RESLA, 11 (1996), 17-49

TESTING ENGLISH AS A FOREIGN LANGUAGE: AN OVERVIEW AND SOME METHODOLOGICAL CONSIDERATIONS

ANTONIO BUENO GONZÁLEZ
Universidad de Jaén

ABSTRACT. *In this paper we present a general view of testing English as a foreign language, including key issues in the field, such as the relationship between teaching and testing, the various purposes testing aims at, together with the different approaches to testing in the history of TEFL. Special emphasis is given to communicative language testing and the requirements to make a test reliable, valid, discriminatory and useful. Different types of tests are discussed and exemplified and methodological hints are given in order to test the four skills and language aspects such as grammar, vocabulary and pronunciation. A proposal to make testing more functional closes the article.*

RESUMEN. *Presentamos en este trabajo una visión general de la evaluación de inglés como lengua extranjera, donde se incluyen temas claves como la relación entre la forma de enseñar y la forma de examinar, los distintos objetivos que la evaluación persigue, junto con los diferentes enfoques que el tema ha recibido en la Didáctica del inglés como lengua extranjera. Se presta una especial atención a las pruebas comunicativas y a los requisitos que hacen que una prueba sea fiable, válida, discriminatoria y útil. Se discuten y ejemplifican distintos tipos de pruebas y se ofrecen sugerencias metodológicas para evaluar las cuatro destrezas y aspectos lingüísticos tales como la gramática, el vocabulario o la pronunciación. Termina el artículo con una propuesta para hacer las pruebas más funcionales y comunicativas.*

1. INTRODUCTION

1.1. *Measuring language ability*

Testing is one of the most controversial areas related to any kind of teaching and at the same time something that is necessary as a sort of completion of the teaching and learning progress. In fact, if adequately focussed, it checks the effectiveness of the whole process. L. F. Bachman (1990: 55) insists on these aspects mentioning that the

information provided by testing is essential to effective formal education and that this feedback conveys appropriate changes in the program that improve learning and teaching. If we speak about language testing, and more specifically, foreign language testing, the first problem arises because language is both the instrument and the object of measurement (Bachman 1990: 2), that is to say, we use language to measure language ability.

We are speaking of testing as a kind of measurement but what should we measure? A first obvious answer is: language ability, or in a more concrete way, language performance. In this sense *"one of the most important and persistent problems in language testing is that of defining language ability in such a way that we can be sure that the test methods we use will elicit language test performance that is characteristic of language performance in non-test situations"* (Bachman 1990: 9). This refers to 'authenticity', 'measuring authentic (use of) language', a more than difficult issue in testing which will be considered in some detail when dealing with communicative testing. The term 'authentic' has been used by Spolsky (1985). Other terms which refer to a similar idea are 'pragmatic' (Oller 1979), 'functional' (Carroll 1980), 'communicative' (Morrow 1979), 'performance' (Jones 1985). To put it in a nutshell, we should test language as authentic as possible. In addition, language ability involves not only linguistic competence (the only concern of many traditional tests) but also communicative performance (increasingly included in many actual tests in order to make them resemble as closely as possible real life -the RL approach).

In the same vein, language makes reference both to language skills (listening, speaking, reading, writing) and linguistic components (grammar, vocabulary, phonetics). There should be a combination of both if we really want to test language as a whole. Many of the seminal books about testing present tasks for the individual skills and for the linguistic components (Heaton 1975, 1988[2]; Madsen 1983; Hughes 1989).

So we can attempt a composite definition of what we should measure when we test our students: the ability to use the different linguistic components both in a receptive and productive way in oral and written media in order to achieve a purpose as communicative and authentic as possible. An additional requirement seems to be lacking in the previous statement: the need for language to be contextualized. The more contextualized the different items of a test are the more authentic the test is. Furthermore, there is a golden rule as to what to test: *'Test what you teach'*. This leads us into the next section.

1.2. *Testing and teaching*

> *Both testing and teaching are so closely interrelated that it is virtually impossible to work in either field without being constantly concerned with the other.*
>
> (Heaton 1988[2]: 5)

We have already pointed out this fortunate relationship and it is from this perspective that we concentrate our attention from now onwards on classroom tests. We will make this issue more concrete when speaking about the purposes of testing as regards the classroom. Suffice it to say that apart from being used formally for assessment, a test can be used *"as essentially a constructive and practical teaching strategy giving learners useful opportunities for the discussion of language choices"* (Hubbard *et al.* 1983: 256). By 'backwash' or 'feedback' we understand the effect of testing on teaching, which can be either harmful (according to Heaton 1988[2]: 5, in the past even good tests of grammar, translation or language manipulation had a negative effect on teaching and he proposes communicative tests which should generally result in improved learning habits) or beneficial. In the second case, they have positive effects both for students and teachers. Madsen (1983: 4-5) suggests at least three ways in which well-made tests can help students: creating positive attitudes towards the class, helping them master the language and giving a sense of accomplishment.

In a similar way, tests help the teacher know if his/her teaching was effective and can provide insights into ways to improve the testing process. From this perspective, we can consider the relationship between teaching and testing as one of partnership: *"we cannot expect testing only to follow teaching. What we should demand of it, however, is that it should be supportive of good teaching and, where necessary, exert a corrective influence on bad teaching"* (Hughes 1989: 2). *"When there is a serious discrepancy between the teaching and the means of evaluating the teaching, then some-thing appears to be amiss"* (Alderson 1981a: 6).

There is another important conclusion which derives from here: teacher-made tests can be superior in certain respects to their professional counterparts. More often than not, teachers believe that writing tests requires a sort of expertise they lack. As we said before, we are here concerned with classroom tests; by following the aforementioned golden rule (*'Test what you teach'*) a classroom test is given validity. And it is obvious enough that it is the teacher who best knows the teaching profession: *"a test is seen as a natural extension of classroom work, providing teacher and student with useful information that can serve each as a basis for improvement ... it follows that the person best prepared to set the test is the teacher"* (Harrison 1983: 1). In this sense, some of the recommendations we will make will help teachers to write better tests themselves and they can put pressure on others to improve their tests. In a way, this parallels what is happening with ELT research in general: more importance is increasingly being given to classroom-centred research and the researcher who is an 'outsider' and loses contact with the classroom is highly questionable.

In this respect, classroom tests differ from external examinations in that the latter are generally concerned with evaluation for the purpose of selection while the purpose of the former is to enable teachers to increase their own effectiveness by making adjustments in their teaching to enable certain groups of students or individuals in the class to benefit more (cf. Heaton 1988[2]: 6). So, a well-constructed test is one which covers an adequate and representative section of the intended areas and skills and

reflects the actual teaching being followed. This test is useful to locate precise areas of difficulty, to evaluate the effectiveness of the syllabus, methods and materials and can even be a source of motivation for students.

Last but not least, testing should always be followed by remedial teaching, a very clear example of the fruitful relationship of both: *"Learners will gain more from feedback of a more personal nature which gives credit for what they have got right, as well as help for what they have got wrong"* (Williams 1985: 143). Responding to students is, then, essential.

