MINISTRY OF HIGHER AND SECONDARY SPECIALIZED EDUCATION OF THE REPUBLIC OF UZBEKISTAN

NUKUS STATE PEDAGOGICAL INSTITUTE NAMED AFTER AJINIYAZ

FACULTY OF FOREIGN LANGUAGES

ENGLISH LANGUAGE AND LITERATURE DEPARTMENT

QUALIFICATION PAPER

On the theme: "The role of Assessment in Language Education"

Fulfilled: the IV-th year
student Z. Kdrniyazova
Scientific supervisor:
Ph.d. D.Dj. Mamirbaeva
Reviewer: Ph.d. A. Tajieva

Qualification paper is admitted to the defence

Protocol № _____ «___» ____________ 2015

Nukus - 2015

МИНИСТЕРСТВО ВЫСШЕГО И СРЕДНЕГО СПЕЦИАЛЬНОГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ УЗБЕКИСТАН

НУКУССКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ ИНСТИТУТ ИМЕНИ АЖИНИЯЗА

ФАКУЛЬТЕТ ИНОСТРАННЫХ ЯЗЫКОВ

КАФЕДРА АНГЛИЙСКОГО ЯЗЫКА И ЛИТЕРАТУРЫ

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

НА ТЕМУ: «РОЛЬ ОЦЕНИВАНИЕ В ОБУЧЕНИЕ ЯЗЫКАМ»

Выполнила: студентка
4 г курса З. Кдрниязова
Научный руководитель:
к.п.н. Д.Дж. Мамырбаева
Рецензент: к.п.н. А. Тажиева

Выпускная квалификационная работа допускается к защите.

Протокол №___ «___»______________2015

Нукус – 2015

Content

Introduction

After the collapse of the Soviet Union in 1991, the Russian language, which was a dominant language in all 15 former Soviet republics, was overtaken by English and has become the most widely learned foreign language in the Republic of Uzbekistan. It has become an important vehicle of promoting an open society in the newly independent state and the number of users and learners of English grew rapidly because of its important place in the job market and global education. The emerged need for integration to world community led to increased popularity of English in Uzbekistan and its use in different domains like, business, culture, academia, medicine and many others significantly increased. Especially, it became popular and prestigious among the youth, interested in an access to world technology and culture.

The role of English can be viewed from the perspectives of Braj Kachru s model of Concentric Circles. He has described the spread of English in terms of three concentric circles: the inner circle, the outer circle, and the expanding circle. Kachru s grouping is based on the following criteria: “types of spread, the patterns of acquisition [and] the functional domains in which English is used across [cultures] and languages” Kachru, 1985). The inner circle consists of countries where English is spoken as a native or primary language: the UK, the USA, Canada, Australia, and New Zealand. The outer circle countries are mainly former British colonies, like India, Nigeria, Singapore, Kenya, etc.

As for the Expanding Circle, it comprises countries where English is learned as a foreign language. These are: China, European countries, Latin American countries, and former Soviet Union countries. According to Kachru, “it is the [users] of this circle who actually [further] strengthen the claim [of] English as an international [or] universal language” Uzbekistan belongs to the Expanding Circle as English is not officially used, and its functions in various fields is rather limited. In the Constitution of the Republic of Uzbekistan adopted on December 8, 1992, it is stated that Uzbek is the only state language and no other indigenous or foreign languages are given any official status in the country. Thus, English is primarily learnt through formal schooling the quality of which still leaves much to be desired. More and more adult EFL learners are taking English classes at both private and public language learning centers. The increased demand for learning English raised the question of reforming the foreign language learning system in Uzbekistan.

The government of Uzbekistan has always acted as a primary initiator of educational reforms, and special attention has been paid to foreign language learning. The importance of the reform of the education system should be considered in the specific demographic context of Uzbekistan: about 41% out of total population of about 30 million people are under 17 years of age. Radical reform of the education sector in Uzbekistan started in 1997 with the adoption of the Education Act and the National Program for Personnel Training (NPPT). The two documents have provided a legal basis for higher education and further development of the educational system in Uzbekistan. National policy in the field of education, including legislation on higher education, is based on the Constitution of the Republic of Uzbekistan, Decrees of the President of the Republic of Uzbekistan and Resolutions of the Cabinet of Ministers of the Republic of Uzbekistan.

In the framework of the Law of the Republic of Uzbekistan On Education and the National Program for Personnel Training in the country, a comprehensive foreign language teaching system aimed at creating a harmoniously developed, highly educated, forward-thinking young generation, further integration of the country to the world community, has been created. During the years of independence, over 51.7 thousand teachers of foreign languages graduated from universities, English, German, and French multimedia tutorials and textbooks for 5-9 grades of secondary schools, electronic resources for learning English in primary schools were created, more than 5000 secondary schools, professional colleges and academic lyceums were equipped with language laboratories.

On December 10, 2012, President of the Republic of Uzbekistan Islam Karimov signed the Decree on measures for the further Improvement of the

System of Foreign Language Education. According to the decree, starting from 2013/2014 school year, instruction of foreign languages, mainly English, will be offered from the first grade of schools. Moreover, English will be a primary language of instruction at the institutions of higher learning, especially those providing technical and international education. Besides, as declared by the Decree, the salaries of foreign language teachers are expected to be increased of 30% and 15 % in rural and urban areas (respectively).

The National TeleRadio Company, State Committee for communications, information and telecommunication technologies, Agency for Press and Information of the Republic of Uzbekistan are appointed to prepare and broadcast language-learning programs, significantly increase access to international educational resources via *Ziyonet* educational network, promote publication of foreign language textbooks, magazines and other materials. The State Testing Centre, along with other relevant agencies, is requested to prepare draft proposals on introducing foreign languages testing to the entrance examinations for all higher educational institutions.

Although much has been done in this domain, the analysis of the current system of organizing language learning shows that learning standards, curricula and textbooks do not fully meet the current requirements, particularly in the use of advanced information and media technologies. Education is mainly conducted in traditional methods. Though English is taught at school from the first or second grade through ninth, and continued in colleges, lyceums, and institutions of higher education in Uzbekistan, students can hardly produce any English by the time they graduate. Very often the Language program encourages students to memorize isolated lexis items and grammar rules and often it fails to build communicative abilities. Even those students, who had quite good knowledge of grammar and vocabulary, actually, cannot speak English fluently.

The actuality of this research work is that it describes reasons for the partial implementation of innovative ELT methods and changes that fit with the paradigm shift, most often described as communicative language teaching.

English language learners (ELLs)—students who are still developing proficiency in English— represent a large and rapidly growing subpopulation of students in English language classrooms.

The aim of this Qualification Thesis is to highlight the role of Testing and Evaluation as one of the important aspect of the process of language learning and teaching. A Linguistics approach to language teaching is a scientific and objective approach and is based on the theoretical knowledge of Linguistics. Since language testing involves language, one cannot ignore the assumptions of Linguistics. Linguistics has to offer many things to the teaching of native and foreign languages. Similarly it is also recognized that Linguistics can be of great help in evolving the methodology of the construction of language tests.

The subject of the paper is the process of Assessing and Testing students' knowledge.

Testing and evaluation has attracted much attention of the Scholars of Linguistics as well as language teachers. Much research has been done in this particular area. There have been revolutionary changes in the method & procedure of language testing. Research has been conducted in order to find out a scientific and standard method for language testing. Various linguists have suggested the measures which one has to take into account while constructing a language test. There are various requirements of a test, without which a test cannot be considered a valid and standard. Ingram (1974:319) observes that the search for 'objective' testing methods is the direct out come of dissatisfaction with the unreliability of the marking of traditional examinations.

Like in any other field, change seems to be constant in education. Although the paradigm shift was initiated several years ago, it still has been only partially implemented. Even when teachers and other stake-holders are striving to go toward the new paradigm, in many cases, while teaching methods have become more or less communicative, teaching methodology itself remains with the traditional paradigm, very often as teacher-centered and the ways of assessing and testing is traditional as well.

At a typical English lesson delivered at school, lyceum, or even institution, a teacher explains the rules and then students usually translate sentences from Uzbek/Russian into English or reverse. Moreover, when they use a communicative textbook, which became possible owing to the new secondary textbooks recommended by the Ministry of Education, they tend to have students read the texts aloud taking turns, translate them into their mother tongue, make up sentences with the new words, etc. and consequently break the idea of the communicative textbook.

The Practical Value of this work is that the methods and strategies which we are going to investigate in this paper will help to the teachers and students of English Language Department in their Professional Development.

This situation corresponds to the simile of Jim Scrivener, who claims that "traditional teaching [is imagined to work as] 'jug and mug – the knowledge being poured from one receptacle into an empty one." This widespread attitude based on the precondition that "being in a class in the presence of a teacher and 'listening attentively is enough to ensure that learning will take place" (Scrivener) is quiet popular in Uzbekistan. Mainly the responsibility for teaching and learning is put on the teacher and it is believed that if students are present in the lesson and listen to the teacher's explanations and examples, they will be able to use the knowledge.

An interesting issue is that there are many teachers in Uzbekistan who claim that they use modern methodology. However, most teaching at our schools and universities is described as traditional by students and observing teachers. Even in the case when interactive methods are effectively used, the assessment remains as traditional, mainly in the form of traditional tests which simply assess bits of memorized knowledge and ignore the use of language skills.

The Theoretical significance of the research work is that data collected for this paper can be used as the resource during teaching practice, designing portfolio, manuals and lectures on Methodology. Obviously, many teachers are encountering resistance, challenges and frustrations and there are many who are confused by changes and simply do not know what to do or do not want to accept the change.

And number of questions arises regarding the teaching practices in Uzbekistan: Is the teaching these days traditional, or modern, or something unspecific? Do we perceive it to be ‘traditional because of our expectations? And if the teach ing really is traditional, why do so many teachers claim they use the modern methodology?

It is difficult to find the answers to these questions, because people have different backgrounds and language learning experiences, and share different beliefs about language learning and classroom practices. Also, all methods have some positive as well as negative aspects, which are highlighted by professionals in their publications.

The Qualification Paper consists of Introduction, two Main Chapters, Conclusion and the list of Literature used.

