Avrasya Dil Eğitimi ve Araştırmalar Dergisi

Dergi Web sayfası: http://dergipark.gov.tr/ader

# LANGUAGE ASSESSMENT: NOW AND THEN

Pınar Uyanıker*

*İngilizce Öğretim Elemanı, Milli Savunma Üniversitesi

*Gönderilme Tarihi: 23.01.2017*
*Kabul Tarihi: 14.08.2017*

**Abstract:** This article is a historical overview of English language testing. It starts with a history of testing and offers a discussion of changing trends that have occurred during the last 25 years starting from communicative language testing to learner oriented language testing. In this article, changing trends in psychometric principles will also be discussed in-depth.

**Keywords:** language assessment, history, trends in assessment

## Introduction

The educational reforms since the second half of the last (20th) century and their striking influences on the theories and principles of teaching and learning brought about a movement in assessment paradigms. Developments in language testing research in the past 25 years have brought language testers into closer contact with applied linguists, as well as with measurement specialists. The blossoming of language testing research provided us with a rich variety of research approaches and tools, at the same time broadening the research questions that are being investigated (Hamidi, 2010). Recent trends also have led to a more contextualized, communicative, and authentic assessment.

This revolution, owing to social constructivist framework, proposed to move toward helping learners to make their own decisions in learning.

**A Brief History of Assessment**

To understand how language assessment has evolved, it is important to take a historical view of it. Spolsky identified three periods of language testing; pre-scientific, psychometric-structuralist, and integrative sociolinguistic periods (1978). These periods are mainly distinguished from each other in terms of how language is defined. Each will be mentioned briefly.

Tests are ancient practices. Language tests, in particular, are as old as language teaching itself.So, language tests are an integral part of teaching. Naturally, trends in language testing follow trends in language teaching (Giri, 2010).The relation between language teaching and assessment of language can be explained as follows; test developers need to base language tests on a theory of language proficiency. As it will be mentioned, there are different ways of looking at language and therefore different definitions of language. As our understanding and explaining of language differs, so do the way test developers measure it.

The pre-scientific period of language testing refers to the time before standardization. In this period, language experts were considered to be testing experts. These experts made the decisions about teaching and testing but the decisions which the experts made were intuitive and based on personal judgment. One important characteristics of this period is that there was no concern for reliability or validity (Madsen, 1983). Although the earliest formal testing is known to have begun in China known as Imperial Examination about 1500 years ago (Spolsky, 1978), language testing in the West began with the adoption of English as a royal language during the reign of Henry V in the 15th century. It was not until 17th century when large numbers of people immigrated to Britain that testing became widespread. With the publications of Fick and Miller, language teaching and naturally testing took a significant turn. Public examinations were carried out by universities (Giri,2003). This can be considered as a first step towards standardization. As mentioned earlier, language testing is not independent from the way language is taught. The prominent language teaching method in the pre-scientific era was grammar translation method. So, tests included translation, composition writing tasks which aimed at measuring testers' knowledge about language. One advantage of language testing is that it allowed a global evaluation of the learners' ability in the target language (Giri, 2003).

In the psychometric-structuralist period, language was defined by structural and behavioristic theories of language. Language ability is seen as the ability to handle discrete elements of the language and develop language skills. According to these theories, language can be broken into its components (phonemes, morphemes, and sentences) (Farhady, 1997). Discrete point analysis breaks the elements of language apart and tries to test them separately (Oller, 1979). So, testing aimed at measuring language through discrete elements known as "discrete point testing". The main concern was psychometric reliability (Bachman, 2000). Therefore, tests included multiple-choice items to retain reliability. The advantages of "discrete-point" testing

are easy quantification of results, wide coverage of items, and objective scoring. However, this type of testing undermined context of communication in which language is used. In other words, it can be suggested that measuring "discrete" elements of language may not give testers reliable information about language use because it might be misleading to make generalizations about testee's language use based on a test through which language is tested separately. Similarly, Farhady cautions against content validity problem in discrete point testing because test items are not adequate (2014).He also highlights the complexity of language: "the use of language relates to sociolinguistics, its changing nature to linguistics, its acquisition to psycholinguistics, and the interpretation aspect to discourse and pragmatics. That is why a comprehensive treatment of language through one single dimension is neither easy nor acceptable" (Farhady, 1999).