1.3. *Purposes of testing*

When speaking about the relationship between testing and teaching we mentioned some of the aims testing fulfils. Hubbard *et al.* (1983: 255) propose up to ten in the following chart, which we use for discussion with our students in class:

Identification of problem areas for remedial attention.
Giving each student a course grade.
Assessment of your own effectiveness as a teacher.
Checking on general progress and obtaining feedback.
Course or syllabus evaluation.
Preparation for public examinations.
Institutional requirement for student promotion.
Measuring what a student knows.
Identification of levels for later group-work.
Reinforcement of learning and student motivation.

In fact, except for 'Preparation for public examination' which, although occasionally, is not always the case of classroom testing, we can say that all the other purposes show how useful testing can be for teaching and in this way it supports what has been previously said. Globally, it can be truly stated that in many ways, testing helps to know about the effectiveness of the teaching and learning process.

Next, we can distinguish different tests according to different purposes. We will deal with them in detail later. In this sense, in the case of placement tests we want to measure ability before the start of a programme. Aptitude tests aim at providing information on the likelihood of success or failure. If we want to predict language behaviour in a real-life situation or when an individual is capable of doing now we use a proficiency test. The purpose of achievement tests is to establish what a learner has learnt; it can be said that they have a past orientation. In order to discover areas and causes of failure (future, predictive orientation) we use diagnostic tests.

1.4. *Testing* vs *evaluation*

We are primarily concerned here with testing as different from evaluation, which is a wider concept. Testing is only one component of the evaluation process. For an

ample discussion of evaluation we recommend the reading of the book (1992) by P. Rea-Dickins and K. Germaine *Evaluation* (Oxford University Press).

By testing we understand any formal or informal task set at a given moment for one or several purposes. It may be more structural or more communicative, longer or shorter, but always given as a precise means to provide assessment. We consciously neutralize the terms 'examination' and 'test' because, in fact, the difference is not very clear in the specialized literature and there seems to be no consensus on what the distinction is. Pilliner (1968: 21-22) tries to establish the difference in terms of time (examinations take longer), hierarchy (examinations seem to be more important and set at a more advanced level) and assessment (examinations favour subjective scoring while tests are more objective).

Evaluation comprises not only tests but also the continuous assessment which is made daily by the teacher, together with the analysis of the personal elements (students and teacher), and the material ones (school, classroom, syllabus, textbooks, readers and other materials, teaching aids, etc.). It also includes the methodology used and self-evaluation. In sum, it provides information about the whole process of teaching and learning by studying the different aspects it involves.

1.5. *Some other key issues*

Testing is often the basis for taking important decisions. These decisions affect people (students in the first place and also parents and other people who are responsible for education). So, there should be a high component of fairness and to a certain extent some degree of tactfulness in testing. The information upon which we base decisions should be as reliable and valid as possible. This is the reason why we will mention reliability and validity as essential requirements of a good test.

We can help testees by identifying clearly the area or problem we want to test. Unfortunately, this is not always easy. It is very clear in discrete-point tests, there is considerable variation in 'editing tasks' (underlining errors to be corrected or no underlining) and there is no identification at all in some 'integrative' tests (oral interview, composition or dictation): "*Input in which the information is either highly compact or highly diffuse will be relatively difficult to process*" (Bachman 1990: 135).

It is also the duty of the tester to reduce as much as possible what is known as 'test anxiety'. A sort of 'humanization' of the testing environment has been suggested. There is no doubt that factors such as the familiarity of the place and personnel administering the test and even the personal qualities of the tester(s) affect performance con-siderably and can influence the reliability and validity of a test. In addition, we should not use tests as the only basis to evaluate our students. Some of the issues mentioned here will help to overcome the apparently eternal discontent about language tests.

1.6. *Extended definition of a test*

Brown (1987: 219) gives the following definition of a test in plain, ordinary words: *"a method of measuring a person's ability or knowledge in a given area"*. Let us expand on it. It is a method, which means that it belongs to the conventional aspects of teaching, not to the essence, and, consequently, it may be different from teacher to teacher, from school to school, from period to period. It reflects an underlying set of theories or *approach* and, in that sense, it emphasizes one aspect or another. When we speak of measuring we mean that some objective quantification needed, apart from obvious subjective valuation. We deal with people, which implies that the test should be as impartial and human as possible and justifiable on all grounds. We measure ability, that is to say, the practical mastering of the language, its functional and communicative nature, together with knowledge for which a certain theoretical background is needed as well as some formal checking of the grammatical skeleton and some memorization. Finally, our area refers to English as a foreign language, bearing in mind that this is different from English as a second language.

2. APPROACHES TO TESTING

From Spolsky (1975) it has been customary to distinguish three chronological periods in the history of testing: pre-scientific, psychometric-structuralist and psycholinguistic-sociolinguistic. K. Morrow (1979) calls them metaphorically *'The Garden of Eden'*, *'The Vale of Tears'* and *'The Promised Land'*. A fourth one seems to have been added: the so-called communicative approach to language teaching and its counterpart in language testing, assessing communicative competence. As happens in ELT history, the different approaches can be broadly traced to a certain period but the chronological limit is not clear-cut. On some occasions there seem to be features belonging to more than one approach, and even an eclectic approach has often been suggested. Let us deal with each one in turn.

2.1. *The pre-scientific period*

Chronologically, it is prior to the early 1950s, when there was virtually no language testing research. Teachers constructed their own tests, basically following the general principles of humanities and social sciences. The exercises included in these tests were grammar-translation or reading-oriented, such as translation, essay-writing, testing knowledge of grammar -often with incomplete sentences to be completed. This is the reason why Heaton (1988[2]) calls it 'the essay-translation approach'. It has also been termed 'traditional' and had a highly subjective character; no attention was paid to reliability, objectivity or statistics.

On the credit side, we have to recognize the acknowledgement of personal responsibility on the part of the teacher. On the debt side, this testing method is within the traditional way of thinking and it has been criticised as elitist and authoritarian. It is rooted on the techniques used to test classical languages.

Madsen (1983: 5 ff) calls this first period 'intuitive' because of its subjective character and its dependence on the personal impressions of teachers. In fact, one of the main problems of this approach, apart from those mentioned, is the one derived from subjective marking.

All in all, some of the exercises are still widely used today and prove to be successful with certain modifications, such as the concern for language to be authentic and contextualized, the emphasis on the communicative and interactional aspects or the use of marking bands for compositions or oral interviews, along with a more humanized attitude on the part of the tester.

2.2. *The psychometric-structuralist period*

The excess of subjectivity in the previous period led to the need of more objective means to measure language ability. Historically, this approach can be dated in the 50s and 60s, coinciding with the structuralist views of language (Fries), the contrastive analysis hypothesis and behaviourism (Lado). Testing is now focussed on specific language elements (discrete points, each item tests an element), especially centred on the contrasts between the mother tongue and the target language.

This period is also called 'modern' and 'scientific' because with the help of measurement experts and statistical procedures it is demonstrated that testing can be objective, precise, reliable and, in sum, scientific. Testing specialists with linguistic training entered the scene. It is the time of multiple-choice questions centered on structures or vocabulary items, together with tests devised to measure performance or recognition of separate sounds. Tests which refer only to one skill or one linguistic component are frequent, following a simplistic, static and analytical conception (the testing competence), completely different to that of Chomsky (dynamic, creative, synthetic, the testing of performance). Robert Lado stressed two points that have become very important: tests should test language *usage* and not knowledge about language; the structures to be tested should be valid structures in colloquial language *use*. With the passing of time there have been some adaptations of this approach, moving from smaller to larger units and asking for responses not only linguistically correct but also situationally appropriate.