Chapter I. Theoretical Aspects of Testing and Assessment

The increasing demand for evaluation, assessment, and accountability in education comes at a time when the fastest growing student population in Uzbekistan is learners whose home language is not English. This presents several challenges to practitioners and school systems generally when they lack familiarity with important concepts such as second language acquisition, acculturation, and the role of socioeconomic background as they relate to test development, administration, and interpretation. Because assessment is key in developing and implementing effective curricular and instructional services that are required to promote student learning, English language learner (ELL) students have the right to be assessed to determine their educational needs. Through individual assessments, teachers can personalize instruction, make adjustments to classroom activities, assign students to appropriate program placements, and have more informed communication with parents. They can also identify learning problems that may require additional outside assistance. And educational systems need to know how ELLs are performing in order to make proper adjustments to their programs and to effect necessary policy changes. Notwithstanding the increasing need, states have struggled to develop strong assessment programs with appropriate instruments for use with young ELLs.

1.1. Factors Influencing the Assessment of English Language Learners

Different linguistic backgrounds ELLs possess a wide range of linguistic backgrounds. While the majority of ELLs come from Uzbek- or Karakalpak-speaking backgrounds, it has been estimated that approximately 400 different native languages are spoken by ELLs nationally. This is particularly important to keep in mind when considering the use of native language testing accommodations, since it may not be possible to provide assessments in all native languages represented in a large school district or a state. Varying levels of proficiency in English—ELLs vary widely in their level of English language proficiency, and furthermore, ELLs may have varying levels of oral and written English proficiency. Do not assume that students who can converse easily in English will have the literacy skills necessary to understand the written directions for a standardized test. Some ELLs may be proficient in the English used for interpersonal communications but not in the academic English needed to fully access content-area assessments. Studies show that the level of language proficiency has an influence on processing speed. In other words, compared with native speakers, ELLs generally take longer on tasks presented in English. This is important to keep in mind when designing and scoring the assessment, as well as when making decisions about testing accommodations. Varying levels of proficiency in native language - ELLs also vary in their levels of proficiency and literacy in their native languages. Therefore, do not assume that speakers of other languages will be able to understand written test directions in their native languages. In fact, a large proportion of ELLs were born in Uzbekistan and may not have had any formal schooling in their native language. This is important to keep in mind when considering the use of native language accommodations. Educational Background Factors

Varying degrees of formal schooling in native language—As mentioned previously, ELLs vary widely in the level of formal schooling they have had in their native languages. The degree of native-language formal schooling affects not only native language www.ets.org Guidelines for the Assessment of English Language Learners proficiency—specifically, literacy in the native language—but also the level of content-area skills and knowledge. For example, students from refugee populations may enter the Uzbekistan educational system with little or no formal schooling in any language. These students must learn English and content-area knowledge simultaneously, while also being socialized into a school context that may be extremely unfamiliar. Other ELLs may come with more formal schooling and may have received instruction in the content areas in their native languages. The primary challenge for these students is simply to transfer their existing content knowledge into English. Again, these factors come into play when making decisions about appropriate accommodations.

ELLs also vary in the number of years they have spent in schools where English is the language of instruction. A distinction may also be made between students who have studied English as a foreign language while in their home countries. Furthermore, ELLs differ in the type of instruction they have received. Bilingual, full English immersion, and English as a second language are but three of the many existing instructional programs for non-native English speakers, and there are great variations in how these programs are implemented.

It should not be assumed that all ELLs have had the same exposure to the standardized testing that is prevalent. Students in some countries may have had no exposure to multiple-choice questions, while those from other countries may never have seen a constructed-response question. Even ELLs from educationally advantaged backgrounds and with high levels of English language proficiency may not be accustomed to standardized, large-scale assessments and may be at a disadvantage in these testing situations. Guidelines for the Assessment of English Language Learners www.ets.org

Cultural factors can also be potential sources of construct-irrelevant variance that add to the complexity of appropriately assessing ELLs. Varying degrees of acculturation to U.S. mainstream—ELLs come from a wide range of cultural backgrounds, and cultural differences may place ELLs at a disadvantage in a standardized testing situation. Lack of familiarity with mainstream American or British culture, for example, can potentially have an impact on test scores for ELLs. Students who are unfamiliar with American and British culture may be at a disadvantage relative to their peers because they may hold different assumptions about the testing situation or the educational environment in general, have different background knowledge and experience, or possess different sets of cultural values and beliefs, and therefore respond to questions differently. Students from cultures where cooperation is valued over competition, for example, may be at a disadvantage in those testing situations in the United States where the goal is for each individual student to perform at his or her best on his or her own. Students from economically disadvantaged backgrounds may also respond to questions differently and may have background knowledge and experiences that are different from those presumed by a test developer.

1.2. Types of Tests and their Requirements

Tests are set up so as to eliminate any differences in results due to variations in the judgment of one marker at different times. The objective type test derives its name from the objectivity in scoring. Since there is only one correct answer to such a test and since in most cases the answer is given along with the test and the candidate is simply required to indicate the correct answer with a tick or a number, the subjective judgment of the examiner cannot vitiate the scoring. As far as scoring is concerned, these tests are highly reliable. Some of the popular types of objective type tests are: Constant-Alternative, Rearrangement type, Multiple-Choice type, Matching type, True-False type, Yes/No Answer type, Completion type (Fill-in the blank type) etc.

The objective type tests have certain distinct advantages over the traditional essay type tests, i.e. (a) objective type tests can cover a large area of syllabus in a relatively short time and (b) objective type tests can be scored easily & objectively. Traditionally, the system of evaluation was subjective in nature. Bhat (1992) is of the view that the present examination system is conducted to determine fail or pass of the participants. The examinations test the knowledge of textbooks and the competence of the teacher rather than the competence of the student. Major portions of syllabus are deleted while setting up question papers. This infused a stigma of “Choice Making” on the part of learners and the aim is only to pass the examination rather than master the course. This contributed to the degeneration of the evaluation system. A subjective test is based on an opinion or judgment on the part of examiner, which is expected to match with that of an examinee. It involves more of memorization on the part of learner, while an objective test is scored mechanically and involves measurement. Some of the popular types of Subjective type tests are: Simple Question type, Short Answer type, Long Answer type, Problem solving, Completion etc.

A test has been defined as a “measuring device”. Measurement is the process of assigning numerical value to the response for a given task to each of the members or a set of objects or group of persons normally examinees.

Ingram (1974:313) is of the view that “tests, like examinations, invite candidates to display their knowledge or skills in a concentrated fashion, so that the result can be graded, and inferences made from the standard of performance that can be expected from the candidate, either at the time of the test or at some future time”. A test is conducted to measure the knowledge of an individual and to compare him with other individuals who belong to the same group. According to Carrol (1965: 364), “the purpose of testing is always to render information to aid in making intelligent decisions about possible courses of action. Some times these decisions affect only the future design or used of the tests themselves, in which case we are dealing with solely experimental uses of tests. Some times the decisions have to do with the retention or alteration of courses of training, as when one decides that poor tests results are due to in effective training”. Pit Corder (1973:351) is of the view that “language tests are measuring instruments and they are applied to learners, not to the teaching materials or teachers. For this reason they do not tell us ‘directly’ about the contribution of the ‘teacher’ or the ‘materials’ to the learning process. They are designed to measure the learners ‘knowledge of’ or ‘competence’ in the language at particular moment in his course and nothing else. The knowledge of one pupil may be compared with the knowledge of others or with that of the same pupil at a different time, or with same standard or norm, as in the case of height, weight, temperature etc.” According to Halliday, et al., (1966:215), “tests are an attempt to construct an instrument for measuring attainment, or progress, or ability in language skills.” Thus, testing is a set of techniques of questioning and observing to find out how far learning is taking place, whether the students are following the teacher or instructor, and what are the problems of the students? It is also used to assess the knowledge of the students in order to compare one individual to another individual in the same group. The term evaluation in modern educational practice is used for “tests” and “examination”. It is a general term that covers both. It is a much more comprehensive term than either test or examination. The term test refers to the measurement of the competence of the learners with reference to the particular area of knowledge, whereas the term examination refers to particular standard that is to be achieved by the learner after a particular level. A test is regarded as an attempt to see whether the things taught have been learned, while examination is regarded as an attempt to find out whether the students have attained certain predetermined standard. Thus, a test is directly concerned with teaching while an examination is linked with an externally fixed standard of achievement. However, since both tests and examinations have the same common function, namely evaluation, it has become conventional to call them ‘tests’.

In the area of testing and evaluation, evaluation refers to the judgment of performance as process or product of change. In other words, it is the process of testing, appraising and judging achievement, growth, product, process or changes in these, through the use of formal and informal tests and techniques. The process of evaluation is global in conception and application. There are three major components that constitute the concept of evaluation and testing, i.e. Content, Method and Purpose. Content: Content has different connotation in testing. The general assumption is that whatever has been taught is to be tested. Hence, whatever is assumed as content for teaching will become the content for testing too? In Second Language teaching, structure gets focused as main content. In First Language teaching, meaning gets the main focus and in the area of education, the traits of personality. Method: A means or manner of procedure, especially a regular and systematic way of accomplishing something. In other words, it refers to the plans or procedures followed to accomplish a task or attain a goal. In testing, it refers to the procedure to be followed according to a definite, established, logical or systematic plan. Purpose: In the field of testing it is defined as the reason for which something exists/happens. It is synonymously used to represent the terms goal, aim and objective. Goal refers to a very broad and ultimate category, aim to a more specific set of purposes, and objective as the most precisely defined ends which can be described in terms of behavioral outcomes in the field of education.

Tests are designed for different purposes, which help in making decisions about possible course of action. Keeping in view the purpose, tests may be characterized as follows:

Proficiency tests: It is designed to find out how much of a language a person actually knows. As Davies (1977:46) suggests, “proficiency tests, as we see it, are concerned with assessing what has been learned of a known or an unknown syllabus”. This test may be used generally before language teaching programme in order to prepare the teaching materials for the learning programme. On the basis of the information collected through this test, language teachers prepare their teaching materials. This test is very helpful in order to organize teaching materials according to the current need.