The integrative sociolinguistic period emerged as a reaction to the psychometric-structuralist period, and"discrete-point" testing. First, the integrative aspect of this period will be discussed. With Chomsky's definition of language, language learning was seen as a process of acquiring conscious control and understanding of language systems. This definition highlighted language as an interactive phenomenon(Farhady, 1997). Test developers rather than focusing on accuracy, tried to measure functional ability (Giri, 2003). In testing, cloze tests, dictation, and oral interviews became popular in assessment (Farhady, 1997). In language testing research, Oller was a prominent figure who proposed Unitary Competence Hypothesis. According to Oller, there is a single unitary factor that underlies language proficiency and four skills are closely interrelated. As language is seen as indivisible, then, language tests need to be integrative. That is, language tests need to consider the relation between language elements. Soon this would be proven wrong; Oller used test analysis to explain his model but Farhady applied factor analysis to examine the components of language proficiency and reached a different conclusion (Stansfield, 2007). He showed that correlation analysis was not as strong as Oller suggested because discrete items have equally high correlations as integrative ones (Skehan, 1988). So, it may be suggested that the distinction between discrete and integrative tests might be a "false assumption". Farhady also points out that integrative tests are problematic in reliability because item independency is violated (2014).Cloze tests are important in integrative testing because the ability to supply appropriate words in the blanks requires grammar knowledge, knowledge of vocabulary, discourse structure, reading skills and strategies and internalized expectancy grammar. Dictation tests are also useful as they correlate strongly with other tests of proficiency. Carrol similarly points out that integrative tests are more valid than discrete-point tests (Carrol, 1986) Although these tasks do not test learners' communicative ability, they are better guide to learners' aptitude and potential communicative ability (Giri, 2003). Furthermore, they are economical to set and mark and have a respectable degree of reliability. But Carrol cautions about the problem in developing effective integrative tests because there is still little known about factors in language production (1986).

The sociolinguistic aspect of this period lies in the work of Hymes who underlined the importance of context in communication. According to Hymes, missing socio-cultural elements may cause misunderstanding or communication breakdown. Following Hymes, a number of studies were published on learners' communicative needs. So far as language testing is concerned, revealing learners' abilities in communicative settings gained importance. These works eventually gave way to communicative language testing (Bachman, 1990).

**Communicative Language Teaching &Testing**

Although communicative language testing is presented as a separate title here, it falls into the category of the integrative sociolinguistic period (Giri, 2003).It deserves a separate title because what it brings to language teaching and assessment is important. In the past twenty years, language assessment has evolved to a great extent. At the first Language Testing Research Colloquium, unitary trait hypothesis received criticism and a broadened view of language and language definition was put forward based on the works of Canale& Swain, Widdowson (1983) and Savignon (1983). With the introduction of "communicative competence", language testing entered a new phase. These works highlighted multi-componential and dynamic nature of language involving discoursal and sociolinguistic aspects.One of the commonly recognized models is by Bachman and Palmer in 1996. What made their model stand out among other models is the skill and method factors which means the model places competence in a wider performance framework (Skehan, 1988). However, as Farhady argues, communicative competence is complex and vast in domain. He further suggests that, communicative competence compromises many functional competences within specific areas of language use which may develop by educational and professional careers (1983). Although these models can be considered revealing, except for Bachman's model, none of these models included measurement aspect (Farhady, 2005). Nevertheless, there are some implications of "communicative approach to testing; test developers discussed about the nature of communicative language tests as well as sociolinguistic aspects. Using authentic materials in reading and listening tests, requirement of language production appropriate to specified purposes (using language for specific purposes such as giving directions to the airport) , and recognition that a test can yield valid results without inclusion of components such as grammar and vocabulary are some features that have originated from communicative language testing (Bachman, 2000).

It can be suggested that tests of "language" have become tests of "reading, speaking, writing, and listening". As the definition of competence is too complex some issues are still on debate; one of the most important questions is how do we achieve a single measure of proficiency which leads us to the question of how do testers ensure reliability and validity? Last but not the least, how do we disentangle other variables (culture, mood)? Another point worth mentioning here is about the aforementioned

issue on "specified purpose". How many purposes are there and how can we make generalization from students' performance on a particular task?(Coombe et.al,, 2012).Fulcher similarly criticized communicative language testing on the grounds that "using content validity as a major criterion in test design and evaluation has been mistaken" (1999). These questions indeed are questions of validity which will be discussed in the upcoming section of this paper.

***Assessment in Cross-Cultural Pragmatics***

Attempts of defining language competence and the importance of context gave way to Cross-cultural pragmatics. Contemporary models of language proficiency agree that language involves linguistic, sociolinguistic, pragmatic and strategic competencies (Bachman, 1990). There are methods for measuring grammatical and textual competence (syntax, vocabulary, cohesion, etc.), but no generally accepted measures of cross-cultural communicative ability (Hudson, Detmer& Brown, 1992). Drawing on research in linguistic pragmatics, SLA and sociolinguistics, Bachman notes the advances in assessing cross-cultural pragmatics; Assessment procedures focused on varieties in the social properties in the speech event and on variability due to the particular types of data collection procedures and associated instruments'.The researchers included a wide variety of task types – including multiple-choice or cued response items, structured oral interviews, self-assessments and direct observations – as part of the assessment procedure (Bachman, 2000). But more research is called for in assessment of cross-cultural pragmatics.

***ESP***

Before moving on to validity and reliability discussions, it is worth mentioning ESP and its assessment. As mentioned earlier, when communicative language teaching emerged, tests were developed to meet the requirements of communicative competence. Tasks became authentic and include real-life activities and centered on contextuality (contex-dependent), productivity (requires production of the language), and interactivity (Farhady, 2005). From these growing trends, and the focus on learners' needs, ESP has emerged. According to Hutchinson & Waters, if the language varies from one situation to another, determining the features of specific situations and making these features the basis of learners' course is feasible (1987). Actually, ESP is not different from other forms of language teaching but the content of learning may vary (Hutchinson & Waters, 1987). And this content determines what to assess. One of the principles of ESP testing is that test tasks mirror candidates' target language use situation, so content and test methods are more narrowly defined (Tratnik, 2008). Figure 1 shows the stages in ESP process;

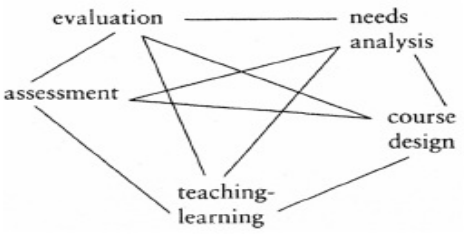


Figure 1.Stages in ESP process. Source: Dudley – Evans & St. John, 1998, p. 121.