The main criticism comes from the fact that it is based on an atomistic view of language (isolated segments) and on the idea that knowledge of the elements of a language is equivalent to knowledge of the language. But as Morrow (1977) points out, we cannot forget that synthesis is necessary. By considering answers as either right or wrong, the concept of 'transitional competence' (Chomsky) -not only right or wrong but intermediate stages- or 'interlanguage' is lost and responses are assessed

only quantitatively, not qualitatively. Lado's objectivity accounts for reliability but Morrow (1977) questions this equation since these tests are objective only in terms of actual assessment or scoring but in terms of the construction of the test itself, subjective factors play a considerable role.

In order to be fair, we cannot forget how much easily quantifiable data (such as those obtained with these tests) help the teacher at the scoring stage. In the same way we have to admit with Morrow (1977) that though not sharing many of Lado's theories, test writers have accepted some of his influential ideas, such as the importance of reliability and validity and the advantages of the directly quantifiable modes of assessment. Anyway, objectivity results in higher reliability but what about validity? Do these tests really measure what you want to measure?. Most multiple-choice tests measure the recognition level but not that of production. Madsen (1983: 6) mentions aptitude tests, designed to predict success in learning a second language, as one of the interesting by-products of this approach.

2.3. *The psycholinguistic-sociolinguistic period*

On the one hand, it is a reaction to the previous period and on the other, a prelude of the communicative era. The growing dissatisfaction with structuralism and behaviourism led test writers and teachers to consider the need to test the whole of the communicative event. This has to do with Oller's 'communicative performance', measuring the total communicative effect of an utterance. In this sense, integrative tests, such as cloze, dictation, composition, oral interview and translation, should be used.

We are dealing with the concept of 'unitary competence' as overall language proficiency, based on an underlying linguistic competence and related to what Oller (1978, 1979) calls 'pragmatic expectancy grammar':

> *The term expectancy grammar calls attention to the peculiarly sequential organization of language in actual use. ... The term pragmatic expectancy grammar further calls attention to the fact that the sequences of classes of elements, and hierarchies of them which constitute a language are available to the language user in real-life situations because they are somehow indexed with reference to their appropriateness to extralinguistic contexts.*
>
> (Oller 1979: 24)

The best exams, then, are those which combine various skills as we do when exchanging ideas orally or in writing. This is what is understood by pragmatic tests, that is to say, a test which requires from the student the use of more than one skill and one or more linguistic components. Discrete-item tests are not pragmatic. Language is a whole, with the purpose of communication. Hence, full attention to meaning is another feature of pragmatic tests. As we said at the beginning of this section, the communicative period is just round the corner.

But before dealing with communicative testing we have to mention that on many occasions tests combine discrete-item exercises (typical of the psychometric-structuralist period) for diagnosis purposes and in order to test structural aspects and linguistic competence in general and global exercises (peculiar to the psycholinguistic-sociolinguistic one) to test general knowledge, emphasizing communicative use and linguistic performance. In fact, many external examinations (TOEFL, Cambridge Examinations) include several papers or sections as an example of this sort of eclectic synthesis (cf. Martínez Haro 1984: 79 ff): Listening Comprehension, Multiple Choice, Vocabulary, Grammar or Use of English, Reading Comprehension, Composition.

Needless to say many classroom tests take this form (with different parts) in a more or less balanced way. If the language is contextualized, the topic relevant and interesting and the expected answers require students to use authentic or semi-authentic language in a to-a-certain-extent communicative situation, we will have an ideal picture of a good test. What we are doing is just combining the three approaches.

2.4. *The communicative period*

2.4.1. *The controversy*

Communicative testing is something controversial because, although it seems to characterize current trends, in accordance with the correspondent theories of language teaching that emphasize communication above all, at the same time it gives the impression of being something unattainable or attainable only up to a certain extent. As Alderson (1981b: 48) has clearly said, the settting of assessment disauthenticates most language tests. Opinions range from total commitment to it to the conviction that communicative testing is just impossible. A clear example to show the opposite feelings is Section 1 ('Communicative Language Testing') of *ELT Document No. 111, Issues in Language Testing*, where the seminal paper by K. Morrow (1979) 'Communicative language testing: revolution or evolution?' is discussed and reactions to it expressed by C. J. Weir, A. Moller and J. C. Alderson. Although a little old now, these articles are worth reading in order to understand the controversy that more than a decade later still exists.

To begin with, let us define what is understood by communicative testing and we will spend some time later discussing the problem of authenticity.

2.4.2. *What is communicative language testing?*

A. Moller (1981: 39) provides the following definition:

> *An assessment of the ability to use one or more of the phonological, syntactic and semantic systems of the language 1) so as to communicate ideas and information to another speaker/reader in such a way that the intended meaning of the message communicated is received and understood and 2) so as to receive and understand the meaning of a message communicated by another speaker/writer that the speaker/writer intended to convey.*

The difference with discrete and integrative tests is that in them the candidate is an 'outsider', while in communicative performance tests the candidate is an 'insider'. Another difference is that in the former we test 'usage' but in the latter it is 'use' which is tested. Of the several characteristics Morrow (1977) gives to identify a situation as communicative, the following can be listed for a testing activity to be communicative: search for information, creativeness, a purpose and authenticity; that is to say, students communicate something in the test. In addition, tests should be criterion-referenced (please see 4.2.2.) and assessment should be based on quality and not quantity. Bachman (1990: 107) defines communicative language ability as both knowledge of language and the capacity for implementing that knowledge in communicative language use. Brown (1987: 230 ff) mentions some primary criteria for the construction of communicative tests: concentration on content, providing something motivating, interesting and substantive and at the same time integrated and interactive, and grading the difficulty of the items (from easier to more difficult). Bestard Monroig and Pérez Martín (1992: 201 ff) emphasize the importance of providing students with a physical context (the house, the bus...), a clear communicative activity and the sociocultural context, insisting on the relationship between the participants. As regards the difficulty of offering a completely real context (only possible in the foreign country) they suggest the use of an imaginary context in the classroom by means of drama, simulation, problem-solving activities and role-play and they insist on the need for a global, qualitative assessment.

