1981:05).

The first person to claim the need was Robert Lado, who was responsible for the discrete point approach. According to this approach the target language was broken down into small testable segments using contrastive analysis. A test in each test item was intended to give information about the candidates’ ability to handle that particular point of language.

Morrow (1979:144-145) calls the approach of this era ‘atomistic’ due to the assumption that ‘knowledge of the elements of a language is equivalent to knowledge of the language’. Though this approach provided easily
quantifiable data for testing, it carried a major drawback as pointed out by Morrow (1979). He stated that the knowledge of discrete elements is worthless unless the user can synthesize those elements according to the linguistic demands of the situation. Oller (1979:212) rightly claims that “the whole is greater than the sum of its parts …”

iii) The Psycholinguistic-Sociolinguistic Era: This era began with the emergence of the global integrative testing. Oller’s (1979, cited in Miyata-Boddy & Langham, 2000:76) argument on the utility of global integrative testing provided a closer measure of the ability to combine language skills in the way they are used for actual language use rather than discrete point testing. However, even this view has been refuted by Bachman (1990), Alderson (1981) and Morrow (1979). This era believed that “the whole is greater than the sum of its parts …” (Weir, 1990, in Miyata-Boddy & Langham, 2000:76).

About the developments in this era, Wesche (1983:41) says that “in the past decades or so, linguistic science has paid increasing attention to the communicative – or message transmission – aspect of language as opposed to an earlier almost exclusive emphasis on grammatical forms and the ways in which they could be combined to form grammatical sentences. … More recently, attention has increasingly been given to how the communicative
ability of second language speakers can be tested.” More discussion on this aspect and the other aspects of this era will be done in the next chapter.

So far through the historical overview discussed above, an attempt has been made to survey under what circumstances, how and when the language tests have developed. But a discussion on historical overview remains unjustified, if the various kinds and approaches to tests and testing are not mentioned. Therefore, the following section of this chapter will deal with the types, approaches, Criteria, and test techniques language tests that have taken shape so far, since its infancy.

2.3. Types of Language Tests

Over the years along the research developments various kinds of tests have been formulated in different situations for different purposes. Some major ones are being mentioned discussed below:

2.3.1 Proficiency Test:

Proficiency tests are intended to measure one’s overall, general ability in a language in order to assess whether the learners are proficient enough to handle the language in a given situation. It is not content or objective specific. These tests are conducted, for example, to test whether a student’s English is good enough to follow a course of study at a British university.
That is, these tests take into account the level of proficiency and the kind of English needed to follow courses in particular subject areas. Cambridge examinations (First Certificate Examinations and Proficiency Examination) and the Oxford EFL examination (Preliminary and Higher) are a few examples of other types of Proficiency tests which are more general in their target. They try to establish whether candidates have reached a certain standard with respect to certain specified abilities.

There are other proficiency tests too, which do not have any occupation or course of study in mind. Here proficiency in a language is perceived in a more general term. Existing tests like the Cambridge First Certificate in English Examination (FCE) and the Cambridge Certificate of Proficiency in English Examination (CPE) are appropriate examples for this type of test. The function of such tests is to show whether students have reached a certain standard with respect to a set of specified abilities. The examining bodies of such tests are independent of teaching institutions. The FCE and CPE, as mentioned above, are linked to levels in the ALTE (Association of Language Testers in Europe) framework, which draws heavily on the work of Council of Europe.

2.3.2 Placement Test:
These tests are conducted at the beginning of a course of study in order to group the learners, for instance, to elementary, intermediate or advanced levels. Placements tests are used to find the initial level of proficiency of learners who have enrolled for a particular course. These tests are generally situation based. A Placement test, therefore, designed and conducted at one institution cannot be used in totality at another institution. This test is generally tailor-made, an in-house product. Well et al (1994) and Fulcher (1997) provide a detailed discussion on the issues in placement test development.

2.3.3 Progress Test:

These tests are also termed as Mid term or Mid year tests because they are conducted during the continuation of a course of study. Such tests help the teachers and learners in assessing how far they have been able to meet the teaching and learning targets, respectively. The results of these tests help in modifying the teaching methodology or materials, if there is a need for it. Some institutions conduct two or three progress tests spread over one year. Ideally tests should reflect a clear progression towards the final achievement based on course objectives.