ESP assessment process starts with need analysis. Students' needs for that matter, shape course design and materials, the analysis of target language use (in analysis of language use, there has been different approaches; in 1960's the analysis focused on field-related words, then syntactic analysis become prominent followed by discourse analysis in 1980's and as in the aftermath of 1980's genre analysis were on the focus of ESP in analysis of target language use) facilitated selection of suitable test tasks which will result in valid assessment.

**Validity**

Traditionally, validity is defined as "the extent to which a test measures what it is supposed to measure" and it is related to content and form of the test (Lado, 1961). Validity is considered as being the most important characteristic of a test (Bachman & Palmer, 1996). In this part of the paper, the aim is to see how the term "validity" has evolved during the past 25 years. Traditional view of validity has three components;

***a. Criterion-oriented validity***

In criterion-oriented validity, the tester is interested in the relationship between a test and criterion to which predictions will be made. To illustrate, we may want to make predictions from scores of L2 academic reading ability to see whether test-takers can read undergraduate business texts. If the test predicts test-takers reading ability, then it is said to have criterion- oriented validity (Fulcher& Davidson, 2007).

***b. Content Validity***

Content validity is defined whether the content of the test is a representative sample of the domain to be tested (Fulcher& Davidson, 2007). If we extend our example in criterion-oriented validity, we can suggest that the test has content validity if the texts used in L2 academic reading test are typical of the texts used in first-year undergraduate business course.

***c. Construct Validity***

Farhady defines construct validity as an underlying structure which investigates whether the test measures the predefined ability. However, our construct, which is the language, as an abstract trait makes it difficult to measure directly. To achieve construct validity, predefined ability should be "measurable" or "operational". In other words, we should measure language by something observable (1997).

The above characteristics of validity are seen as distinct and independent in traditional view of validity. The way linguists define language and the way teaching of language has changed caused changes in the way we assess learners. These changes made it necessary to review test qualities;

Oller, in 1979, stated that "reliability and validity are bound together and tests must range over a variety of situations to achieve validity, and even then there is no assurance that language elements are adequately sampled" (p.240).

Following Oller, Davies presented a scheme for determining validity and listed five types of validity; face, content, construct, predictive and concurrent (1968).

According to Henning, "even an ideal test which is perfectly reliable and possessing perfect criterion-related validity will be invalid for some purposes"(1987) highlighting that there is not a clear-cut distinction between validity and reliability.

In 1989, Messick sat out a "unified validity framework". He defined validity not as a test property of the test but "the degree to which we are justified in making an inference to a construct from a test score". He highlights content aspect (relevance and appropriateness of the content of the test), the substantive aspect (empirical evidence to the content appropriacy of the test), the structural aspect(relationship between the scoring system and the internal structure of the domain being tested), the generalizability aspect(scope of the interpretation of the scores), the external aspect (the relationship between outside criteria and the test in question), and the consequential aspect of validity (the evaluation of the consequences of the test results on the test takers) (Farhady, 1997;Messick, 1989). Bachman explains the rationale behind this unified validity framework by pointing out the use of tests; validity cannot be limited to collecting factual evidence to support a given interpretation or use because testing takes place in social context (and the interpretation as well as the different use of tests may not be equally valid for all context and abilities) and it is necessary to consider the educational and social consequences of the test (Bachman, 1990). So, it can be suggested that by the term validity, what counts for validity is not the test content or the scores but the way scores are interpreted (Bachman, 1990).

This paradigm shift in language testing is welcomed but there are some criticisms and doubts about its implementation; Popham states that adding such social consequences to validity makes the term more complex, he suggests that social consequences of tests should be systematically addressed apart from validity. Another

concern for “unified validity framework” is about its implementation; Popham thinks this view of validity might not be feasible for educators (1997).

More recently, Bachman& Palmer brought upon the term “test usefulness” through which test development and use can be evaluated. Test usefulness involves six qualities; reliability (consistency of scores), construct validity(meaningfulness and appropriacy of test score interpretations as well as generalization), authenticity(the degree of correspondence of the characteristics of a given language task to the features of task in real-life), interactiveness (the extent of test takers’ involvement in accomplishing the task type), impact (on individual and system) and practicality (required resources to develop a test that has necessary qualities) (1996). It was suggested that “test usefulness” is important in that it ties notion of usefulnessto specific testing situations and it provides a principled basis for importance of all qualities (Bachman& Palmer, 1996).