2.4.3. *The problem of authenticity*

All the authors coincide in saying that language tests are by definition inauthentic: *"Does not the very fact that the setting is one of assessment disauthenticate most language tests? Are there not some language tasks which are authentic in a language test, which would be inauthentic outside that domain?"* (Alderson 1981b: 48). This is what Davies (1978) calls 'the quimera of authenticity' because *"the conditions for actual real-life communication are not replicable in an artificial and idealised test situation"* (Weir 1981: 29). In addition, the more authentic the language task we test, the more difficult it is to measure reliably. We can say that the development of the communicative theory in language teaching and language materials, such as text-books, seems to have no parallel in communicative testing. Alderson (1981b: 54) even speaks of failure:

> *Testing is the testing ground for any approach to teaching. If we cannot get the tests our theories seem to require, then we have probably not got our theories right (unless, of course, the theory implies the impossiblity of testing). Why has there apparently been such a failure to develop tests consistent with theories of communicative language use?*

We think we should not probably go so far as to consider communicative tests as either inexistent or impossible. First of all, as Moller (1981: 83) recognises, some of the most traditional forms of language testing, the viva and the dissertation or essay, are both forms of communication. Secondly, if we agree that communicative tests should be an assessment of what a candidate can actually do with the language, performance-based tests containing the characteristics we mentioned for a communicative situation (Morrow 1977) will be communicative. Thirdly, we agree with Bachman (1990: 315) that authenticity should not be strictly identified with 'natural situations' and the RL approach but that test language is different from real-life language and, in this sense, language tests have an authenticity of their own:

> *I find the authenticity argument somewhat sterile since it seems to assume that the domains of language teaching and language testing do not have their own set of specifications for authentic language use which are distinct from the specifications of other domains. Thus 'What is this? -It's a pencil' is authentic language teaching language, and so on. If one does not accept this, then authentic tasks are in principle impossible in a language testing situation and communicative language testing is in principle impossible.*
>
> (Alderson 1981b: 48)

Authenticity has to do with interaction and negotiation of meaning and, in this sense, it is very similar to Oller's description of a 'pragmatic test':

> *... any procedure or task that causes the learner to process sequences of elements in a language that conform to the normal contextual constraints of that language, and which requires the learner to relate sequences of linguistic elements via pragmatic mapping to extralinguistic context.*
>
> (Oller 1979: 38)

In sum, tests that make students relate form and meaning in a relevant context and that contain meaningful and interesting tasks similar to those in real life can be considered communicative, although not completely authentic or real. As examples we can mention split dialogues, problem-solving activities or those tasks in which students have to choose from a series of communicative choices, according to the appropriate register.

3. REQUIREMENTS OF A GOOD TEST

Above all, a test must be reliable and valid. In addition, it should be feasible, that is to say, it should have practicality. It should discriminate and be useful for students and teachers ('feedback'). Let us consider each requirement in detail.

3.1. *Reliability*

A test is reliable if it is consistent and dependable, in other words, we can rely on the information it provides. This reliability refers to the test in itself and also to the external factors implied. In this sense, we should try and make surrounding conditions and personal elements as reliable as possible, considering aspects such as the place of the examination (conditions of noise, heat, etc.), the person giving the test, the instructions (rubrics) on the examination paper and the amount of time allowed. Apart from that, a test should be reliable in the way of correcting and marking ('scorer reliability'). It is highly recommendable to reach agreement among several scorers. Likewise, scoring directions should be clear and specific.

Among other factors that account for reliability we have to mention objectivity and length. The longer the test is the more reliable it becomes. Reliability is a pre-requisite for validity. For a test to be valid it must be reliable. It is not the case the other way round: a test can be reliable (you can depend on it) but not valid (it does not measure what it is intended to measure). For example, a multiple-choice test can be reliable but not valid to test oral or written production.

Some formulae to reveal internal consistency have been proposed but for the layman suffice it to say that *"a test is unreliable if it provides very different results when administered to two different groups of equal ability"* (Leeman 1981: 119). Once we discover that a test is unreliable we should try and look for its weaknesses or external conditions which are responsible for it. Among other factors that can affect performance, and consequently the reliability of a test, we can mention the following: the relative importance of the parts of a test, the testing environment, the test rubric, the kind of input provided, the format (cf. Bachman 1990: 119-122), the number of people taking the test and the amount of language given. Tension or test anxiety is one of the main sources of unreliability and may be caused by the inappropriateness of the previous factors. Heaton (1988[2]: 167-170) emphasizes the importance of clear instructions and other practicalities to ensure reliability, mentioning concrete, apparently insignificant, details of administration, such as the fact that confusion may result in a multiple-choice exercise if the items are numbered vertically on the question papers and horizontal numbering is adopted for the corresponding answer sheet.

Hughes (1989: 36-42) provides us with the means to make tests more reliable in what may be a very good guide for teachers and other test writers: taking enough samples of behaviour, not allowing candidates too much freedom, writing unambiguous items, providing clear and explicit instructions, ensuring that tests are well laid out and perfectly legible and that candidates are familiar with format and testing techniques. In relation to scoring reliability, he suggests the following: using items that permit scoring as objective as possible, providing a detailed scoring key (with agreement if there is more than one scorer on the key and the acceptable responses) and employing -if possible- multiple, independent scoring.

3.2. *Validity*

A test is valid if it actually measures what it is intended to measure. On a classroom footing, it must test what you have taught your students, what they have studied, and in this way, the exercises should be of the same type and approximately the same level as the practical activities in class. Another idea follows from this statement: together with the results of the test, everything done in class counts. With no doubt, this is the most important requirement of a test. As we said before, a test, in spite of being reliable, can have no validity. In fact, many authors speak of the reliability-validity tension (Davies 1978). A test can be perfectly reliable but students may produce no language at all. If our objective is production, this test is not valid. We can distinguish several types of validity.

3.2.1. *Content validity*

Some authors (Bell 1981) refer to it when the tasks of the test reflect truly the skills in real life. In a more concrete way, we can consider that a test has content validity if it measures the contents of a teaching programme (especially with achievement tests) or the specifications of any external examination. In other words, content validity refers to the fact that the selection of tasks one observes in a test-taking situation is representative of the larger set (universe) of tasks of which the test is assumed to be a sample (Bachman and Palmer 1981: 136). From the previous definitions it follows that the content of a test, in order to be valid, should constitute a representative (relevant/ specific) sample of the language skills, structures, etc., with which it is meant to be concerned. This means that major areas and areas which have received special teaching emphasis should be present. Many tests have no content validity because their content is determined by what is *easy* to test rather than what is *important* to test (cf. Hughes 1989: 22-23).

3.2.2. *Construct validity*

A test is said to have construct validity if it accurately reflects the construct or theory underlying it, that is to say, if it is able to satisfy some previously stated theory against which we validate it. For example, in a test of reading comprehension, if our underlying conviction is that vocabulary is more relevant than syntax, the scoring procedure should reflect this theoretical assumption and vocabulary and syntax should be weighed accordingly. So, the concept of construct validity assumes the existence of certain learning theories or constructs underlying the acquisition of abilities and skills. Taking again the classroom as a basis, the test should be consistent with the approach used during the course. If the teaching has been clearly structural, with a lot of emphasis on grammar and translation, we would not expect a communicative test to be used.

3.2.3. *Criterion-related validity*

The validity of a test can be validated by comparing its results to another assessment (criterion). If the test and the criterion are administered at about the same

time, we speak of 'concurrent validity' (for example, if a group of students is given the test and is immediately rated by an experienced teacher or immediately given a longer test). If it concerns the degree to which a test can predict the candidate's future performance, it is called 'predictive validity' (for instance, with placement tests: once courses are under way, we can check validation by establishing the proportion of students who were thought to be misplaced -cf. Hughes 1989: 23-25). This type of validity has also been called 'empirical', 'pragmatic' or 'statistical' (Bell 1981: 198).

3.2.4. *Face validity*

Simple though it may appear, it is of paramount importance, especially for the testees: it refers to the layman's impression of what a test measures, that is to say, if the test is accepted as appearing to be appropriate by those who administer it and those who take it, the extent to which a test looks like it measures what it is supposed to; in sum, the appearance of validity.