Achievement tests: This test is constructed to find out that how much of a course a learner has actually mastered. Paterno (1965: 376) is of the view that “An achievement test is an inquiry to see if what has been taught is retained”. It determines that how much of the material of a course has actually been mastered by the learner. This includes only what has been taught to him. These tests are generally given at the end of the teaching programme.

Prognostic tests: This test is designed to predict the knowledge of a person, that, up to which level he is able to learn. This test is very useful for selecting the students in any language learning programme and the material of teaching, which is sufficient to the learner in any teaching programme.

Diagnostic tests: Diagnostic test differs from other tests on the basis of the use of the information obtained and to the absence of a skill in the learner. The purpose of this test is to find out what remains to be taught during the course of learning. As Davies (1977:47) points out that, “a diagnostic test may be constructed for itself or it may be an additional use made of an achievement or proficiency test. If it is specially constructed it could perhaps be argued that some element of learner’s skills, or rather absence of skill, is involved because the tester is concerned with discovering what might be termed non-achievement”. Through this test a teacher can know where the learner needs more attention and which area of language skills has to be practiced more. This test also points out the shortcomings of the learner and of teaching materials. And if teacher will know shortcomings of the learner and of teaching materials, he could adopt certain remedial methods in order to remove the learning difficulties.

Objective tests are designed to elicit specific responses. It can be quickly judged as correct or incorrect. Objective tests can be of two types: discrete items and passage items. Tests can be constructed successfully only after the objectives of the course are finalized. An instructor has to determine the objectives of the test as well as prepare a general plan in advance. In modern language teaching programmes, an equal emphasis is given to all the basic language skills i.e. Comprehension, Speaking, Reading and Writing, from the beginning of the course, and language tests are to be prepared accordingly. In language testing, time is an important factor. Both the duration of a test and its proper administration at regular intervals are essential factors to be kept in mind. There are three main stages of preparation of a test.

(i) Planning: It covers outlining test, listing of topics, casting of ideas for items and material collection.

(ii) Composition: This includes the composition of actual items and choice for objectivity.

(iii) Analysis: It consists of determining difficulties and discrimination of test items. Speediness of test and scope for its improvement.

As Bachman (1992: 119) suggests, a language test can be classified in terms of five characteristics, which are as follows:

- Test can be distinguished according to their intended use, such as selection, entrance, readiness, placement, diagnosis, progress, attainment and mastery.
- Tests can differ in content; Achievement tests are based on syllabus, while a proficiency test derives a theory of language ability.
- Different frames of reference can provide the basis for test development and score interpretation norm referenced tests are developed to maximize differences among individual test takers and a test score is interpreted in relation to the score of the test takers.
- Tests can be classified according to the scoring procedure (the act or process of evaluating responses to test situations or evaluating characteristics of whomever or whatever is being rated. It consists of checking the student's response to each item to see if it is correct. Scoring objective tests is purely mechanical process which requires no special skills);

Objective tests require no judgment on the part of the scorer but in subjective tests, the scorer must judge the correctness of the test taker's response. Tests may employ different testing methods, such as dictation, cloze, multiple choice, completion, composition and interview.

A test has been characterized by certain features, which can be termed as "requirements of a test". Ingram (1974:313) has discussed these requirements of a test. A good test must meet at least six requirements, which are as follows:

(i) Discrimination: It is one of the most important requirements, which is necessary for a test. It means that a test must be designed in such a way that it can discriminate among the students. If we want to measure the height of the school children, we should use such a measuring device, which is suitable for the students whom we are going to measure.

(ii) Reliability: Reliability refers to the accuracy of a measuring instrument that is if a student is tested again and again the result or score must always be the same, regardless of who is giving and marking it. As Paterno (1965:379) suggests that, “A test that lacks reliability is as useless as a thermometer that gave different reading when the temperature of the air was the same. A test is reliable if it will always give the same results under the same conditions”. A test must have consistency in it. As Davies (1977:57) is of the view that, “ A reliable test possesses consistency of results. An inconsistent test would give meaning less, random results. Before looking at the meaning of results, it is important to ensure that they are reliable”.

(iv) Validity: It means that the test should measure the same for which it has been devised. If it does, it is a valid test. If a test of pronunciation tests only pronunciation and nothing else, it is a valid test of pronunciation. Paterno (1965:378) suggests that, “Validity can only be obtained when we state clearly the objectives of our teaching, break them down into skills and abilities involved, and define them in separable elements; and then to measure each in situations which comes as close as possible to the real circumstances in which they will be used”. For instance if listening comprehension in English is aimed at, it must be tested in a variety of ways that approach the actual, normal use of language. Ingram (1974:315) is of the view that, “the most obvious way of achieving validity is to arrange for a job sample. If you want to know how good a person is at writing essays, you ask him to write an essay, if you want to know how fluent he is in a foreign language, you ask him to talk to you. The trouble is that, validity is limited by reliability; no test or examination can be anymore valid than it is reliable. So if it turns out that the reliability of marking essays or of rating command of spoken language is low, then validity of the marks or ratings must be correspondingly low”. If the reliability of a test ensures its consistency, validity ensures its meaningfulness. A test is meaningful, within the terms of what is wanted from the test.

(v) Scorability: It refers, that the test should be scored with ease so that the user may be able to handle it. Subjective tests are not easy to score as compared to objective tests. Secondly there should not be any differences in scoring. The difference will affect the accuracy of the test. (v) Economy: This is practical criterion the test should measure what it wants to test and it should also measure in a reasonable time. If it does, the test is practical and economical.

(vi) Administerability: It means that a test should be such that it may be given under the conditions that prevail and the personnel (person who is conducting the test) that are available. For instance, if a test requires electronic equipments and the service of highly trained technician, then it is not administrable since these facilities are not available in most of the school and even most colleges and Universities.

Chapter II. The Role of Assessment in Language Education

Assessing the development of ELLs demands an understanding of who these learners are in terms of their linguistic and cognitive development, as well as the social and cultural contexts in which they are raised. The key distinguishing feature of these learners is their non-English language background. In addition to linguistic background, other important attributes of ELL children include their ethnic, immigrant, and socioeconomic histories (Abedi, Hofstetter, & Lord, 2004; Capps et al., 2005; Figueroa & Hernandez, 2000; Hernandez, 2006). Though diverse in their origins, ELL students, on average, are more likely than their native English-speaking peers to have an immigrant parent, to live in low-income families, and to be raised in cultural contexts that do not reflect mainstream norms (Capps et al., 2005; Hernandez, 2006).

Decades of research support the notion that learners can competently acquire two or more languages (Garcia, 2005). Relationships of linguistic properties between languages are complex, and several theories have been presented over the years to explain how language develops for young bilingual learners. Among the major theoretical approaches, available empirical evidence suggests that transfer theory best explains the language development of students managing two or more languages (Genesee, Geva, Dressler, & Kamil, 2006). This theoretical position asserts that certain linguistic skills from the native language transfer to the second. In like manner, errors or interference in second language production occurs when grammatical differences between the two languages are present. Language that is contextually-embedded and cognitively undemanding—or automatic, over-learned communication—does not lend itself well to transfer. Contextually-reduced and cognitively demanding language skills, on the other hand, tend to transfer more easily between languages. Higher order cognitive skills relevant to academic content are more developmentally interdependent and, therefore, amenable to transfer (Genesee, Geva, Dressler, & Kamil, 2006). In the process of crosslinguistic transfer, it is normal for children to mix (or "code-switch") between languages. Mixing vocabulary, syntax, phonology, morphology, and pragmatic rules serves as a way for young bilingual children to enhance meaning.

2.1. Using Statistics to Evaluate the Assessment and Scoring

Alderson et al (1995: 9 – Document 4.1 Pack b)) define test specifications in the following way:

> A test's specifications provide the official statement about what the test tests a ow it tests it. The specifications are the blueprint to be followed by test and ite vriters, and they are also essential in the establishment of the test's construct validit
>
> Deriving from a test's specifications is the test syllabus. ... A test specification detailed document, and is often for internal purposes only. It is sometim onfidential to the examining body. The syllabus is a public document, often mu implified, which indicates to test users what the test will contain. Whereas the t pecification is for the test developers and those who need to evaluate whether a t as met its aim, the syllabus is directed more to teachers and students who wish repare for the test, to people who need to make decisions on the basis of test scor nd to publishers who wish to produce materials related to the test.

McNamara (1996: 96 – Document 0.4, Pack b)) gives a similar definition, but adds several other features relating to assessment criteria, training of raters and the reporting of results.

Based on the definitions, and with a model of specifications for Cambridge CAE (Document 4.2, Pack b)), a first attempt was made to convert the needs identified in step 3 into a set of specifications for the ESP graduating exam (see Appendix D). A number of points emerge from these specifications:

- These are very much draft specifications, but the process was important as it gives an indication of as the product gives an indication of `what the test tests and how it tests it'.
- While these are draft specifications which would have to be revised in the light of more reliable needs analysis data and a fuller understanding of the practical constraints, they are sufficient to move on to the next stages of the test construction process – drafting of a pilot version and benchmarking.

McNamara (1996: 96) provides an introduction to this stage of the test construction process:

> A pilot version of the test instrument or procedure must be developed (from the specifications) and properly trialled on an appropriately representative population (this may be difficult in the case of a highly specific test, that is, one where the test population is small, if test security is not to be compromised).

At this stage of the ESP graduating exam development, it has not been possible to develop a full sample or pilot version of the test. The emphasis so far has been on developing sample writing and speaking items. The reason for this is the generic/specific nature of the test. As was stated earlier (see also West & Tompos 2003 – Document 0.3), the aim is to have 'narrow' versions of the exam for different academic/professional specialities. The ways in which this might be done are set out below, skill-by-skill.

Writing Each test item would begin life as a 'generic' item, i.e. an item not related to any specific speciality or discipline. This generic or 'template' item is then converted into a number of parallel items for different specialities. For example:

Each version of the item is similar in terms of:

- Genre tested (in this case, a professional report) is the same.
- Length of written output is the same (500 words).
- Rubrics are very similar.
- The amount of time available for answering the question is the same.
- The criteria and levels for grading are the same.
- The procedures for administration, evaluation, rater training, etc are the same.