### *d. Construct Irrelevant Factors*

There have been a number of identified variables that are assumed to have no direct relation to language ability but may influence test taker performance and may alter the interpretations of assessment outcomes. Some of these factors are washback, ethics, bias, politicizations of the tests, standardization, and the power of the tests.These terms, however, may cause confusion because the context in which these concepts have been used, are not at all clearly identified in the field. For instance, fairness is discussed in terms of bias, and bias in terms of ethics, and both are considered immoral. (Farhady, 1999).Not surprisingly, moral problems of the late 20th Century caught up with applied linguists and language testers. The 19th Language Testing Research Colloquium in 1997 was held on the theme "Fairness in Language Testing", and issue 14, 3, 1997 of Language Testing was a special volume on ethics in language testing (Fulcher, 1999).

Exams are necessary and indispensable part of educationand have multiple effects on test takers, teachers, curriculum and teaching. The effect of testing (on individual, on teaching, and on society) is called washback. Washback has been discussed by many scholars (Alderson & Wall, 1993; Bachman & Palmer, 1990; Davies; Hamp- Lyons, 1997; Shohamy1997; Spolsky, 1981).When the studies are scrutinized, it can be suggested that the literature lacks a common deficiency in conceptualization of the matter. Some researchers prefer to discuss the effects of testing in relation to validity, like Messick who points out that washback is one form of testing consequence and needs to be weighed in consequential aspect of construct validity(Messick,1996)or like Willingham & Cole who suggest that anything that reduces fairness also reduces validity(1997)and Anderson & Wall who see washback as a “neutral” term which might have positive and negative effects and propose that validity should be measured by backwash (backwash validity) (1993), whereas Bailey takes a difference stance and points out positive washback as a key difference between Communicative Language Teaching and traditional language tests. He suggests that for a positive washback, there should not be differences between learning activities and test tasks,

there should be detailed score reporting and test authenticity (Bailey, 1996; Hamps-Lyons, 1997). For future studies, Hamp-Lyons suggests that there is a tendency to move towards a more complex model in explaining effects of testing (1997).

The debate on washback has not been over and seems to last for a longer period time.More research might therefore help to see this issue in a more clear way and the deepening of the issue might help us develop a logical model for washback.

As mentioned earlier, washback is the effect of testing on individuals and society. If the effects of a test show systematic differences in test performance which is associated with characteristics not related to the ability in question, there is a possibility of bias. Biasis a complex topic;it may include misinterpretation of the scores, sexist or racist content, and inappropriate selection procedures.Differential Item Functioning helps to detect bias (though not in classroom tests). DIF is a Mantel-Haenszel approach which is a chi-squared contingency table that examines differences between the reference and focal groups on all items of the test one by one (Bachman, 1990). IRT can also be used for DIF. However, DIF has certain drawbacks; a single DIF study may answer the question of item bias for certain groups, but not be able to answer questions regarding other group differences.

Early approaches to fairness were evaluated through validity and reliability concepts. The focus of this concern is on whether test-score interpretations have *equal construct validity* (and reliability) for different test-taker groups such as gender, race/ethnicity, field ofspecialization and native language and culture (Kunnan, 2000). Recently the term “fairness” has been used in terms of equity which goes beyond validity.
A definition of fairness is stated by Jensen:

“to the ways in which test scores (whether of biased or unbiased tests) are used in any selection situation. The concepts of fairness, social justice, and equal protection of the laws are moral, legal, and philosophical ideas and therefore must be evaluated in these terms.”(Jensen 1980: 376).

Kunnan states fairness indicates a multi-disciplinary concept; not only based on psychometric view of tests but also on social, ethical, legal and philosophical views (2000). His“Test FairnessFramework” starts with thinking stage of test development and carried out in writing, piloting, analyzing and research stages and therefore can be considered a detailed and rich one. He concludes his paper by new methodologies that can contribute to fairness like item level exploratory and confirmatory factor analysis, structural equation modeling,Multidimensional Item Response Theory for DIF, Rule Space and verbal protocol analysis (2000).

**Reliability**

According to Henning, “reliability isa measure of accuracy, consistency, dependability, or fairness of scores resulting from the administration of a particular examination” (1987). According to Bachman, reliability is concerned with answering the question “How much of an individual’s test performance is due to measurement error, or to

factors other than language ability we want to measure?" (1990, p. 160). Bachman rather than seeing validity and reliability as distinct concepts,recognizes them as complementary which will enable us to identify, estimate, and control factors which affect test scores. Reliability and validity have two complementary objectives; 1. Minimizing measurement error, 2. Maximizing effects of ability we want to measure. However important it is to achieve these, it is not an easy task. So, what are these factors that affect reliability? Bachman (1990) identifies these as characteristics of test methods (such as authenticity, context of language tests) and individual attributes (such as cognitive/affective characteristics of test takers, sex, L1, socio-economic background of test takers). Farhady categorizes factors that affect reliability as; environment (ex.lighting, ventilation) administrative procedures (ex. directions) examinees (ex. fatigue, health, vision) scoring procedures (human errors, variance in judgments) test and test items (unfamiliar format, smudged booklets). (2014) Empirical research at this point helps us to estimate reliability.