3.3. *Practicality*

It concerns the useability of the test. Is it feasible? Can it be used with my students?. According to Bell (1981: 200) two parameters appear to be involved: economy (in terms of money and time) and ease. Likewise, we should refer back to the suggestions we made in order to make tests more reliable as tips to make them more practical too. Perhaps the article by L. Dangerfield (1985a) "Writing achievement tests: practical tips", although confined to one particular type of test, will be a good complement to what was said. It involves questions about time, coverage, format, difficulty, rubrics and marks.

3.4. *Discrimination*

It is considered here in a positive sense and, in some way, it constitutes a feature of validity: any test should offer a range of results (except perhaps for some achievement tests where all candidates are expected to score high). As we will explain later, there are some means to measure discrimination.

3.5. *Usefulness*

A test must have instructional value, it must be useful, first of all for students and teachers and also for institutions and examining bodies. Everything said when speaking about teaching and testing is applicable here.

Putting all the requirements together, we would get the picture of the ideal test (Bell 1981: 200):

> *The ideal test would be one which was reliable in that it provided dependable measurements, was valid in that it only measured what it was supposed to*

> *measure, supported what we already believed about the nature of language and of learning and agreed with trustworthy outside criteria but also looked as though it did all these things. In addition, it would be cheap and easy to use.*

His final comment (*"Such a test is, of course, at the end of the rainbow and we are still looking"*) is probably more than relevant nowadays.

4. KINDS OF TEST

There are different kinds of test according to different criteria. We are going to concentrate on five: purpose, frame of reference, scoring procedure, content and specific testing method or format. Throughout the discussion of this classification we will pay some detailed attention to standard tests, such as multiple-choice, true/ false, oral interview, composition, cloze and dictation.

4.1. *Purpose*

According to the different purposes, we can distinguish proficiency, diagnostic, achievement and aptitude tests.

4.1.1. *Proficiency*

They are not limited to any course or curriculum and they measure global competence in a language. Some of them are external examinations which contain several papers and are machine scorable (TOEFL, Cambridge examinations) and sometimes they add free writing or speaking with the subsequent problem of practicality of scoring. The aim is, then, to assess the student's ability to apply in actual situations what he has learnt (Harrison 1983: 7, *"having learnt this much, what can the student do with it?"*). Consequently, we can say that they have future orientation.

4.1.2. *Diagnostic*

They are also called 'formative' or 'progress' tests and they are used to diagnose a particular aspect of a particular language or to check on students' progress in learning particular elements of the course. Many end-of-unit/ lesson tests are diagnostic. Sometimes, they serve to detect difficulty on some areas and, in this sense, they are a useful help for teaching. They generally refer to short-term objectives.

Some authors (Hughes, for instance) include 'progress' tests under the achievement heading, while others (Harrison) reserve this term for longer periods of learning, that is to say, only for final examinations.

Placement tests provide a special kind of diagnostic tests. They contain a sampling of materials to be covered in one specific curriculum and serve to check where students should be placed. This is why Hughes (1989: 14) calls them *'tailor-made rather than bought off the peg'*.

4.1.3. *Achievement*

The clearest difference with proficiency tests is that in this case they are related directly to classroom lessons, units or a total curriculum. So, they are limited to particular material covered in a curriculum within a particular time frame. As we have already seen, if the period is short they are often called 'progress' tests and included by some authors under 'diagnostic'. Our personal opinion is that any test based on material taught in a given period of time measures what students have achieved in that period, no matter how long it is. In this sense, both final and 'progress' tests should be termed 'achievement' (cf. Hughes 1989: 13, they serve *'to keep students on their toes'*).

4.1.4. *Aptitude*

They are prior to any exposure to the target language and they are intended to show if a person will be successful or not in learning a foreign language, for instance. As Brown (1987) points out, they are rarely used today because they merely predict the general scholarly success of a student without saying anything about the strategies he/ she may use. Apart from that, there are even serious ethical objections because they bias both student and teacher. Some of the best known are *The Modern Language Aptitude Test* and *The Pimsleur Language Aptitude Battery*.

4.2. *Frame of reference*

4.2.1. *Norm-referenced*

One candidate's performance is related to that of other candidates. The score places him/ her in a particular position within the group. Whenever a mark is given, we are dealing with norm-referenced tests. Although based on what the candidate has done, we are not told directly what he/ she is capable of doing in the language.

4.2.2. *Criterion-referenced*

We learn something about what the individual can actually do in the language. Each candidate's performance is described by means of comments which show what he can and cannot do in relation to the purpose of the test. There is no explicit comparison with other candidates. These tests create a more positive attitude and reduce the negative effect of competition. In fact, if we want to make our tests more communicative, this frame of reference is recommended.

4.3. *Scoring procedure*

In general, we can say that tests that refer to discrete items favour objective scoring while global or integrative tests favour subjective scoring. Likewise, certain skills and areas of language may be tested more effectively by one method than by another: objective tests are very useful to test at the recognition level, while subjective tests are better for the production level. Objective tests are easy to mark (many of them are machine scorable) and difficult to write (multiple-choice, for instance). With subjective tests it is just the other way round: setting a composition or an oral interview does not take too long but scoring may be rather complicated and general criteria have to be established. According to Heaton (1988[2]: 27) a good classroom test will usually contain both subjective and objective test items.

4.3.1. *Objective*

Although we speak of objective tests, we borrow Heaton's words (1988[2]) to say that it is only the scoring that can be described as objective, because there is a certain subjectivity in all tests (especially when writing them). We must admit that objective tests can never test the ability to communicate in the foreign language, they cannot evaluate performance. Let us discuss in some detail two of the standard formats for objective scoring: multiple-choice and true/ false items.

4.3.1.1. *Multiple-Choice*

Students have to answer by choosing one of a number of alternatives. The format is familiar to teachers: the basic frame is called *stem*, the alternative possibilities are the *options*, of which the correct one is the *key* and the others are termed *distractors*. To facilitate marking a separate answer sheet is usually given. We must be careful because if the options on the test paper are arranged vertically and those on the answer sheet horizontally (as is usually the case) this can cause confusion.

The main criticism about multiple-choice is that this format does not lend itself to the testing of language as communication but we must admit that it is very useful in order to recognise discriminations and subtle differences in grammar, vocabulary and phonetics. At the same time, many areas of difficulty can be identified. It can also be used for reading/ listening comprehension, appropriateness, error recognition and punctuation. The principal difficulty lies in writing plausible distractors. They have to fulfill some conditions (Heaton 1988[2]: 28 ff provides some of the general principles multiple-choice items have to fulfil) and on some occasions it is just the case that you cannot find three good distractors. Many authors coincide in suggesting errors your students make as good sources to write distractors (give them open-ended sentences to complete and they will give you the distractors). Along with this, Heaton (1988[2]) suggests that plausible distractors can also be based on the teacher's experience and contrastive analysis items.

Multiple-choice items have also been criticised on the grounds that they encourage guessing, although we have to admit that very rarely do students guess with no reason at all. One suggestion to discourage guessing is for every wrong answer to make the student lose one mark. To offer five options (customary 20 years ago) instead of four (now) complicates the writing of the test and it does not make any real difference.