Based on this concept of ‘generic’ testing, a number of items were drafted, with both generic and specific versions (see appendix E).

Speaking The same technique is applied to the speaking part of the exam, although it may remain generic, allowing the candidate to supply his or her own specific realia, picture, etc relating to his/her own field, a technique referred to as ‘portable specificity’. An example item is included in Appendix F.

Reading The problem of how to make a reading test which could be suitable for different disciplines was discussed and the discussion moved beyond the proposals in the specifications. It was suggested that there should be two texts:

Text A: A general text of a factual nature dealing with broad professional issues which might be of interest to students of any discipline. This text would be of a known CEFR level (B2?), calculated using the Oxford Text Checker (Appendix G).

Text B: A specific text selected by each institution or even each department, in order to ensure that the reading component of the exam has the required specificity. Again, the text would have to have the designated CEFR level, calculated by the Oxford Text Checker.

Of course, the difficulty level of a reading test is more than a matter of text difficulty – the difficulty of the task(s) is also a factor. In an institutional test, the overall difficulty level of a reading test is unknowable without extensive trialling or pre-testing, which may not be possible. For this reason, the difficulty level of the reading component would be calculated after the exam, and scaled appropriately to ensure that scores on reading tests which were discovered to be too easy/difficulty were adjusted in order to be certain that no student received a mark which was too high/low. This is discussed further under 'derived scores' in section 7 below.

Listening As for reading.

McNamara (1996: 96-97) discusses this stage only briefly:

It is normally assumed that trialling must take place before the exam goes live, and this is the case with large-scale public tests. However, with institutional testing, this may not be possible but item analysis of even the most basic kind (e.g. facility index) will reveal 'rogue' items which may be eliminated from the final scores or compensated for in some way.

Multiple sources of empirical evidence should be gathered to evaluate the fairness of assessments and scoring. The ETS Standards for Quality and Fairness state that, whenever possible and appropriate (i.e., if sample sizes are sufficient), testing programs should report analyses for different racial/ethnic groups and by gender, and that testing programs should use experience or research to identify any other population groups to be included in such evaluations for fairness. Therefore, we recommend that in K-12 assessments, testing programs should, where possible, report disaggregated statistics for native English speakers, ELLs, and former ELLs, so that the distributions of scores for these groups can be evaluated. Programs should also review differences in scores across testing variations (types of accommodations and test modifications). Whenever appropriate, programs should report analyses for test variations commonly employed with ELLs. These include:

- language of assessment, translated versions of the test or dual language booklets (e.g., English vs. Spanish),
- linguistically modified (or plain English) versions of tests, and
- extended time, reading aloud instructions, and use of bilingual glossary.

Differential Impact For each studied group (or test variation, if appropriate), the following statistical information can provide evidence regarding the validity of an assessment for different examinee groups:

- Performance of studied groups. Provide statistics about the performance of studied groups on the whole test, subtests, and items. Group differences in the distribution of scores and item and test statistics are worthy of investigation in order to determine the underlying causes of these differences.

For the test and, if appropriate, for subtests, compute score distributions and summary statistics—means, standard deviations, selected percentiles (the 10th, 25th, 50th, 75th, and 90th)—and percentages of students in each achievement level. This section assumes familiarity with psychometric and statistical concepts. Guidelines for the Assessment of English Language Learners www.ets.org26

For individual items, report item difficulty, item-test correlations, and item characteristic curves.

- Differential item functioning (DIF). Report DIF statistics, if sample size allows, using ELLs as the focal group and non-ELLs as the reference group. If sample sizes allow, DIF results could also be reported using former ELLs as the focal group. Examine test items that are flagged as exhibiting DIF against one or more examinee groups in order to identify the possible causes, which can be useful in making decisions about possibly removing items from scoring.

• Differential predictive validity. Report statistical relationships among reported scores on tests and subtests and criterion variables (such as scores on other tests given in later years) for ELLs and non-ELLs. Gather information about differences in prediction as reflected in regression equations, or differences in validity evidence for studied groups. Evidence of differential predictive validity indicates that the test functioned differently for different examinee groups and suggests that further investigations into the construct validity of the test for all groups may be warranted.

Reliability - To investigate whether scores are sufficiently reliable to support their intended interpretations, the following statistics for each of the examinee groups are particularly informative:

• If sample size permits, provide the following for reported scores, subscores, and cutscores (if available): Reliability estimates (accounting for a variety of sources of measurement error), information functions, index of classification consistency (consistency of the pass/fail decisions based on cutscores), standard error of measurement (for raw and scaled scores), and conditional standard errors of measurement around cutscores.

• When comparing test reliability across studied groups, evaluate differences in group dispersion (for example, ELLs may be more homogeneous than non-ELLs). If reliability coefficients are adjusted for restriction of range, provide both adjusted and unadjusted coefficients.

• For scoring constructed responses, follow the ETS Guidelines for Constructed Response and Other Performance Assessments (i.e., estimate inter-rater reliability for individual items). Since ELLs' writing skills in English are in most cases lower than those of English-proficient www.ets.org Guidelines for the Assessment of English Language Learners students, evaluate whether there are interactions between rater scoring and ELL membership. Validity The ETS Standards for Quality and Fairness recommend gathering evidence about whether a test is measuring the same construct(s) across different subpopulations. These standards also indicate that, if the use of an assessment leads to unintended consequences for a studied group, the testing program should review validity evidence to determine whether the consequences arose from invalid sources of variance—and, if they did, revise the assessment to reduce, to the extent possible, the inappropriate sources of variance. For ELLs as well as non-ELLs, some methods for investigating validity include:

• Analyses of internal test structure. Report statistical relationships among parts of the assessment (e.g., intercorrelations among subtests, item test correlations, dimensionality and factor structure). • Relations to other variables/constructs. Report statistical relationships among reported scores on the total test and subtests and with external variables. • Test speededness. Because of ELLs' lower reading fluency, test time limits may affect their performance disproportionately relative to non-ELLs. For timed tests, evaluate the extent to which there are differential effects of test speededness on ELLs. Report the number of items not reached and omitted for each examinee group.

2.2. Alternative Assessment and Its Types

Alternative assessment is a blanket term that covers any number of alternatives to standardized tests. While the traditional paper and pencil tests may be effective to assess some of the skills (such as listening), they are not sufficient to assess the productive skills of speaking and writing. The nature of proficiency-oriented language learning calls for a variety of assessment options reflecting the numerous instructional strategies used in the classroom. Authentic assessment, performance-based assessment, and portfolio fall under this category. Assessment, Articulation, and Accountability, 1999. Authentic Assessment Authentic assessment is an alternative assessment, it integrates the assessment of traditional academic content with the knowledge and skills important to lifelong learning using a variety of techniques, including "real world" situations. (McREL Institute, 1993). For an assessment to be authentic, the context, purpose, audience and constraints of the test should connect in some way to real world situations and problems. Tests can be authentic. Performance-based assessment is an alternative assessment, it ‘‘requires students to construct a response, create a product, or demonstrate application of knowledge’’ in authentic context (Authentic Assessment for English Language Learners, p. 239). Performance-based assessment requires the development of new assessment tools and scoring rubrics. Regardless of the terms, these assessments share several commonalties.

1. They are alternative to the traditional testing.

2. They involve some kind of performance or tasks relating to real-life situations.

3. They are process-oriented.

Following are some reasons for incorporating alternative assessment in the foreign language classroom.

o To capture complex outcomes. Alternative assessment goes beyond the assessment of knowledge and facts to the more complex goals of assessing and developing life-long skills of creative thinking, problem solving, summarizing, synthesizing, and reflecting. With authentic assessment, products and processes are equally valued. Assessment, Articulation, and Accountability, 1999

o To address realistic tasks. With authentic and performance-based assessments, students are involved in tasks, performances, demonstrations, and interviews reflecting everyday situations within realistic and meaningful contexts.

o To include good instructional tools. Assessment and instruction interact on a continuous basis. Assessment can be used to adapt instruction and to provide feedback for monitoring students’ learning. Alternative assessment focuses on the students’ strengths, therefore enabling the teacher to get a more accurate view of students’ achievement, of what they can do, and of what they are trying to do.

o To communicate what we value. Assessment and instruction need to be aligned. If we value oral proficiency but only assess through written tests, students infer that only the written language matters.

o To meet the students’ different learning styles. Alternative assessment offers a broad spectrum of assessment possibilities to address the different learning styles. Some students might choose to demonstrate understanding by writing about something while others might prefer to perform, to display visually, or to create a timeline.

o To collaborate and interact with students. Even though schools usually focus on students working alone, the real world allows and encourages people to talk, ask questions, get help and receive feedback. Denying students the right to cooperate and collaborate diminishes the authenticity of the achievement.

Following is a list of possible alternative assessments.

o Performance-based assessments (projects, exhibitions, role playing, experiments and demonstrations) o Open-ended questions

o Writing samples o Interviews o Journals and learning logs

o Story of text retelling o Cloze tests o Portfolios

o Self and peer assessments o Teacher observations

o Checklists

While objective tests are easy to grade, authentic and performance-based assessments require a more subjective judgment on the part of the teachers. Thus, criteria and rubrics need to be developed prior to the students’ assignments (see section on rubrics). These criteria define the standards for acceptable performance and can be used for self-assessment, peer evaluation, or teacher evaluation. Wiggins (1994) feels that making the criteria clear and de-mystifying them helps foster self-assessment.

The following principles, adapted from the ‘‘Principles of Assessment’’ developed during a symposium at the Center for Applied Linguistics, are based on the assumption that the purpose of language instruction is to prepare students to use language with cultural understanding and knowledge in real-life contexts. While these principles were developed for early foreign language learning, they apply to language instruction. Assessment, Articulation, and Accountability, 1999

161. The purposes for assessment should be clear. The purposes determine the frequency and types of assessment to be done.

1. Assessment should be tied to curricular practices that are informed by second language theory and research and should support the goals and objectives of the instructional program as determined by the school, the district, and the state.

2. Assessment should be developmentally appropriate.

3. Assessment should reflect student progress along a continuum from simple to progressively more complex tasks. The designed tasks should be curriculum-embedded and part of the teaching/learning process.