2.3.4 Achievement Test:
The Progress test can also be understood in terms of ‘Achievement’ test. This test is closely related to the curriculum and they establish exactly what a learner has learned in a given teaching context. That is why; most teachers are involved in either framing or evaluating this test. The achievement test can be categorized into two types: ‘Final Achievement Test’ and ‘Progress Achievement Test’.

‘Final Achievement Tests’ are actually final tests conducted at the end of a semester or an academic year after course completion. The content of this test is related to the content of the course as specified in the syllabus. That is why it is referred to as the syllabus-content approach. The major disadvantage of this test is that if the course and materials are not designed properly, the results will have a negative wash back. That is, it will not indicate successful achievement of course objectives. However, the results of these tests help the teachers in correcting or modifying their strategies or materials only to be used in the next academic session. The feedback is not given to the learners who take this test. The amendments felt to be made are experimented on the fresh learners who join the course.

There is an alternative approach too in this test, where the test is directed more on the course objective rather than on the course content. This helps in
making the course designers explicit about objectives; hence this test is able to capture the learners’ achievements in terms of these objectives.

‘Progress Achievement Test’ is practically synonymous with ‘Progress Test’, as discussed in brief above. This test is intended to measure the consistent progress of learners during their academic session. Even here it is intended to make the test focused on the short term course objectives for a better performance and positive wash back. This is because low scores obtained prove to be discouraging for students. Such tests as sessionals, tutorials, quizzes, weekly assignments, and weekly tests are examples of this type of test.

2.3.5 Diagnostic Tests:

Diagnostic tests are meant to identify the learners’ strengths and weaknesses. Unlike achievement tests, diagnostic tests not only provide an impression of the level of proficiency, but they also establish what a learner has or has not mastered. These tests provide data which can be used to adapt a teaching programme through its follow-up techniques. They intend to establish what further teaching is required. Such tests are extremely useful for individualized instruction or self-instruction. This test can be conducted at the beginning or during the course to determine proficiency related to vocabulary, grammatical accuracy, or linguistic appropriacy. But conducting
and administering such a test is difficult from various practical points of views.

Besides, not many diagnostic tests have been designed so far. The availability of computers have come up as a help, and we have witnessed the development of a web-based DIALANG, which offers diagnostic test in fourteen European languages, each having five modules: reading, writing, listening, grammatical structures, and vocabulary.

The above types of tests reflect the very purpose and use of tests in actual practice. Over decades of language testing, the following sets / binaries of approaches to language testing emerged, based on which the tests were prepared:

i) **Direct versus Indirect Testing**

The ‘direct’ versus ‘indirect’ continuum refers to the degree to which a test task approaches the actual criterion performance. As the term suggests, ‘Direct’ testing means a direct assessment of the learners in the skills concerned. That is, if we want to test the ability of composition among the learners, we get them to produce a composition or if we have to evaluate the pronunciation, we ask them to speak. In other words, a ‘direct test’ requires the candidate to perform the skill that is being tested. Direct test is generally
better conducted on the productive skills, like reading and writing, in an authentic situation. Though the tests cannot be called to be authentic situation in the real sense of the term, still the situation can be brought very close to real life. The example of a direct test is a continuous piece of writing within a limited time frame, with an unknown topic.

In this approach the positive points are that we are clear about the test target to be assessed; hence it becomes easier to develop the authentic situation, to assess and interpret learners’ performance, and to get a positive backwash effect.

Unlike the ‘direct test’, the ‘Indirect test’ attempts to measure the abilities which underlie the skills in which we are interested (Hughes 1996:15) like the multiple choice tests of grammar and usage. One good example of indirect testing is Lado’s (1961) paper – pencil test for testing pronunciation ability through the identification of rhyming pairs of words. Another example comes from grammar based test indirectly measuring writing ability, for instance in TOEFL and IELTS, where learners are required to identify erroneous or inappropriate expressions in formal Standard English. The problem with indirect testing is with regard to construct and content validity whether the test reflects what is being tested and what is actually being covered in the test.
This approach to testing lacks in conveying the aim of the test hence creates problems at the level of marking and evaluation, especially when the evaluators / assessors are different from the test designer. Here learners’ performance on the test itself does not confirm the learners’ better performance in the overall skill itself. That is, this approach to testing does not necessarily promise a positive backwash effect. However, the ‘indirect’ testing is best suited for the diagnostic purposes.