A third criticism is that because of its format (three wrong options and only one right) many traditional multiple choice tests expose students to many unlikely errors, that is to say, *"a situation where far more language is wrong than is right"* (Hubbard *et al.* 1983: 261). An alternative has been proposed: three keys and one distractor. Students have to identify the distractor and eliminate it. Thus, the test becomes a 'correction' test.

4.3.1.2. *True/ False*

This type of test is used both for reading and listening comprehension and also for the understanding of specific elements of the language. In fact, it is a multiple-choice test with only two options. This means that there is a fifty per cent of probability of getting the right answer and, obviously, this encourages guessing. Three solutions have been proposed (cf. Dangerfield 1985b: 157-159): by substracting one mark for each wrong answer (the problem is that it discourages guesses based on partial understanding); by requiring the students to write corrections to the false statements or justification for the answer (but this adds the element of subjective scoring, which makes the test somewhat impure); by adding a third element 'Don't know' or 'Not stated in the test'.

Something to be borne in mind is that negative true/ false statements should be avoided because they induce confusion, for instance, the following statement is true, but it may be the case that some students choose 'False' because of the wrong date: *Christopher Columbus did not discover America in 1493.*

4.3.2. *Subjective*

In spite of the scoring difficulty, they are very useful because they test production, both oral and written. In addition, they are very easy to set. We will briefly refer to two traditional tests, oral interview and composition, which are widely used today, once illocutionary and contextual facets have been added; in sum, they have been given a communicative focus. We will also mention two methods to score them, holistic and analytic.

4.3.2.1. *Oral interview*

The main difference between the traditional oral interview and the most recent one lies in the fact that the former generally consisted in a series of questions and answers, while the latter implies the constant interaction of the interviewer and the

student. The result is that the interviewee gets the impression that he/ she is talking to someone. Evidently, a sincere, open, supportive manner on the part of the interviewer helps a lot.

Madsen (1983: 162 ff) proposes the combination of several elements to make the interview interactive, interesting for the student and by using various elicitation techniques. In this sense, he recommends a guided oral interview, preparing some cues beforehand, which will be adapted according to the course the conversation follows. Martínez Haro (1984: 72) suggests three previous steps the teacher has to follow before the interview: determining those aspects of oral production that are to be tested, preparing elicitation questions for those points and preparing a scoring sheet with the previous elements and the marking system.

Questions can be made personal if the interviewer knows something about the students. A range of yes/ no, wh- and either/ or questions should be used, together with statements. It is important to include some questions or statements that require some kind of correction or modification to make students talk. In the same vein, there should be some questions requiring clarification (the teacher makes them ambiguous and not very clear on purpose; it also gives the student the opportunity to make questions, not only answering them).

With regard to the difficulty, easy questions should go at the very beginning and at the very end, and after a rather challenging item or two, one or two easier questions should be inserted. On a methodological footing, students should be given opportunities to talk and for that the teacher must be flexible and ready to change the topic if the student seems to be at a loss. In the same way, an interview should not be stopped when the student keeps silent because he/ she has nothing to say about one particular item. The idea is that he/ she should leave the room with the satisfaction of having said something about the last two or three questions. This is also the reason for these to be easier. Between five and ten minutes per student seems to be recommendable.

The solution of interviewing several students at the same time has the advantage of allowing more interaction and the small group can be engaged in real conversation, but it has many drawbacks: the teacher may not get clear assessment criteria for individual students and one student may monopolize the conversation and harm the others.

4.3.2.2. *Composition*

In the traditional sense it has been one of the easiest tasks to set, both as classroom homework and as a testing exercise. It just consisted in giving the students 'poor titles' which gave them no guidance as to what was expected of them. Some examples of these titles are 'A horrible evening', 'My best friend', 'A good book I have recently read'. In order to make the testing of writing more communicative some elements should be included in the rubrics: meaningful contexts and situations (plausible situations in which students may find themselves); a reason to write (a clearly defined

problem which motivates them); a reader or readers, apart from the teacher, otherwise, the sense of communication is lost (it is advisable to provide as many details as possible about this real or imaginary reader).

The contrast between one and the other type of composition is clearly shown by Heaton (1988²: 137-138) with these two examples and it is to the second type that we turn our efforts nowadays.

> *(a) Write a letter, telling a friend about any interesting school excursion on which you have been.*
>
> *(b) You have just been on a school excursion to a nearby seaside town. However, you were not taken to the beach and you had no free time at all to wander round the town. You are very keen on swimming and you also enjoy going to the cinema. Your teacher often tells you that you should study more and not waste your time. On the excursion you visited the law courts, an art gallery and a big museum. It was all very boring apart from one room in the museum containing old-fashioned armour and scenes of battles. You found this room far more interesting than you thought it would be but you didn't talk to your friends or teacher about it. In fact, you were so interested in it that you left a small camera there. Your teacher told you off because you have a reputation for forgetting things. Only your cousin seems to understand you. Write a letter to him, telling him about the excursion.*
>
> (Heaton 1988²: 137-138)

4.3.2.3. *Holistic/ analytic scoring*

It goes without saying that the main problem with subjective tests is that of scoring. Two procedures have been suggested:

a) *Holistic scoring*: this is also called 'impression marking' or 'impressionistic scoring'. Although especially with expert teachers/ testers it seems to be very reliable, it is recommended that there should be more than one scorer (probably three or four). Banding systems, equivalent to those referred to when speaking about criterion-referenced tests, are used. These bands may include a very short description of the level or a longer specification of the abilities the student shows. An example of the former is the following:

NS	*Native speaker standard*
NS-	*Close to native speaker standard*
MA	*Clearly more than adequate*
MA-	*Possibly more than adequate*
A	*ADEQUATE FOR STUDY AT THIS UNIVERSITY*
D	*Doubtful*
NA	*Clearly not adequate*
FBA	*Far below adequacy*

For more detailed banding systems those by the British Council or the ACTFL (American Council for the Teaching of Foreign Languages) can be good examples (Hughes 1989: 87-91).

b) *Analytic scoring*: it consists in giving a separate score for each of a number of aspects of a task, that is to say, the different aspects are scored separately and, then, the average constitutes the final mark. This system is advantageous in compelling the scorer to consider a number of aspects that might otherwise be ignored. In addition, the higher the number of partial scores, the more reliable scoring is. On the debt side, it takes longer (especially difficult in an oral interview because there is not enough time to mark all the aspects while it is being carried out and, furthermore, it can be discouraging for the interviewee) and it seems not to pay attention to the overall effect.

This can be a possible suggestion for oral interview and compositions. In the first case mechanics refers to pronunciation and organization, while in compositions it refers to spelling and organization:

	1	*2*	*3*	*4*	*5*
Grammar					
Vocabulary					
Mechanics					
Fluency					
Relevance					

(For more details see Hughes 1989: 91-93 and 95-96)

As for the question of whether to use one system or the other, we are with Hughes (1989) in the sense that the choice depends on the purpose of the testing ('analytic' is more adequate for diagnostic tests) and on the circumstances: if we deal with a small, well-knit group of scorers, holistic scoring is preferred, while the analytic system would be recommended for a heterogeneous group of scorers in different places. Conversely, a teacher who tests some students for the first time or with elementary students will prefer analytic scoring, while once he/ she gets to know a small group very well or with advanced students, holistic may be preferred. Anyway, what seems to be clear is that multiple scoring (both methods and different scorers) is desirable.