4.Assessment should be both formative (to continually assess the degree to which short-term objectives have been met so that ‘‘fine tuning’’ can occur with instruction and learning) and summative (to assess the degree to which final objectives have been met).

5. Assessment should allow students to demonstrate their ability to function in a variety of cognitively engaging tasks. When assessment is performance-oriented, the students’ work will result in a product, written or oral performance, or problem solving.

6. Assessment employs a broad range of data-gathering methods over time and should be based on multiple sources of evidence. Multiple sources of evidence allow for assessing a student’s progress in many areas and also take into account the different learning styles.

7. Assessment should be conducted regularly and frequently.

8. Assessment is authentic in that it reflects what students do in the classrooms on a regular basis. The relationship between instruction and assessment is constant. Teachers should assess the effectiveness of what they teach and use the results to improve instruction.

9. Assessment activities should be contextualized and include practical contexts and culturally appropriate situations.

10. Assessment should encourage students to reflect on their own progress. For this reason, it is essential to design assessments which are learner-centered and

to share assessment expectations with students. Assessment, Articulation, and Accountability, 1999

11. Assessment results should be reported in a variety of ways depending on the target audience. 12. Educators should use assessment efficiently so that demands on instructional time are reasonable.

A wide range of factors should be taken into account when assessing students' language, including their maturation level, learning styles, learning disabilities, physical disabilities, and other characteristics affecting their performance.

2.3. Evaluating the Tasks Through Tryouts

Trying out or field-testing items can provide extremely useful information during the test development process. When conducting an item tryout, use a sample of examinees similar to those who will take an assessment once it is administered operationally (for official score-reporting purposes). This step is particularly important for items that will be used with ELLs. Purposes of Item Tryouts There may be several reasons to conduct item tryouts. Data may be collected in order to:

- inform decisions about how appropriate the items are for a sample of examinees similar to the operational population,
- inform content and fairness reviews of the items,
- evaluate timing requirements for new or existing item types,
- evaluate the clarity of instructions to examinees,
- support the scaling or equating of test forms,
- inform the standard setting process by providing performance data, which panelists will receive as feedback on cutscores, on different groups, and
- assess whether ELLs of different proficiency levels can understand the text of the items. This is important when English language proficiency is not the

construct of interest. Guidelines for the Assessment of English Language Learners www.ets.org. Types of Item Tryouts Item tryouts may take several different forms, ranging from one-on-one interviews with students, through small-scale pilot tests, to large-scale field tests. As with other activities described within these guidelines, it may not be possible to implement each of these types of item tryouts in a given testing program because of resource constraints. However, we describe them here so that readers can make informed decisions about when and whether each type may be useful.

One-on-one interviews with students who have been administered the items can provide much useful information. These interviews can take the form of informal debriefings after students have completed the tasks, or more formal cognitive laboratory activities where students are interviewed either while they are answering the questions or afterward. Because individual interviews are time-consuming to conduct, it is usually not possible to involve large numbers of students. The information that such interviews yield can sometimes be idiosyncratic. However, the quality and type of information interviews provide can offset that concern. Interviews allow students to talk about the cognitive processes they employed when answering the item, whether anything confused them, and how they arrived at their answer. The interviewer can also ask students what they think the item is asking them to do or what they think the item is measuring. Qualitative summaries of this feedback can be very helpful for the item review process. This type of item tryout is particularly important for items that will be used with ELLs, since the interviewer can ask them directly about their understanding of and response to the items. For ELLs, interviews are extremely useful for identifying potential threats to the validity of tests that measure knowledge in content areas other than English language arts. To determine whether items require a high degree of English proficiency unrelated to the construct, it is important to assess ELLs’ understanding of the language of the items. While external reviewers with expertise in ELL issues can provide valuable insights, working directly with ELLs to gather their impressions of test materials can generate even more detailed and useful information. Since one-on-one interviews may be costly and time-consuming, it may not be possible to conduct them as part of an ongoing testing program. They may be most useful when trying out a new item type. Testing officials will need to decide whether the information they may gain from these interviews is worth the time and expense. www.ets.org Guidelines for the Assessment of English Language Learners. Small-Scale Pilot Tests Small-scale pilot tests may also provide useful information on how students respond to the items. In this data collection format, test developers administer the items to a larger sample of students than is used for one-on-one interviews, and, generally, one-on-one debriefing does not take place. Because these samples may not be fully representative of the test-taking population, the item statistics provide only a gross measure of whether students were able to answer the item correctly. Including a small-scale pilot with an oversampling of ELLs may prove very helpful during the item development process to discover issues specific to ELLs. Again, however, budgets and schedules may not allow for these types of pilot tests to take place. Such activities may be most appropriate when introducing a new item type. Large-Scale Field Tests In large-scale field tests, test developers administer the items to a large, representative sample of students. Because of the size and nature of the sample, statistics based on these responses are generally accurate indicators of how students may perform on the items in an operational administration. If the tryout items are administered separately from the scored items, motivation may affect the accuracy of the results. When the tryout items are embedded among the scored items, students do not know which items count and which do not, so motivation is not a factor. Consequently, many states conduct embedded field testing and are increasingly moving toward placing the tryout items in random positions within each test form. Conducting a large-scale field test on a group in which ELLs are well-represented will allow for the evaluation of item difficulty and other item characteristics specific to ELLs. Guidelines for Item Evaluation The type of tryout should be tied to the goals of the evaluation. To obtain information directly from students about their thought processes while answering the items, conduct one-on-

one interviews. To obtain information directly from ELLs about their understanding of complex language in items measuring content areas other than English language arts, conduct one-on-one interviews. Evaluate the extent to which complex language generates comprehension difficulties for ELLs relevant to the construct being measured. If there appears to be unnecessary linguistic complexity, review the item and revise it as appropriate before the operational administration. Field test it again if necessary (for example, in the case of pre-equated tests). To inform judgments about how items will work, conduct a small-scale pilot test—but remember that the data from such pilots usually does not come from a representative sample. To Guidelines for the Assessment of English Language Learners www.ets.org18 obtain reliable and valid statistics that can be used when selecting items for test forms or equating, conduct a large-scale field test. Try items out on a sample that is as similar as possible to the population that will take the operational administration. However, oversampling ELLs during pilot testing is recommended; such oversampling increases the likelihood of uncovering issues that may be specific to those students. Document the procedures used to select the sample(s) of examinees for item tryouts and the resulting characteristics of the sample(s). Try out all item types, including both selected-response, constructed-response, and hands-on tasks or activities. If constructed-response items are tried out, score them using scorers and procedures that are as similar as possible to those used for operational administrations (but consider possible security risks engendered by exposing prompts before the administration). Evaluate responses to constructed-response items according to the following criteria, per the ETS Guidelines for Constructed-Response and Other Performance Assessments:

- Do the examinees understand what they are supposed to do?
- Are the tasks appropriate for this group of examinees?
- Do the tasks elicit the desired kinds of responses?
- Can the responses be easily and reliably scored?
- Can they be scored with the intended criteria and rating scale?

• Are the scorers using the scoring system in the way it was intended to be used?

To ensure accessibility for ELLs, it is also important to ensure that rubrics focus on the construct of interest and do not include construct-irrelevant variance by placing inappropriate emphasis on English language proficiency unrelated to the construct. For example, scoring rubrics should state clearly that, when English language proficiency is not defined as part of the construct, raters should ignore errors in English when scoring for content. For more information, see Scoring Constructed-Response Items. Limitations of Item Tryouts Even when field test samples and operational populations seem comparable, differences in demographics, curriculum, and culture may make comparisons difficult. Document any limitations of the representativeness of the field test sample. Such limitations are most likely to be present for one-on-one interviews and small pilot samples. Motivational level can also be a factor as field test participants are often not as highly motivated to do their best as are operational examinees. In sum, www.ets.org Guidelines for the Assessment of English Language Learners field testing is valuable for trying out new tasks and scoring criteria, but use the results of field testing with caution for higher-stakes decisions such as setting the standard for passing the assessment.