Today ‘direct’ testing has attained an upper hand over ‘indirect’ testing. It is important to mention here the existence of ‘Semi-Direct’ testing too. One appropriate example is the ‘speaking test’ based on tape-recorded stimuli, to which learners have to respond.

**ii) Discrete point versus Integrative Testing**

The term ‘discrete point’ refers to the testing of one element at a time. That means one needs to construct separate tests for various language items. Opposed to this, the ‘integrative’ tests combine various aspects of language in one wholesome / integrated test. We can take, for example, the case of writing a composition where the form, content, use of vocabulary, proper use of grammar, etc can be checked.
Discrete point testing (such as diagnostic tests of grammar items) is generally ‘indirect’. It has traditionally been referred to as tests of grammar in which single sentences are the maximum stimulus unit presented. These tests also use separate tasks for or subsets to assess the ‘four skills’. In other words, this type of test selects particular levels of language code organization (e.g. phonology, syntax) to test grammatical points (e.g. speech sound discrimination, correct choice of preposition), while dealing with only one skill (reading, writing, listening or speaking).

Integrative testing, with an exception of the cloze procedures, is basically ‘direct’ in nature. It involves higher level of language organization in a test; hence it comes to be integrative by involving language in discourse, rather than at the word, sentence or syllable level. Such tests integrate various skills for assessment in one activity. These tests evaluate language proficiency of learners keeping in mind their communicative competence and performance. Studies by Carroll (1961) and Oller (1979) are remarkable with regard to this pair of approach to language testing.

iii) Norm-referenced versus Criterion-referenced Testing

‘Norm-referenced’ testing is meant to evaluate and place the learners in terms of other learners. That is, it tells us the learners’ ranking (like top ten or bottom five students) as conducted by TOEFL, IELTS, or The Michigan
Test of English Language Proficiency. It intends to make comparisons between individuals. It is, for instance, necessary in a situation, where a fixed quota of the best achievers, has to be selected.

The ‘criterion-referenced’ testing, on the other hand, intends to tell what a learner can actually do in the target language. That is, it suggests how proficient a learner is in performing certain skills / functions through the target language. This test does not rank-order testees. The simple aim is to find out whether a learner has mastered a set of specific objectives. Here the tasks are generally set and the performances are evaluated. That means learners are supposed to measure their progress in relation to meaningful criteria. Tests conducted by Interagency Language Roundtable (ILR) and Berkshire Certificate of Proficiency in German Level 1, for instance, have developed certain criteria for the language skills to be reflected for being determined as proficient in the respective skills. Some other examples are ACTFL Oral Proficiency Interview, the FBI Listening Summary Translation Exam (Scott et al, 1996), the Canadian Academic English Language (CAEL) Assessment (Jennings et al, 1999).

A meaningful discussion of this binary of approaches can be found in the studies by Popham (1978), Ebel (1978), Skehan (1984), Hudson and Lynch (1984), Hughes (1986), and Brown and Hudson (2000).
iv) **Objective testing versus Subjective testing**

The main distinction between the types of testing is in terms of methods of scoring. If a test provides a scoring scale to be followed by the scorer, the test is claimed to be ‘objective’, otherwise it is ‘subjective’. The ‘subjective’ testing or scoring is actually a sort of ‘impressionistic’ scoring, which varies from the evaluation of a composition task to the scoring of short answers in response to a reading comprehension passage.