4.4. *Content*

We refer here to the type of elements to be tested. If it concerns isolated points, we speak of discrete-item tests but if linguistic competence and performance as a whole, including the different skills and linguistic components, are tested, we have

integrative or global tests. In actual practice, there are no purely discrete-point or integrative tests and we must say that both of them are useful.

4.4.1. *Discrete-item tests*

They respond to the underlying assumption that language can be broken down into its component parts and those parts tested in turn. These components are the four skills (listening, speaking, reading, writing) and the different linguistic components (phonology, graphology -spelling-, grammar -morphology and syntax- and vocabulary), together with subcategories within these units. Accordingly, tests are devised in order to assess just one of these components. They have received several criticisms, especially from Oller (1976, 1979).

It is an analytical conception of language and testing, in the sense of considering only one point at a time. There are arguments 'for' and 'against' this type of tests (cf. Els *et al.* 1984: 321). Among the former, we should mention the following: they are suitable for testing linguistic competence, especially in the initial stages of the learning process; they serve for diagnostic purposes and they are very useful when a high level of accuracy is required.

Arguments against discrete-item tests refer to the fact that language proficiency is more than just the sum of discrete elements. Apart from that, there seems to be no sense in isolating elements from their context because it is impossible both to compile all the elements of a language and to assess the contribution of individual items to the whole.

4.4.2. *Global/ integrative tests*

The underlying theory is clearly expressed by Oller (1976, 1979) who argues that language is a unified set of interacting abilities which cannot be separated apart and tested adequately. So, integration is required. In this sense, integrative tests attempt to assess a learner's capacity to use many elements all at the same time. All components of language are integrated and tested in combination in a meaningful context. Oral interview and composition, previously discussed, are good examples of global tests. We will now consider cloze, dictation, editing tasks and translation.

4.4.2.1. *Cloze*

It is based on the Gestalt theory of 'closure' (closing gaps in patterns subconsciously). Although originally designed to check the reading difficulty of a passage, it soon revealed itself as a useful test to assess overall competence, because it tests linguistic, textual and world knowledge and it is related to global skills, that is to say, it implies the three linguistic meanings Fries (1963) identifies: grammatical,

lexical and socio-cultural, and some abilities (grammar, vocabulary, discourse reference) are required.

It consists in a passage where every nth word (generally every sixth, seventh or eighth) is deleted and the student has to write it. The shorter the distance between deletions, the more difficult the test is. We can make variations on this basic scheme by using what Weir (1988) calls 'selective deletion': the place of the deleted word can slightly change according to the test writer's preferences in order to create interesting items. In fact, this is a sort of 'gap-filling exercise' or 'impure cloze'. For some time, it was considered a language testing panacea but later its validity was brought into question.

The purest form is without any cues but especially with elementary or intermediate students we can use a sort of guided cloze where the deleted words are given after the passage in different order. As far as correction is concerned we can use three methods: a) the exact-word method (objective scoring), b) the acceptable-word method (it is more psychologically reassuring for the testee but agreement should be reached as regards the acceptable alternatives), and c) what is known as 'clozentropy' (Darnell 1970): by using native speakers' responses on a test as the norm (complicated and impractical).

It is basically written though it can also be oral ('cloze dictation'). We can also have several options for each deletion ('multiple choice cloze').

The following advice on creating cloze type passages can be given: the chosen passages should be at a level of difficulty appropriate to the people who are to take the test; some two or three lines at the beginning should have no blanks for students to get some context; scoring is easier if the blanks are numbered and an answer sheet is provided, although filling the blanks in the text itself approximates more closely to the real-life tasks involved (Heaton 1988[2]: 17) and blanks should be of equal length. The length of the text depends on several factors, among them the level of the students and the amount of time and number of marks to be allotted. For a lower-intermediate class Dangerfield (1985b: 157) suggests 20 spaces to fill in 15 minutes in a one-hour test. If blanks are every seventh word the text would be around 140 words. Heaton recommends ideally 40 or 50 blanks. The more blanks contained in the text, the more reliable the cloze test will generally prove.

A variation of the cloze test is the so-called 'C-test' (Hughes 1989: 71). Instead of whole words, it is the second half of every second word which is deleted. Here only exact scoring is necessary and shorter passages possible. Anyway, it runs the risk of looking like a puzzling activity.

4.4.2.2. *Dictation*

Although sometimes offered as a listening test, in fact it also involves at least the skills of writing and reading. The procedure to follow is familiar to many teachers: the text is read three times: the first time, the teacher reads the text throughout and students only listen; the second time, the text is read in chunks (less than seven words

should be avoided) and with punctuation marks, each chunk is repeated and students write; there is a third reading for students to check. They are generally given two minutes for a final checking.

The practice of dictation has been rejected or accepted throughout the history of ELT. Lado and Oller's controversy has been the clearest expression of the matter (referred by Morrow 1977: 20-21). In fact, Oller has given back credibility to dictation.

There are some variations on the traditional system, such as 'cloze dictation' (already mentioned), the 'noise test' (with interferences, resembling real life listening), 'forced imitation' (oral repetition or oral summary) and 'dicto-comp' (students listen to the dictation and later write a composition about it). Martínez López (1989) mentions many of the advantages it may have for students and teacher. For students: practice in note-taking, associating sound and spelling, discovering things which are not heard, learning from errors on a feedback session, reinforcing learning and the possibility of self-correction. Among the advantages for the teacher we can quote the following: it can be used with large classes, it is quick to prepare and administer and easy to score, many things can be asked for in a short time, it constitutes a source of information for problem areas and a good reference of the general progress of students. In this sense, it can be taken as the basis for a list of common errors. Once again, we see the fruitful relationship between testing and teaching.

The usual marking procedure is to allocate a number of marks to the whole dictation and to take one mark off for every error (negatively). Other more positive-oriented systems, though less useful and practical, have been suggested (Harrison 1983: 114): a) dividing the dictation into sense-groups and marking each group on a 2-1-0 communication-correctness basis or giving one mark per group demanding absolute accuracy, and b) awarding one point for each correct word (time-consuming).

Many of the problems students have with dictation are probably caused by inappropriate, irrelevant and unlikely-to-be-dictated passages: *"The worst kind of dictation test is when students have to try to write down a dull and unfamiliar passage and then be insulted by having one mark deducted for each mistake"* (Hubbard *et al.* 1983: 277). Haycraft (1978) recommends that material for dictation should be that which is useful and likely to be dictated: (answer) phone messages, letters, shopping lists, instructions... Anyway, extreme positions (Morrow 1979) consider dictation does not give *"any convincing proof of the candidate's ability to actually use the language, to translate the competence (or lack of it) which he is demonstrating into actual performance in ordinary situations, i.e. actually using the language to read, write, speak or listen in ways and contexts which correspond to real life"*.

4.4.2.3. *Editing tasks*

Also called 'intrusion tests' (Hubbard *et al.* 1983: 281). It is the converse of the cloze test: additional words, alien to the text, are included and students have to delete them. Errors may come from non-identification and misidentification.

4.4.2.4. *Translation*

Severely criticised though it has been, especially if used in an almost exclusive way (Grammar-Translation method), translation, both direct and inverse, is one of the global or integrative tests because it involves several linguistic aspects, as well as the deciphering of the original author's message (reading) and the encoding of this message for other receptors (writing or speaking). Not only the superficial meaning should be translated but also any underlying emotional, aesthetic or cultural meaning.