2.4. Testing Accommodations for English Language Learners

Purpose of Testing Accommodations for English Language Learners The main purpose of providing examinees with testing accommodations is to promote equity and validity in assessment. For ELLs, the primary goal of testing accommodations is to ensure that they have the same opportunity as students who have English as their first language to demonstrate their knowledge or skills in a content area. Reducing or eliminating construct-irrelevant variance from the testing situation increases the likelihood that score users will be able to make the same valid interpretations of ELLs' scores as they make for other examinees. In general, the main sources of construct-irrelevant variance on content area assessments for ELLs are the effects of English language proficiency in answering test items. Unless language proficiency is part of the construct being measured, it should not play a major role in whether an examinee can answer a test item correctly. Accommodations refer to changes to testing procedures, which researchers have traditionally considered to include presentation of test materials, students' responses to test items, scheduling, and test setting. As a general principle, testing accommodations are intended to benefit examinees that require them while having little to no impact on the performance of students who do not need them. At present, the research basis regarding which accommodations are effective for ELLs under what conditions is quite limited. Relative to research on students with disabilities, research on accommodations for ELLs has a much shorter history, with the results from studies often seeming to contradict each other. Some state policies distinguish between testing accommodations (changes in the assessment environment or process that do not fundamentally alter what the assessment measures) and testing modifications (changes in the assessment environment or process that may fundamentally alter what the assessment measures) and refer to both as testing variations. In these guidelines, the term testing accommodation refers to changes that do not fundamentally alter the construct being assessed. Identifying Students Eligible for Accommodations Policies for identifying ELLs who may be eligible for testing accommodations continue to evolve. At present, there are no uniform guidelines or policies at the federal level regarding the use of accommodations for ELLs. For students with disabilities, eligibility for accommodations is part of a student's Individualized Education Plan (IEP); however, ELLs do not have any corresponding documentation. Across states and local school districts, both the eligibility requirements as well as the www.ets.org Guidelines for the Assessment of English Language Learners23 specific accommodations available to ELLs vary widely. In fact, some policies are not transparent with respect to how eligibility for accommodations is determined or who is making the decisions for ELLs. As a general principle, if an ELL's English language proficiency is below a level where an assessment administered in English would be considered a valid measure of his or her content knowledge, then that student may be eligible for one or more testing accommodations. Typically, ELLs who regularly use accommodations in the classroom are usually eligible to use the same accommodations in testing situations. However, some accommodations that may be appropriate for instruction are not appropriate for assessment. For example, some ELLs routinely have text read aloud to them as part of instruction. But if decoding or reading fluency is being assessed as part of reading comprehension, this would not be an appropriate accommodation because it would change the nature of the assessment from one of reading comprehension to one of listening comprehension. Further, an accommodation such as the use of a native language glossary of terms that could be appropriate for certain subjects such as mathematics or science would not be appropriate for English language arts, because the use of a glossary would change what is being assessed and would provide an unfair advantage to those who have access to it. Identifying Accommodations Testing accommodations for ELLs can be broadly grouped into two categories: Direct linguistic support accommodations (which involve adjustments to the language of the test) and indirect linguistic support accommodations (which involve adjustments to the conditions under which a test is administered). To be ELL-responsive, an accommodation should provide some type of linguistic support in accessing the content being tested. To date, the limited number of research studies on accommodations for ELLs indicates that direct accommodations appear to benefit student performance more than indirect accommodations. Examples of direct linguistic support accommodations include providing a translated or adapted version of the test in the student's native language or providing test directions orally in the student's native language. The use of translated tests is a complex issue because questions can arise as to whether the original and translated versions are measuring the same construct in the same manner. Translated versions of items may or may not have the same meaning as in their original versions. Therefore, some educational agencies have created transadapted versions of tests, which are translated versions of tests that have been culturally adapted for the examinees. Furthermore, the use of translated tests may only be of limited benefit to examinees, particularly if the language of instruction and the language of the test are not the same. Furthermore, unless a test can be translated into all of the native languages spoken by the students in a school district or state, questions of equity may arise. In addition, in some states, public policy may prohibit the assessment of students in languages other than English. Examples of indirect linguistic support accommodations include extended testing time or having the test administered individually or in small groups. Some of these accommodations do not address construct-irrelevant variance due to language; however, they may be useful or necessary to facilitate test administration for ELLs or for all students. Because state and local policies are evolving at a rapid pace, we have not provided with these guidelines a complete list of accommodations that state or local school districts allow for ELLs. Test developers and interested readers should contact the appropriate educational agencies to obtain the most current assessment policy and list of accommodations available to ELLs. Some states have simply extended to ELLs the use of accommodations originally intended for students with disabilities. However, some of these accommodations are clearly inappropriate when applied to ELLs (such as the use of large print versions of tests, which are appropriate only for students with a relevant disability such as a visual impairment). Recent reviews indicate that fewer than two thirds of the accommodations for ELLs found in states’ assessment policies address the unique linguistic needs of ELLs exclusively.

At present, there are no existing standards that can definitively guide the use of testing accommodations for ELLs. The appropriate use of accommodations depends on a number of factors including: a student’s proficiency in English as well as his or her native language, the academic subjects being assessed, the student’s familiarity with the accommodations, the language in which the student receives instruction, and the range of available accommodations for examinees. To the extent practical, decide on accommodations for individual students, not as a collective group. The accommodation or combination of accommodations that may be most appropriate for one ELL may or may not be the best choice for another student.6 Within the past decade, some progress has been made in developing systems for making decisions on testing accommodations for ELLs, but additional work is necessary before any of these systems are ready for use by administrators or teachers. Currently, without sufficient research findings to inform appropriate use of accommodations for ELLs, accommodation decisions are best guided by the following operating principles: Most importantly, accommodations for ELLs should not alter the construct being assessed; this is particularly critical when students are tested on their academic content knowledge and skills. In 6 Status as an ELL is much more dynamic than disability status or cognitive status, and a student’s ELL proficiency level may change from one year to the next. For this reason the student’s need for a given accommodation may change from one year to the next due to increased English language proficiency. This means ELLs should receive the greatest degree of linguistic support accommodations—such as a glossary or bilingual dictionary—necessary in order to ensure this outcome.

2.5. Different Test-taking Skills

Testing skills (also known as test taking strategies) are often seen as a sneaky shortcut learners can use to get around their lack of English. However, on a well-designed standardized test it's actually not possible to spoof your way into a score significantly higher than your actual ability.

That being said, I do believe test taking skills are a valid and important part of an effective exam preparation course.

The reason for this is that language ability is not the only challenge learners face when taking a standardized test. The design and format of the test, the way the information is presented, and especially the timed nature of the questions all present additional difficulties for learners. In fact, after several decades of being involved with testing and test preparation, I am convinced that these factors cause a very large percentage of test takers to significantly underperform relative to their actual language knowledge and ability.

So what are test taking skills? In a nutshell test taking skills are nothing more than effective language skills (reading, writing, listening, speaking) that are chosen to overcome challenges related to specific aspects of the test design or timing. These include:

• *Familiarity with the test format, instructions and question types* – Although on most standardized tests the instructions are printed and exemplified at the start of each test part, familiarity with these will help the test taker avoid confusion on exam day.

• *Time management* – The tight timings found on most standardized tests are one of the most challenging factors for many learners. Those not accustomed to dealing with a lot of language in a short time often have difficulty completing all the questions. Encouraging learners to monitor their time carefully and teaching skills such as skimming (quick reading to get a general idea) and scanning (quick reading to pick out specific details) can help learners budget their time effectively and allow them to allocate more time to those places with the best potential score payoff.

• *Efficiency of information processing* – The lengthy listening and reading passages commonly found on standardized tests make it essential that learners are able to read and listen efficiently within the allotted time. This can include (where possible) previewing the questions and answer choices before listening/reading to allow them to predict what they will hear/see. Pre-reading the questions also allows them to focus their attention only on what is needed to answer the question.

• *Awareness of features that can make incorrect answer choices attractive* – With any multiple-choice test, incorrect answer choices are designed so as to be in some way attractive. Being aware of the common forms these ‘distracters’ take will enable learners to avoid them and choose the correct answer.

Developing test taking skills maximizes the chances that learners will be able to fully demonstrate the extent of their English ability. Time spent working on these areas can result in significant gains in a reasonably short time, especially with learners unfamiliar with the test format. For example, with inexperienced test takers, score increases of up to 100 points on the TOEIC are not unusual within the span of a typical 20 hour course.

Linguistic skills

Although significant short-term gains can be made by focusing on test taking skills, once a learner has become familiar with the test format and the related approaches to deal with its challenges, further significant progress requires a major increase in overall language competence.
Key areas for linguistic development include:

• *Understanding language in use* (conversational English) – The English used in standardized tests like the TOEIC, TOEFL and EIKEN reflects everyday usage as encountered in offices, on the street, or (in the case of the TOEFL) in academic situations. An important element of this is understanding the ways native speakers appropriately handle such common functional situations as requests, complaints, suggestions, etc. For example in Part 2 of the TOEIC we may hear a question based around a conversation like this:
Q - *"Can you help me fill out this form?”*
A - *“Sure, it's actually pretty simple.”*

In this example there are no clear grammatical or lexical links between the initial question and the (most appropriate) response. In order to do well on questions like this, learners must be familiar with the common ways that native speakers make and respond to requests, and other high-frequency language functions.

• *Familiarity with different native speaking accents* – The listening sections of the TOEIC and TOEFL tests include US, Canadian, British, and Australian speakers. Due to their past learning experiences many learners develop a bias for a given style of English and have difficulty understanding speakers of other dialects. To overcome this it is important that learners be exposed to English from a variety of English-speaking countries.

• *Awareness of the sound changes that occur in natural English speech* – Many learners are unaware that the sound of words spoken in natural conversation can differ dramatically from words spoken in isolation (e.g. 'going to' often sounds like 'gonna' in natural spoken English). In the past, learners may have had instructors who shielded them from exposure to such natural language on the grounds that it was only suitable for advanced learners, or even that it was representative of slang or 'lazy' English. This has led to many learners finding it very hard to follow natural conversation. Helping learners become aware of the ways that sounds are combined, dropped and changed in natural speech can significantly improve their listening comprehension on standardized tests.

• *Vocabulary and grammatical understanding* – Vocabulary is arguably the single most significant factor in doing well on standardized tests, both in terms of individual words, and multiword groups such as stock phrases and expressions. Any effective study program should include an organized system for noting and reviewing words and phrases that learners encounter as they study. Grammar, though of course an important aspect of language, has tended in the past to be somewhat over emphasized in test preparation courses. Because the current trends in standardized testing are moving away from an overt focus on accuracy, I generally devote little class time to discrete grammar study.

The bottom line on helping learners improve their scores on standardized tests is that there is no single ideal approach. In order to make significant gains learners must work on a number of separate but highly interrelated areas.

Conclusion

Thus, testing and evaluation are very useful in the preparation of language teaching materials as well as after the actual teaching has taken place. These tests are used to place the students into categories as well as to judge the problems of teaching. On the basis of these tests, language teacher focuses his or her attention towards the areas of difficulties which the learner faces in any language teaching programme and these areas of difficulties can be predicted by the effective use of language tests. So, without effective testing no language teaching programme can be successful. We must mention here another important matter: a comprehensive language testing covers all the levels of Linguistics such as phonology, morphology, syntax, lexicon, grammar and semantics and without the knowledge and application of Linguistics these areas cannot be tested properly. A linguistic approach to language testing, therefore, is an approach, which makes use of the theoretical knowledge of Linguistics. For instance, for testing the listening and speaking skills, the knowledge of the phonological system of the target language is essential. Only then appropriate tests can be constructed and the required skill can be tested properly.

So why should we bother ourselves thinking about testing and assessment?

Well the obvious answer for many students and teachers alike, is that we are forced to. Most test takers take these tests because they absolutely need the scores either for school or work. Failure to get the desired grade can result in the student being barred from the university of their choice, or not getting a desired job or promotion. Similarly, most teachers end up teaching these types of courses because their faculty head or Director of Studies assigns them to teach the class, rather than out of personal choice. In fact, these sorts of test preparation courses are often seen as boring, ineffective and generally unrelated to the practical real-world skills they purport to assess. Sadly this perception is often not far from the truth. Many test prep courses are approached with a mind-numbingly large number of practice questions, followed by a painfully detailed breakdown of errors, and often with a hefty dollop of discrete grammar to cap it off.