There are four major theoretical approaches. Each will be mentioned briefly:

### *a. Classical Test Theory (CTT)*

In Classical Test theory, observed score (examinee's score) on a test comprises two factors or components: a true score that is due to an individual's levelof ability and an error score, that is due to factors other than theability being tested (Bachman, 1990). Error of measurement is calculated by discrepancy between an examinee's observed score and true score. In CTT, item difficulty is defined as proportion of examinees who answer an item correctly (p-value). But if the test is administered to a higher proficiency group, item difficulty would be different when it is administered to lower proficiency group. This problem will be resolved in Item Response Theory as will be discussed.

### *b. Generalizability Theory (GT)*

GT can be considered as an extension of CTT. In GT, an individual's performance on a test is generalized to her performance in other contexts. It is grounded in the framework of factorial design
and the analysis of variance(Bachman, 1990) . Generalizability theory estimates the components of variance in the error portion. The variance components depend upon the research design. G theory operates at two levels:

G study and D study. G study is analogous to pretest in CTT where the researchers try to identify potential sources of variation. D study is similar to the main administration in CTT when the parameters of variation are determined (Farhady, 2014).The application of G-theory thus enables test developers and test users to specify the different sources of variance that are of concern for a given test use, to estimate the relative importance of these different sources simultaneously, and to employ these estimates in the interpretation and use of test scores (Bachman, 1990).Although G

theory is a powerful extension of CTT, both have certain shortcomings in dealing with item and test characteristics as well as the comparability of test scores (Farhady,2014).

### *c. Item Response Theory*

In CTT, a test should be administered to everybody. Test A to person Aand test Bto person B would not be comparable whereas in IRT, it is possible to compare abilities of two persons using different tests by referring to small bank of common items or common persons. This is called test free person ability in IRT (Farhady, 2014). Another advantage of IRT is discussed by Bachman; assuming that a large number of items that measure the same trait, an individual's ability estimate is independent of the particular set of items that are taken. (Bachman, 1990).

The aim of these three approaches is similar but moving from CTT to IRT, we refine our explanations of the variance with more power. However, in classroom settings, CTT is commonly used due to limited knowledge of psychometrics and limited access to technology (Farhady,2014).

### *d. Structural Equation Modelling (SEM)*

SEM enables us to investigate both the factor structure of the measures we use and the relationships among these factors, or latent variables. Furthermore, SEM can be used to investigate directional relationships among sets of independent and dependent latent variables. SEM has been used in a wide range of studies, including the investigation of test takers' background characteristics and strategy use. Bachman predicts that recent work in latent trait approachesto generalizability theory promises to bring the technologies of G-theory, IRT and SEM together into a single analytic paradigm (Bachman, 2000).

## Alternative "s" (in) Assessment: Beyond Testing

When we look at the history and development of assessment, we can see that there are two paradigms; positivist and constructivist. The positivist paradigm sees assessment as a kind of scientific study; objective and independent of the context. It insists on standardizing testing. Individuals are ranked or compared for course of studies. However, the reality is not that simple. Learning is a complex phenomenon and there is a need of a more flexible measurement that can accommodate the complexity. The constructivist paradigm is a hermeneutic approach to assessment. It leads to an assessment that incorporates evaluation into assessment (Hamp-Lyons & Condon, 2012). Rather than ranking people, constructivist paradigm values progress. This paradigm shift manifests itself through different mediums. In the next section, samples of constructivist assessment will be discussed.

Alternative assessment can be considered as an alternative to traditional testing.Alternative assessment is different from traditional testing in that alternative assessment shows what students can do, how they can integrate and produce rather than recalling and producing. So, the main point is to gather information on how

students are approaching, processing and completing real-life tasks. (Huerta-Macias, 1999).
So, what makes a procedure alternative assessment is that; they require problem solving and higher level thinking, use real world contexts and focus on process as well as products (Ascbacher, 1991).

***Dynamic Assessment***

Dynamic assessment is a relatively new concept in assessment and it appliesVygotsky's sociocultural theory into assessment. Dynamic assessment is not a specific tool for assessment, rather an umbrella term which aims to find out how much learning can take place in the ZPD. In other words, it is a method of conducting a language assessment which seeks to identify the skills that an individual child possesses as well as their learning potential. Vygotsky's theory suggests that if we want to understand learning and development, we have to focus on process instead of product (Yıldırım, 2008). Lantolf and Thorne comment on the nature of dynamic assessment (2006, p.331):

"What makes a procedure dynamic or not is whether or not mediation is incorporated into the assessment process. In other words, fill-in-the-blank, multiple-choice, open-ended essay, or even oral proficiency tests in themselves may or may not be dynamic. Their status is determined by the goal of the procedure and the format in which it is subsequently administered. In other words, there are no dynamic assessment instruments per se; there are only dynamic assessment procedures".

**Standard Assessment vs. Dynamic Assessment (adapted from Pena)**

| | |
|---|---|
| Static | Dynamic |
| Content (What) | Process (How) |
| Norm Based | Test-Teach-Retest |
| Snap shot | Continuous |
| Passive Participation | Inter-Active Participation |
| Standardized Administration | Mediated Process |
| Absence of Feedback | Teaching Strategy |
| Deficit Based | Learning Strategies |

Figure 2..Standard Assessment vs Dynamic Assessment Characteristics, (Pena,2000) .