Objectivity in testing is given priority for better reliability and greater agreement between various scores / evaluators of the same test. This means, it is better if a test is constructed with its specifications and scoring scale. However, in EFL and ESL countries, in general, such things exist only in principle.

v) **Computer adaptive testing**

Since the last quarter of the 21st century, a consistent effort (successful in some instances) is going on to develop computer adaptive language tests. This type of testing mainly helps in collecting information on learners’ ability. It also helps in conducting and aligning the difficulty levels of tests as per the ability of the learners. That is, to begin with all learners are given an average level test, where the learners’ abilities are identified and they are
further tested accordingly. For instance, those who respond correctly (in the introductory average-level test) are presented with more difficult items, while those who respond incorrectly are presented with easier items.

With the introduction of such practices as e-learning, e-content generation and on-line teaching and testing, the tests are required to be adapted as per the new challenges of testing. Chalhoub-Deville (1999), Chalhoub-Deville and Deville (1999), and Fulcher (2000) discuss issues relating to the role of computers adaptation for language testing.

vi) Communicative language testing

The next chapter will deal with a detailed description of this approach. In a nutshell, this approach actually came up as a reaction almost revolutionizing the earlier practices with regard to the matter and manner of the language tests. It suggests the need to test learners’ ability to communicate. In order to do so, the test designers had to replace the aspects of grammatical items and linguistic competence by communicative grammar, functions and notions of the target language and the communicative competence. Although it has its own limitations, it had a considerable influence. Such studies as Morrow (1979), Brumfit and Johnson (1979), Canale and Swain (1980), Alderson and Hughes (1981, part 1), Hughes and Porter (1983), Davies (1988), and
Weir (1990) are most representative contributions with regard to Communicative language testing.

2.4 **Characteristics of language tests:**

So far we saw that language testing over the last one century has been approached differently and has produced various test types. These tests carried certain typical features of their own. But they all were expected to carry the following three characteristics: namely, validity, reliability, and practicality. That is, any test that we use must be appropriate in terms of objectives, dependable in the evidence it provides, and applicable to our particular situation. The development of the above concepts is a typical feature of the ‘psychometric’ phase that emerged after the ‘pre-scientific’ days.

i) **Validity:**

If a test is found to be based upon a sound analysis of the skill(s) intended to be measured, and if there is sufficient evidence that test scores correlate fairly highly with actual ability in the skills area being tested, then we may call that test to be valid. In simpler words, a test is said to be valid if it measures accurately what is intended to measure.
By definition, though the concept of validity seems to be quite simple, the researchers have often found it complex and difficult to separate it from the concept of reliability. General discussion on validity and ways of measuring it can be studied in Messick (1989), Bradshaw (1990), Wall et al (1994), Cumming and Berwick (1996), Fulcher (1997), Anastasi and Urbina (1997), and Nitko (2001).

In recent times, the term ‘construct validity’ has been increasingly used in place of the general notion of validity. This is so because language tests are generally created in order to measure some theoretical constructs in the form of language items like reading ability, fluency in speaking, accuracy in writing, written composition, control of grammar and so on. This term was first used in the context of psychological tests of personality to measure psychological constructs. Bachman and Palmer (1981) is one of the early attempts to introduce construct validity to language testing. Crippen and Davies (1988) and Hughes, Porter and Weir (1988) made further significant contributions on these lines.

However, it is not enough to say that a test needs to have construct validity. Rather it is required to reflect the construct validity in the form of some evidences, like ‘Content validity’, ‘Criterion-related validity’, and ‘face validity’.
The notion of validity is reflected in the following forms:

**a) Content Validity:** If a test is designed to measure mastery of a specific skill or the content of a particular course of study, it is expected that the test is based upon a careful analysis of that skill or the outline of the course. That is, a test must take into consideration the language item (a skill or some grammatical structure), the aspects of this skill to be tested, and its extent / proportion as per the level of the learners (intermediate level or advanced level, for instance). A mere inclusion of a language item from a course of study is not enough to evidence the ‘content validity’. In other words, it is important for the test designers to consult the test specification, if at all it exists. The availability of ‘Content validity’ can be judged on the basis of the proximity of ‘test specification’ and ‘test content’. ‘Content validity’ is directly proportionate to the backwash effect and the accurate measure of what the test is intended to measure. The higher is the ‘content validity’, the more accurate (higher construct validity) will be the test and the backwash effect.

It has been often observed that the test designers ignore the ‘content validity’ and give preference to simpler and easier content over the important content.