It's for these reasons I decided to take on this column. First of all, the stakes for test takers are so high, and the typical expenditure of time and money so significant, that the need to make the process as quick and efficient (and successful!) as possible becomes obvious. Secondly, given the amount of time that teachers and students spend preparing for these tests, I think we really owe it to all concerned to take a long hard look at whether the currently common, painful and often demotivating study methods are in fact, the best way for students to get the scores they need.

The sort of methodology described above is absolutely not the best way to approach the situation, and we plan over the course of this column to present some alternative approaches that are not only more effective at raising test scores, but will also help to make test prep classes more motivating, practical and yes, even fun! I don't claim however, to hold the final word on the issue of testing and assessment, and definitely welcome comments, input and alternative ideas from readers.

Bibliography

1.Allen, J.P.B. and S. Pit Corder (eds.) 1974. Techniques in Applied Linguistics: the Edinburgh course in Applied Linguistics. London: Oxford University Press. Vol. 3.

2.Bachman, L.F. 1992. “Assessment”. In Bright, W. (ed.) International Encyclopedia of Linguistics. London: Oxford University Press. Bhat, R. 1992. “Role of testing in Language Teaching”.

3.In Khan, I.H. and Hasnain, S.I. (eds.) Linguistics and Language Teaching. Department of Linguistics, AMU, Aligarh. Carrol, J.B. 1965. “Fundamental Consideration in Testing for English Language Proficiency of Foreign Students”.

4.In Allen, J.P.B. (ed.) Teaching English As a Second Languages. New York: McGraw Hill, Inc. Davies, A. 1977. “The construction of language tests”.

5.In Allen, J.P.B. and Davies, A. (eds.). The Edinburgh Course in Applied Linguistics: Testing and experimental methods. Vol. IV. London: Oxford University Press. Halliday, M.K.A. et al. 1966. The Linguistics Science and Language Teaching. London: Longman.

6.In Allen J.P.B. and Corder, S.P. (eds.). Techniques in Applied Linguistics: The Edinburgh Course of Applied Linguistics. Vol. 3. London: Oxford University Press. Paterno, A. 1965. “Foreign Language Testing”.

7.In Allen, J.P.B. (ed.) Teaching English As a Second Languages. New York: McGraw Hill, Inc.

8.Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. Educational Assessment, 8, 231-257.

9.Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. 10.M. Haladyna (Eds.), Handbook of test development (pp. 377-398).

11.Abedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large□scale assessment: Interaction of research and policy. Educational Measurement: Issues and Practice, 25(4), 36-46.

12.Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. Review of Educational Research, 74, 1-28.

13.Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. Applied Measurement in Education, 14, 219-234. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999).

14.Standards for educational and psychological testing. Washington, DC: American Psychological Association.

15.Bailey, A. L. (2007). The language demands of school: Putting academic English to the test. New Haven, CT: Yale University Press. Educational Testing Service. (2002).

16.ETS standards for quality and fairness. Princeton, NJ: Author. Educational Testing Service. (2003).

17.ETS fairness review guidelines. Princeton, NJ: Author. Educational Testing Service. (2006). ETS guidelines for constructed-response and other performance assessments. Princeton, NJ: Author.

18.Kopriva, R. J. (2008). Improving testing for English language learners. New York: Routledge. Kopriva, R. J., Emick, J. E., Hipolito-Delgado, C. P., & Cameron, C. A. (2007). Examining the impact of improved decision making on scores for English language learners. Educational Measurement: Issues and Practice, 26(3), 11-20.

19.Martiniello, M. (2008). Language and the performance of English language learners in math word problems. Harvard Educational Review, 78, 333-368. Martiniello, M. (in press). Linguistic complexity of math word problems, schematic representations, and differential item functioning for English language learners (ETS Research Report).

20.Princeton, NJ: Educational Testing Service. Rabinowitz, S. N., & Sato, E. (2006). The technical adequacy of assessments for alternate student populations: Guidelines for consumers and developers.

21.San Francisco: WestEd. Rivera, C., & Collum, E. (Eds.). (2008). State assessment policy and practice for English language learners: A national perspective. Mahwah, NJ: Erlbaum.

22.Thurlow, M. L., Thompson, S. J., & Lazarus, S. S. (2006). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 653-673).

23.Mahwah, NJ: Erlbaum. Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. Educational Assessment, 13, 170-192.

24.Young, J. W., & King, T. C. (2008). Testing accommodations for English language learners: A review of state and district policies (College Board Research Report No. 2008-6; ETS Research Report No. RR-08-48). New York: College Entrance Examination Board.

Appendix A Glossary - assessment in language learning

Achievement tests Tests which look back over a long(ish) period of language learning to test how much of the syllabus has been learnt. Internal end-of-year and external school-leaving examinations are both examples of achievement tests. Also called **attainment** tests.

Aptitude tests Tests which are usually given before any of the foreign language has been learnt in order to discover which potential learners are likely to be good at learning languages. Therefore aptitude tests have to attempt to measure probable language learning ability in the future rather than actual learning achievement in the past. This is done by identifying factors which are likely to contribute to success in language learning, e.g. sound discrimination, memory, etc.

Attainment tests See **achievement tests.**

Backwash effect. See **washback effect.**

Battery tests Tests made up of large number of (usually) discrete-point items or a series of sub-tests assessing different aspects of language performance. The purpose of such tests is to take a wide sample of language so as to assess learners' overall language proficiency.

Bottom-up processing

Bottom-up processing refers to the readers' or listener's ability to interpret the language of the text, i.e. the words, grammar, discourse markers and sounds. The other component in processing is **top-down** processing.

C-test

A form of **cloze** in which the second half of every second word is deleted from the text.

Can-do statements

See **descriptors** and **scale.**

Cloze tests

Tests in which words are deleted from a text. Learners must use a range of language skills to restore the deleted words. In conventional cloze tests, the words are deleted at regular intervals – every 6th, 7th or 8th word, typically. In **modified cloze** the tester deletes the items s/he is interested in testing, e.g. all the articles or prepositions. Cloze has been used as a test of reading or listening (by deleting words from a spoken test) but is now used mainly as a test of grammar or overall language proficiency. See also **C-test**

Concurrent validity

This asks whether a new test produces results which are similar to those produced by an established test or a teacher's predictions. If it does, it is said to have good concurrent validity. Concurrent validity is assessed by correlating results from the experimental test and an established test or a teacher's rankings. See also **validity.**

Construct validity

This asks whether a test is theoretically sound, i.e. is it constructed according to sound theories of language and language learning? Obviously a test devised according to audio-lingual or structural theories is unlikely to have good construct validity if one believes in communicative language learning. This can be a major problem if teaching and testing are based on different theories or ‘constructs’. Construct validity is a matter of personal opinion and cannot be measured. See also **validity.**

Content validity One of the criteria for judging whether a test is good or not. It refers to the extent to which a test actually assesses relevant language content. Relevance may be determined in two broad ways: backwards-looking relevance (i.e. whether the test assesses the language taught on a syllabus) and forwards-looking relevance (i.e. whether the test assesses the language needed in the target situation). Content validity is a matter of personal opinion rather than measurement. See also **validity.**

Criterion-referenced testing Tests in which learners are assessed according to a criterion rather than comparing them with other learners. For example, eye tests are criterion-referenced – can you read certain print or not? In language testing, the criteria might be, for example, can learners produce correct past tense forms? business letters? acceptable consonant sounds? See also **norm-referenced testing** and **descriptors.**

Descriptors Brief descriptions of real-world language performance, usually arranged on a **scale.** Learners’ language proficiency is described in terms of what they can do at each stage on the scale, and so descriptors are also known as ***can-do* statements.** Descriptors are normally used with **criterion-referenced testing.**

Diagnostic tests

Tests which look back over a long(ish) period of language learning to establish how effective that learning has been. The purpose of diagnostic testing is to establish areas of weakness or deficiency, so that future learning/teaching can focus on these areas. Traditional diagnostic tests focused on vocabulary, grammar and pronunciation, but communicative diagnostic tests usually use **self-assessment** to determine areas of weakness such as speaking on the telephone, writing letters, reading journals or listening to announcements.

Dictation

A testing technique in which the tester/teacher reads a text in short sections and the learners then write down the spoken text as accurately as possible. Originally intended as a listening test, dictation has been used more recently as a test of **integrative** language proficiency.

Direct testing

Testing which consists of assessing real-world language performance. For example, asking learners to write a business letter is a direct test of this real-world language task, whereas asking them to answer a series of multiple-choice grammar questions (for example) is an **indirect test** because this is not a real-world language task.

Dirty testing

Testing which makes only a quick, approximate estimate of learners' language level, often by assessing only a narrow range of language skills. **Placement tests** are often 'dirty'.

Discrimination The ability of a test to separate strong students from weaker students. Most **norm-referenced tests**, especially **achievement tests**, should discriminate well, whereas **diagnostic tests** need not have this quality.

Discrete-point 'Discrete' means 'separate, single, distinct'. A discrete-point test, therefore, is made up of a large number of separate items, each testing a different part of the language, for example grammar points. **Battery tests** are usually discrete-point tests.

Distracter See **multiple-choice.**

Dual-choice A test item which offers learners two choices, e.g. true/false, same/different. Such items are easy to construct but learners have a 50% chance of guessing the right answer and so most public tests prefer **multiple-choice** items with four or more distracters, but classroom tests may make use of dual-choice items.

Face validity This asks whether a test seems like a good and relevant test of language. Face validity is established by relating the test either to the syllabus (does it test language which has been taught?) or to the real world. Learners who complain that a test is irrelevant to their needs are complaining about its face validity. Face validity is a matter of opinion rather than measurement. See also **validity.**

Formative Formative tests are administered during the process of

assessment language learning and are designed to assist that process. **Placement, diagnostic** and p**rogress** tests are all formative. Contrasted with **summative tests.**

Fragmentary comprehension Comprehension of isolated details of a text rather than its overall message. Some texts (e.g. telephone directories, dictionaries) are fragmented and so we want only fragmentary comprehension when we refer to them. Other texts (newspaper stories, novels) demand **global comprehension**. Fragmentary comprehension can be assessed through **discrete-point** items; global comprehension requires global or **holistic** items. Whether a test assesses fragmentary or global comprehension depends on the original purpose of the text.