The figure compares the properties of Standard and dynamic assessments. It can be suggested that dynamic assessment is a process oriented therefore a continuous

procedure. This property of dynamic assessment enables teachers to have a more valid understanding of students' performances compared to single-shot assessments. Another point to highlight is on teaching strategy; strategies might be useful for revealing specific teaching strategies for different classrooms, which is more like a "tailored" teaching. Because dynamic assessment enables the assessment of cognitive processes, it can be seen that studies were mostly conducted with bilingual children, children with learning difficulties or diverse children (Botting, 2012; Clellen, 2001; Gorman, 2015; Gilliam, 1999; Pena, 2007; Bedore& Pena, 2008). These studies reported promising results.

There are two approaches to dynamic assessment: interventionist and interactionist. The main difference between these two is that interventionist approaches quantify performance as an "index of speed of learning" (Brown and Ferrara, 1985, p. 300) in terms of the amount of help required for a learner to quickly and efficiently reach a prespecified end point. Interactionists focus on ensuring individual development regardless of the effort required and without concern for the endpoint of development. In language assessment, studies on dynamic assessment is relatively new (Lantof&Poehner, 2008).

However, dynamic assessment can be considered appropriate for communicative language teaching. As mentioned earlier, recent communicative competence definitions and socio-linguistic theories of language underlines multidimensionality of the language rather than a single concept, that's why dynamic assessment can be used effectivelyin helping measure this aspect (Kantar &Özgür, 2012).

A type of commonly used dynamic assessment is test- teach-retest method through which a child's ability to learn after a predesigned learning opportunity is assessed (Kantar &Özgür, 2012). However, because of its theoretical assumptions on the nature of development, dynamic assessment can be criticized as a process which lacks validity, but according to Guterman, any assessment is valid when it is relevant to instruction and useful and beneficial to learners. In other words, if we take 'validity' and 'reliability'out of the context of standardized testing and look at the underlying meaning of these two concepts, we can see that they are both realized in the procedures of dynamic assessment (2002).Some researchers discuss that concerns of validity and reliability should be addressed by the term "trustworthiness"(Huerta-Macias, 1999). Büchel and Scharnhorst suggest that dynamic assessment researchers can link assessment and measurement through "standardization of the examiner–subject interaction," a characteristic of interventionist approaches to dynamic assessment(1993).

When dominant assessment methods are examined, it can be seen that majority are battery of standardized and norm referenced tests. Considering the recent developments in defining language and its aspects (multidimensionality), it can be suggested that dynamic assessment can be used for further information about processing or learning potential not for replacement of standardized tests.

**Learning Oriented Assessment**

In the previous part of this paper, basic considerations about dynamic assessment have been discussed. What dynamic assessment provides us mainly deals with what students can do and how they approach, process and complete the tasks. Traditional assessment, on the other hand, focuses on the summative aspect of assessment.

Learning-oriented assessment attempts to reconcile alternative and traditional assessments (Carless, 2009). As the name suggests, in learning-oriented assessment, learning comes first. It primarily focuses on promoting productive student learning. That is, the essence of learning-oriented assessment is that all assessment whether predominantly summative or formative is focused on developing effective student learning processes (Carless, 2015). The figure below indicates three main components of learning oriented assessment; assessment tasks, student involvement, and feedback. It is, however, important to note that these three strands are not independent from each other. For example, assessment tasks are most effectively focused on learning when they incorporate student involvement and how feedback loops can be closed; feedback is likely to be more effective when students are cognisant of criteria and are monitoring their progress(Carless, 2015). Each of these strands will be mentioned briefly;

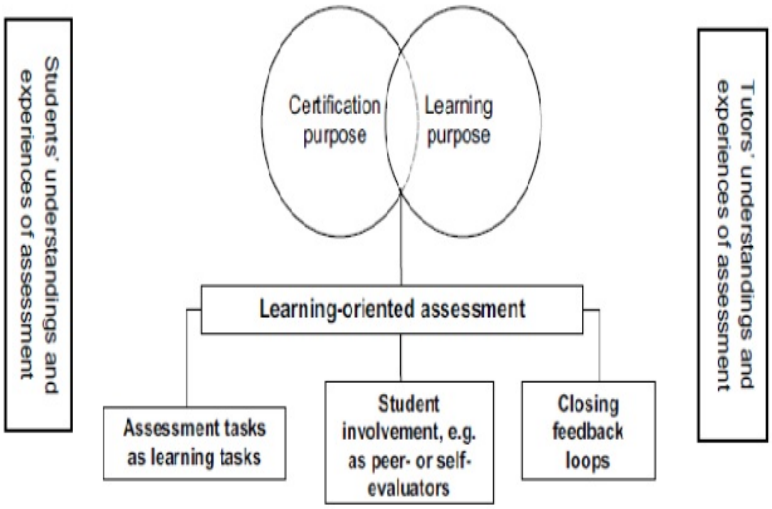


Figure 3.Framework for Learning Oriented Assessment, (Carless, 2009)Learning-oriented assessment: Principles, practice and a project. In L. H. Meyer, S. Davidson, H. Anderson, R. Fletcher, P.M. Johnston, & M. Rees (Eds.), *Tertiary Assessment & Higher Education Student Outcomes: Policy, Practice & Research* (pp.79-90). Wellington, New Zealand: AkoAotear

In learning-oriented assessment, tasks are designed to foster student learning. So, assessment tasks are also considered as "learning tasks" and these tasks are required to mirror curriculum. Another quality of the tasks is that they should encourage higher order thinking. Active participation as highlighted in dynamic assessment, also counts for learning oriented assessment(Carless, 2015a).