**b) Criterion-related validity:** This evidence of construct validity is often termed as ‘empirical validity’, because here the correlation between
the test scores is established by comparing them to independent, outside criterion such as marks attained at the end of a course or instructors’ or supervisors’ ratings. If there is a high correlation between the two sets of test scores, that means the ‘criterion-related validity’ is also high. This validity is of two types:

- **Concurrent validity:** It is established when both the test and the criterion are administered at about the same time. This can be exemplified through the ratings of the two types (in terms of duration, coverage, or mode, etc) of achievement tests for the same content. If the comparison between the two tests shows a higher agreement, then both the test are considered to carry higher concurrent validity.

- **Predictive Validity:** This refers to the degree to which a test can predict a candidates’ future performance. This can be well understood in terms of a proficiency test, which is given to predict a learner’s ability to take up a graduate course at a British or American university.

c) **Face Validity:** A test is considered to carry ‘face validity’ if it is confirmed that it measures what it is supposed to measure. A pen-pencil test to measure a learner’s phonetic / pronunciation ability, for example, lacks in ‘face validity’. Face validity actually depends on the content and criterion-
related validities. That is if a test lacks in content, it automatically promises to lack in face validity.

ii) **Reliability:** The term ‘reliability’ refers to the stability of test scores. That is, a test is considered to be reliable only if the test scores do not vary significantly due to the change in such external factors like day, date and place, or due to the evaluation conducted by two or more sets of independent examiners, or even due to the two parallel forms of the same test.

In order to attain reliability one needs to ‘construct’, ‘administer’, and ‘score’ tests in such a way that the scores obtained on a test on a particular occasion are similar to the scores obtained by the same students with the same ability, but at a different time. The more similar the scores, the more reliable the test is supposed to be. Reliability is perceived in terms of ‘Test Reliability’ and ‘Scorer Reliability’.

- **Test Reliability:** Test reliability is attained by conducting sampling of tasks on varieties of students. Sometimes test reliability is affected by the change in administration or the lack of motivation among students.

- **Scorer Reliability:** It concerns the stability or consistency with which test performances are evaluated. Scorer reliability is often affected in subjective tests or impressionistic evaluations on such test items like
composition task, or free response tests, etc. To achieve this reliability, an attempt is often made to develop a scoring scale to be followed by all evaluators. Scorer reliability is nearly perfect in the case of multiple choice test, if the options to the test are carefully drawn.

Studies by Lado (1961), Feldt and Brennan (1989), Anastasi and Urbina (1997), Nitko (2001), and Brown and Hudson (2002) are some representative contributions with regard to reliability of tests.

iii) **Practicality:** A test may be very valid and reliable and yet may not be practicable in its application. The first and foremost aspect is the cost effectiveness of the test. Besides, economy of time is also significant. Aspect of time is to be managed keeping in mind the level of students. Hence this is also related to the difficulty level of the test, length and breath of the task. In addition, ease of administration and scoring also becomes important for practicality of the test.

2.5 **Summing up:**

The present chapter has mainly tried to capture the advancements made in the field of language testing over the last one century. In the process it also takes into consideration its overall development in terms of its types, approaches and characteristics.
References:


Brinsmade, C. 1928. Concerning the College Board examinations in Modern languages. MLJ, 8, 87-100, 212-227.


Chedyleur, F. D. 1928. Results and significance of the new type of modern language tests. MLJ, 12, 513-531.

Chedyleur, F. D. 1931. The use of placement tests in modern languages at the university of Wisconsin. MLJ, 15, 262-280.


Clapp, H. L. 1947. Meditations on a placement programme or when should a foreign language be studied?. *MLJ*, 31, 203-207


Clarke, F. M. 1931. Results of the Bryn Mawr Test in French administered in New York City high schools.*Bulletins of High Points*. 13, 4-13.


Matheus, J. F. 1937. Corelation between psychological test scores, language test scores, and semester grades. MLJ, 22, 104-106.


Michel, S. V. 1936. Prognosis in German. MLJ. 20, 275-287.

Miller, M. M. 1948. Test on Spanish and Spanish American Life and Culture. MLJ, 32, 140-144.