Free-response item A test item which elicits an uncontrolled sample of language, i.e. the learner can say/write whatever s/he wishes within the constraints of the test question. Writing tests such as compositions are usually free-response. All free-response items are **subjective** in their scoring.

Gate-keeping Refers to the function of certain tests, e.g. **placement** and **proficiency** tests, to permit or restrict access to opportunities, e.g. entry to further education, employment or immigration.

Global comprehension See **fragmentary comprehension.**

Global item Another term for an **integrative** item; the opposite of a **discrete-point item.**

Global testing Another term for **integrative testing.**

Holistic comprehension The opposite of **fragmentary comprehension.**

Indirect testing The opposite of **direct testing**. Indirect tests sample a learner’s underlying language competence or knowledge (e.g. mastery of the grammar or other formal parts of the system) rather than its application in the performance of real-world language tasks. Most ‘language’ tests, as opposed to communicative tests, are indirect. Indirect tests lend themselves to **discrete-point items** and **objective** scoring.

Information transfer An integrative testing technique which requires the learner to transfer information from a verbal to a non-verbal/visual form. The transfer can work in both directions: transferring information from a non-verbal source to a verbal output (Describe the person in this picture) is composition; transferring information from a verbal source to a non-verbal output (Draw a picture of the person described in this text) is comprehension.

Integrated testing Tests which assess two or more of the traditional language skills (reading, writing, listening, speaking) in a single item. Integrated testing is generally regarded as bad practice because of the danger of penalising the learner twice but it is used in some EAP tests (e.g. 1 *Make notes on the following three texts about witches.* 2 *Use the notes that you made earlier to write 500 words about ‘The roles* *of witches in society'*).

Integrative testing Test items which seek to draw on a wide range of language knowledge rather than those which test isolated parts of the language (= **discrete-point items**). In the past, all integrative items were **subjective** (e.g. writing a composition or speaking during an oral interview). More recently, **objective** techniques of integrative testing have been developed, notably **cloze** and **dictation**. Note: Some people prefer the term **global testing** to avoid confusion between the totally unrelated terms **integrative testing** and **integrative motivation** (= motivation to learn a foreign language because you identify or are sympathetic with the target language and its people and culture), and **integrated testing** and **integrative testing.**

Item An individual question in a test. Some people restrict the term to questions in a **battery test** of **discrete-point items**, but others also include items in **integrative tests.**

Modified cloze A form of **cloze testing** in which items are deleted at irregular intervals so as to test a particular aspect of language, e.g. articles, irregular past tenses, etc.

Multiple-choice A test item which presents the learner with a question or an incomplete statement (= **stem**) and several choices or options from which to choose the best answer (= the **key**), while rejecting the incorrect ones (= the **distracters**).

Multiple- A form of multiple choice in which learners are presented

matching with a number of questions or statements, and a list of possible answers. They then have to match the best answer(s) to each question or statement. Multiple-matching offers a number of advantages over traditional multiple-choice, especially where a question may have several answers. For example:

Norm-referenced testing The opposite of **criterion-referenced testing.** Tests which compare students with each other. Norm-referenced tests rank the students (i.e. put them in order from top to bottom or first to last). Results are then determined in various ways: a) a learner may be given a rank (e.g. 12th out of 43; b) a learner may be told that s/he is above or below ‘average’; c) a learner may be placed in a particular band (e.g. the top quarter, the 43rd percentile); d) above or below a ‘pass mark’, which is usually calculated in relation to the average or ‘mean’.

Objective tests Objective tests are those which can be scored on a right/wrong basis. The scoring is therefore said to be ‘mechanical’ and can, in most cases, be carried out by a computer if required. The answers should not be a matter of dispute as long as the item has been properly constructed. Objective tests have high **reliability**. Objective tests are contrasted with **subjective tests.**

Option One of the possible answers in a **multiple-choice item.**

Performance tests Tests which require learners to produce a sample of language, either in writing (e.g. a composition) or in speech (e.g. an interview). Such items are often designed to try to replicate language performance in the real world. Performance tests are usually marked subjectively, using **scales** with **descriptors.**

Placement tests Tests designed to arrange learners into groups or classes, usually by establishing their relative language levels, so that learners of roughly the same level can study together (= **streaming**). Placement tests usually have to be carried out in a hurry and so they tend to be **objective** and **'dirty'**. A **selection test** is a special kind of placement test in which lower-scoring learners are not accepted by the institution or organisation.

Portfolio At its simplest, a portfolio is simply a file in which to store evidence of a learner's language proficiency, including samples of his/her language output and test scores and certificates. More sophisticated portfolios, such as the European Language Portfolio produced by the Council of Europe include scales for self-assessment and scales which have been broken down into different series of learning objectives.

Practicality One of the criteria by which a test is assessed – see also **validity** and **reliability.** Practicality is the extent to which a test is quick, convenient and cheap to construct, administer and score. Tests with high practicality usually test large numbers of students at the same time, in a short time, with little equipment, are easy and economical to administer, and are quick and simple to score. They tend, therefore, to be **discrete-point** and **objective**, and to test receptive language skills rather than productive ones.

Predictive validity The extent to which a test predicts future language performance accurately. Predictive validity can be measured by correlating the results of two tests over a period of time: the first might be an **aptitude** or **placement** test administered at the start of a course; the second an **achievement** or **proficiency** test at the end of the course.

Proficiency A student’s present level of language ability, particularly his/her ability to use that language in real-world communication.

Proficiency test A forward-looking test which assesses whether a learner has the necessary language skills to undertake a task in the future, e.g. studying at an English-medium university or working in an English-medium environment. Proficiency tests therefore usually have an ESP flavour and test the language of the target situation. Note: The Cambridge Certificate of Proficiency in English (CPE) is not really a ‘proficiency’ test in this sense.

Profile A statement of a learner’s test results or language proficiency which shows not a single mark/grade but a series of assessments in various language areas, typically reading, writing, listening and speaking.

Progress tests Small-scale test which look back over recent language learning/teaching to assess how effective this has been. Progress tests may be used for continuous assessment, but their main use is to provide information so that the teacher can decide whether further teaching is needed in that area of language or whether the class can progress to the next area on the syllabus.

Rater Someone who scores, marks, grades or rates a test or part of a test. The term is used particularly with raters of writing and speaking tests, where a rigorous process of **rater training** may be required.

Reliability One of the criteria by which tests are assessed. Reliability refers to the consistency with which a test can be scored – consistency from person to person, time to time, place to place. We can distinguish **inter-rater reliability** (the closeness of the scores when a test is marked by two markers at much the same time) and **intra-rater reliability** or **mark-remark reliability** (the closeness of scores when a test is marked by the same person on two different occasions). **Objective** tests have better reliability than **subjective** ones.

Rubric The instructions for a test and each section/item in that test, including the time available. Rubrics may be given in English or the mother tongue and are frequently accompanied by an example to show how the item should be answered.

Scale A ‘ladder’ describing language performance, one which a number of ‘rungs’ or levels are identified, each with accompanying **descriptors** or **can-do statements.** Perhaps the best-known scales are those of the Council of Europe’s *Common Framework of Reference* (2001), which have six main levels. Scales have various uses, including assessment, self-assessment and setting of learning objectives.

Selection test A selection test is a special kind of **placement test** in which lower-scoring learners are not accepted by the institution or organisation.

Self-assessment Tests in which learners are asked to assess their own language level (self-**placement tests**) or language difficulties (**diagnostic tests**). Self-assessment usually involves some sort of questionnaire or asking learners to estimate their level on a **scale**.

Specifications A test’s specifications provide the official statement about what the test tests and how it tests it. The specifications are the blueprint to be followed by test and item writers, and they are also essential in the establishment of the test’s construct validity. (Alderson et al 1995: 9)

Stem See **multiple-choice.**

Streaming Dividing learners into groups, classes or ‘streams’, usually with each group at the same level. This is usually achieved through a **placement test.**

Subjective tests Subjective tests are those which require a teacher's/tester's judgement in their scoring. The scoring is, therefore, a matter of opinion and variable. Subjective tests normally have lower **reliability** than **objective** tests, but this has been improved recently through the use of **scales** and **descriptors**. All tests of productive skills (writing and speaking) are subjective. Contrasted with **objective tests.**

Summative assessment Summative tests are administered at the end of a stage of language learning and are designed to assess how much has been learnt or what level has been reached. They are often used as 'high-stakes' tests for purposes of certification or selection for further education or employment. **Achievement** and p**roficiency** tests are both summative. Contrasted with **formative tests.**

Technique A method of testing, e.g. multiple-choice, cloze, etc. Techniques can be broadly divided between **discrete-point/objective** and **integrative/ subjective.**

Top-down reading/ listening Top-down means using our prior knowledge and experiences; we know certain things about certain topics and situations and use that information to understand reading or listening texts.

Utility One of the criteria by which a test is assessed. Utility refers to the amount of information that a test supplies to the teacher/tester for the purpose of planning future learning/teaching. Raw scores provide little such information; tests which give **profiles** indicating areas of strengths and weaknesses in particular language areas have higher utility.

Validity One of the criteria by which tests are assessed. Broadly, validity means the extent to which a test actually assesses what it was intended to assess. Traditionally, there are five types of validity: **concurrent, construct, content, face** and **predictive** (all glossed separately in this glossary). Generally, different types of tests require different types of validity:

Washback effect One of the criteria by which tests are assessed. The washback effect (also called the **backwash effect**) is the effect that tests, especially **achievement** and **proficiency tests**, have on learning/teaching. Traditionally, tests such as multiple-choice have been thought to have a poor or negative washback effect, so that changing a high-stakes test such as the school-leaving exam into a communicative test may be the most effective way of reforming language teaching practices, so that the washback effect in such cases would be positive.

Washforward effect One of the criteria by which tests are assessed. The term was invented jokingly but it aptly captures a positive feature of many tests. It refers to the extent to which a test includes and tests language which is relevant to the post-learning target situation. **Proficiency tests**, therefore, should have a good washforward effect.