Students are required to be a part of assessment by peer or self-evaluation so that self-regulation takes place, that's why, self-evaluation is an important skill and a key element in learning-oriented assessment. Peer- evaluation or peer feedback, on the other hand, provides an important role in learning from each other (Carless, 2015a).

Another strand, feedback, is necessary for students to close the gap between current and desired level of performance. By closing the feedback loop, Carless refers to providing feedback which is acted upon by the student to enhance their learning and that the giver of feedback(Carless, 2015a).

To summarize, there are three principles in learning-oriented assessment;

Principle 1: Assessment tasks should be designed to stimulate productive learning practices amongst students;
Principle 2: Assessment should involve students actively in engaging with criteria, quality, their own and/or peers' performance;
Principle 3: Feedback should be timely and forward-looking so as to support current and future student learning(Carless, 2015a).

**Conclusion**

In this paper, the aim was to present an in-depth review of language assessment. First, the foundation of language assessment was presented followed by more recent approaches and their effects on assessment procedures discussing the changing trends and theories in "validity" and "reliability". This review is important to understand how language assessment evolved and what future holds.

So, where are we now? The point we have come is explained by Boud; assessment has "double duty". It is about grading and about learning(2000).It is both a technical matter and one that impacts on students emotional lives (Carless, 2009). But the need for assessment to carry multiple functions is a major challenge to the improvement of its practice.
Furthermore, the tension between the two paradigms (positivist and constructivist) has not been solved and seems to last longer. But constructivist paradigm led alternative assessment to become a complementary type of assessment and in the future it seems that alternative(s) in assessment will gain ground. With current developments, we can speak of a more homogenous and a humanistic view of assessment.

As for reliability and validity, it can be concluded that the concepts are no longer seen as two distinct qualities of a test. With the introduction of new models, reliability ceded ground to validity, making validity a larger term regarding the social factors added to it.

In the future, as highlighted by Bachman, development of authentic tests of communicative language tests and validation research for providing insights to nature of communicative language use should be prioritized. “The complexities of both the language abilities we wish to measure and the facets of the procedures we must use to measure these abilities, along with the need for language tests that are usable, the challenges facing language testers are immense” (Bachman, 1990, p.357), however, what we have at our disposal is great, too.

**References**

Alderson, J.C.,& Wall, D. (1993). Does Washback exist? *Applied Linguistics*., Vol. 14, pp. 15-29.

Aschbacher, P.A. (1991). Performance Assessment.State Activity, interest, and concerns.*Applied Measurement in Education*, Vol. 4, pp. 275-288.

Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford University Press: Oxford.

Bachman, L.F., & Palmer, A.S. (1996).Language Testing in Practice. Oxford University Press: Oxford.

Bachman, L.F. (2000). Modern Language Tests at the turn of the century: Assuring that what we count counts. *Language Testing*, 17 (1), pp. 1-42.

Bailey, K. (1996). Working for Washback: A Review of the washback Concept in Language Tests. *Language Testing*, Vol.13, pp. 257-279.

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, *22*(2), pp. 151-167.

Büchel, F.P., &Scharnhorst,U. (1993). 'The learning potential assessment device(LPAD): Discussion of theoretical and methodological problems', in J.H.M. Hamers, In K. Sijtsma, and A.J.J.M. Ruijssenaars (eds.), Learning Potential Assessment: Theoretical, Methodological and Practical Issues, Swets&Zeitlinger, Amsterdam.

Canale, M., &Swain, M. (1981). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1(1), 1–47.

Carless, D. (2007). Learning-oriented assessment: Conceptual basis and practical implications. *Innovations in Education and Teaching International, 44*(1), 57-66.Carless, D. (2007b). Conceptualizing pre-emptive formative assessment. *Assessment in Education, 14*(2), pp.171- 184.

Carless, D. (2009) Learning-oriented assessment: Principles, practice and a project. In L. H. Meyer, S. Davidson, H. Anderson, R. Fletcher, P.M. Johnston, & M. Rees (Eds.), *Tertiary Assessment & Higher Education Student Outcomes: Policy, Practice & Research* pp.79-90. Wellington, New Zealand: AkoAotear

Carless, D. (2015a). Excellence in University Assessment: Learning from award-winning practice. London: Routledge.

Carless, D. (2015b). Exploring learning-oriented assessment processes. *Higher Education*, 69(6), pp.963-976.

Carrol, B.J. (1986). LT + 25, and beyond? Comments. *Language Testing*, Vol.3, pp. 123-129.

Coombe, C., Davidson, P., O'Sullivan, B., &Stoynoff, S. (2012). The Cambridge Guide to Second Language Assessment. Cambridge: Cambridge University Press.

Cronbach, L.J, &Meehl, P.E (1995). Construct Validity in Psychological tests. *Psychological Bulletin*, 52, pp. 281-302.

Davies, A. (1968). Language Testing Symposium: A Psycholinguistic Approach. Oxford University Press: UK.

Dudley-Evans, T., & St. John, M. J. (1998).Developments in English for Specific Purposes. Cambridge: Cambridge University Press.

Farhady, H. (1983a). On the plausibility of the unitary language proficiency factor. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp.11–28). Rowley, MA: Newbury House.

Farhady, H. (1997). Construct Validity in Language Testing. *The Proceedings of the conference on language, cognition and interpretation*. Islamic Azad University: Khorasgan.

Farhady, H. (1999). Ethics in Language Testing. *Moddaress Journal of human sciences* 3 (11), pp.447-464.

Farhady, H. (2005). Language Assessment a Linguametric Perspective *Language Assessment Quarterly*, *2*(2), 147–164.

Farhady, H. (2014).Language Assessment [Powerpoint Slides].

Fulcher, G. (1999b) 'Ethics in language testing.' *TAE SIG Newsletter Vol.*1, (1), pp.1–4.

Fulcher, G., & Davidson, F. (2007). Language Testing and Assessment: An advanced Resource Book. Routledge: New York.

Giri, A. (2003). Language Testing: Now and Then. Journal of NELTA, 8(1), pp.49-67.

Guterman, E. (2002).Toward dynamic assessment of reading: applying metacognitive awareness guidance to reading assessment tasks. *Journal of Research in Reading*, Vol.25, pp. 283-298.

Hamp-Lyons, L. (1997). Washback, impact and Validity: Ethical Concerns. *Language Testing*, Vol.14 (3), pp. 295-303.

Henning, G. (1987). A guide to language testing: Development, evaluation and researchCambridge, Mass: Newbury House

Hudson, T., Detmer, E., & Brown, J.D. (1992).A Framework for Testing Cross-cultural Pragmatics. Second Language Teaching and Curriculum Center: University of Hawaii.

Huerta-MaciasA. (1999). Alternative Assessment: Responses to commonly asked questions. *TESOL Journal*, Autumn, pp.8-11.

Hutchinson, T., & Waters, A. (1987).English for Specific Purposes. Cambridge: Cambridge University Press.

Kunnan, A.J. (2000). Fairness and Justice for All.19th Language TestingResearchColloquium, Orlando, Florida: Cambridge UniversityPress.

Lado, R. (1961). Language Testing. Longman: London.

Lantolf, J. P. andThorne, S. L. (2006).*SocioculturalTheoryandtheGenesis of Second Language Development*. Oxford: Oxford UniversityPress.

Lantolf, J.,&Poehner, M. (2008). DynamicAssessment in E. Shohamyand N. H. Hornberger (eds), Encyclopedia of Language andEducation, 2nd Edition, Language TestingandAssessment, Vol. 7, pp. 1–12.

Madsen, H.S. (1983). Techniques in Testing. Oxford University Press: Oxford

Messick, S. (1989).“Validity” In Linn, R.L (Ed.) *Educational Measurement*. MacMillan New York.

Messick, S. (1996).Validity and Washback in Language Testing. *Language Testing*, Vol.13, pp.241- 256.

Oller, J. (1979). 'The Psychology of Language and Contrastive Linguistics: The Research and the Debate'.ENC EJ206643.

Özgür, B.,& Kantar, M. (2012). Dynamic Assessement and Zone of Proximal Development. *EALTA*, Ankara.

Peña, E.D.(2000). Measurement of modifiability in children from culturally and linguistically diverse backgrounds. *Communication Disorders Quarterly*, 21(2), pp. 87-97.

Popham, W. J. (1997). Consequential Validity: Right Concern Wrong Concept. *Educational Measurement.,*9, pp.9-13.

Savignon, S.J. 1983: *Communicative competence: theory and classroom practice*. 1st edn. Reading, MA: Addison-Wesley.

Shohamy, E. (1997). Testing Methods, Testing Consequences: Are they ethical? Are they fair? Language Testing, Vol. 14, pp.340-349.

Skehan, P. (1988). Language testing, part 1: state of the art article. Language Teaching, Vol. 21, (4), pp.211-218.

Spolsky, B. (1981). Some ethical questions about language Testing. In C. Klein-Braley and D.K: Stevenson (eds.), Practice and Problems in Language Testing 1(pp.5-21). Frankfurt, Germany, Verlag Peter Lang.

Spolsky, B. (1978). Linguistics and language testers. In B. Spolsky (Ed.) Papers in applied linguistics: *Advances in language testing*, Vol.2, pp.v-x. Arlington Virginia: The Center of Applied Linguistics.

Tratnik, A. (2008). Key Issues in Testing English for Specific Purposes. *Scripta Manent,* Vol, 4, (1), pp. 3-13.

Underhill, N., 1987. Testing Spoken Language. Cambridge University Press, Cambridge.

Widdowson, H.G. 1983: *Learning purpose and language use*. Oxford: Oxford University Press.

Willingham, W.W.,&Cole, N.S. (1997). Gender bias and fair assessment. Hillsdale: NJ: Erlbaum.

Yıldırım, Ö. (2008). Vygotsky's Sociocultural Theory and Dynamic Assessment in Language Learning. *Anadolu University Journal of Social Sciences*, Vol. 8, (1), pp.301-308.