Apart from other teaching functions and uses, we will emphasize its usefulness for testing. With translation from the target language to the mother tongue we can test comprehension of details, logical connectors and abstract concepts. When translating into the target language, grammatical rules and vocabulary can be tested. A contextualized text is always preferred, although some isolated sentences can be very useful to test specific grammar points.

The main criticism has come from the fact that translation establishes an intermediate process between the concept and the form in which it is expressed, thus hindering the development of the ability to think directly in the foreign language. This is true, but, in fact, this intermediate process is really difficult to avoid, especially in beginning and intermediate levels. From our personal point, translation should have its place in testing, especially with advanced students.

4.5. *Specific testing method*

According to the format used, several kinds of tests can be distinguished. Some of these formats have already been mentioned. First of all, tests can be oral or written. Features of the spoken and written language obviously affect performance in tests. With regard to the type of response we have the following types of tests:

a) Close-ended: true-false, matching, transfer of information from tables or charts, multiple-choice, cloze (if the exact-word scoring is used).
b) Restricted-response: gap-filling, cloze (with acceptable-word scoring).
c) Open-ended tests: extended writing/ speaking, completion, questions (yes/ no, wh-, open-ended).

More often than not, good tests include tasks with different formats.

5. TESTING THE SKILLS

5.1. *Testing Listening*

It can be tested alone, though very often it also involves speaking (think of oral answers to listening comprehension) and it always has a spoken (live or recorded) stimulus. Some recommendations are going to be made (cf. Hughes 1989: 134 ff) but,

first of all, we should say that when testing listening this must be our primary purpose (for example, answers to listening comprehension tests can even be given in the mother tongue just to check understanding, not language production).

The material should be as authentic as possible and the recordings should be natural (with fillers and pauses) and with good quality. In order to write the items, we should keep in mind that with extended listening items should be kept sufficiently far apart in the passage and that students should be warned by key words. Next, time should not put pressure on candidates. If we just want to test oral comprehension, items and responses can be written in the native language. We should try and avoid setting questions which require the memorisation of individual words in sentences. When administering the test, it is helpful if the speaker can be seen by the listeners.

Among the possible techniques, we have to distinguish between sound discrimination and sensitivity to stress and intonation exercises, on the one hand, and listening comprehension on the other. The former will be dealt with in some detail when studying 'testing pronunciation'. Useful though they are, *"the ability to distinguish between phonemes does not in itself imply an ability to understand verbal messages"* (Heaton 1988[2]: 64). Hubbard *et al.* call this type of exercises 'pure' listening tests and express the main use of them: *"one reason for testing is not so much to discover error as to bring predictable errors to the surface for remedial attention"*.

For listening comprehension we can use the following types of exercises: multiple choice (with short and simple options), question and answer, statements for completion, information transfer (helped by visuals: labelling of diagrams or pictures, completing forms, following directions on a map -*'taking a pencil for a walk'*, cf. Hubbard *et al.* 1983: 265), note-taking (integrated with writing) and (partial) dictation.

With beginners we can use task responses (students do something after listening: drawings, following directions, physical response -TPR-), choosing the best statement about a picture, choosing the best figure from some statements.

In order to test extended communication, we can, if constraints permit, use talks and lectures, movies, radio and TV programs. For short lecture contexts, Madsen (1983: 138 ff) recommends beginning with a reading and build in natural hesitations, rephrasings, little disgressions and some redundancy. Three or four brief lecturettes are more effective than one long lecture. Radio and TV commercials are very useful but we should ask general questions and avoid small details.

We can use similar techniques for testing listening and reading comprehension but for the former, texts should be shorter and questions (not too many) simpler. In addition, we should avoid giving students too much to write (cf. Doff 1988: 262-263).

5.2. *Testing Speaking*

Many testing experts and teachers coincide in mentioning the difficulties in testing the speaking skills. Madsen (1983: 148) mentions some of them: how to test

fluency, how to get students to speak, how to evaluate so many things at once and, in addition, the practical problem of having to test each student individually. As we said with listening, sometimes it is neither possible nor desirable to separate the speaking skills from the listening ones. In spite of the obvious problems of scoring (highly subjective) and administration, we have to admit the necessity of its testing, especially nowadays when the ability to produce language is a requisite of the communicative trend. Apart from its importance, as Doff (1988) suggests, oral tests should be given from time to time to give seriousness to this skill and also to parallel the importance given to it in class and in our methodology.

The oral test should not be improvised and we should try and make students feel at ease, including major areas and interesting topics and not talking too much ourselves (cf. Hughes 1989: 105-107). For beginners we can use imitation exercises (repetition of sentences), directed requests, reading aloud and directed-response role-play. Paraphrase (combining speaking with either listening or reading and with the help of pictures), guided role-play (with prompts) and split dialogues are useful with intermediate students. At an advanced level we can set oral interviews, speaking from tape-recorded stimuli, short talks, group discussion (especially with consensus-seeking activities) and role-playing.

5.3. *Testing Reading*

The same as with listening, we have to bear in mind the different reading subskills, such as scanning the text to locate specific information, skimming to obtain the gist, identifying stages of an argument or identifying examples presented in support of an argument. Although we can test the reading ability and pronunciation through reading aloud, we are going to concentrate our attention on reading comprehension tests. The Barrett taxonomy of skills distinguishes five levels of comprehension (cf. Hubbard *et al.* 1983: 266-267): literal (information explicitly stated in the text); reorganization (to summarize information or handle it in a different sequence); inferential (to go beyond the immediate text); evaluative (to make judgements about the text) and appreciative (emotional/ personal/ aesthetic/ literary appreciation of the text). Tests can focus on one or more of these levels.

A general principle, also applied to listening, should be borne in mind: if the aim is to test reading or listening skills, students should not be asked to write too much. The questions should test the main message, not details, and students should not be able to guess the correct answer without understanding the text. Texts should be interesting, not too culturally loaded and sufficiently general while at the same time not allowing comprehension to be shown simply by students' reference to their own general knowledge.

Some formats are the following: matching exercises (especially in the initial stages and intermediate levels: word matching, sentence matching, pictures and

sentence matching), reading comprehension questions of different types, information transfer with the help of visuals (tables, maps, pictures), completion exercises, cloze test, identifying order of events, topics or arguments, identifying referents, guessing the meaning of unfamiliar words from context and editing texts.

As for scoring, we must test only comprehension and the reading ability. We should not test productive skills at the same time (grammar, spelling, pronunciation).

5.4. *Testing Writing*

There are several things to test within the writing skill: language use, mechanics (punctuation, spelling), content, stylistic skills, judgement skills (*"the ability to write in an appropriate manner for a particular purpose with a particular audience in mind, together with an ability to select, organise and order relevant information"*, Heaton 1988[2]: 135). This is why writing makes such considerable demands on students.

A general principle when testing the skills is not to expect from students skills they do not possess in their own language. This is particularly relevant when dealing with writing. In addition, all the recommendations we made about 'composition' are applicable here, that is to say, meaningful situations, something to say, a purpose and an audience. To these we will add some others. As far as possible, we should test the students as regards the writing ability and nothing else (many times we also ask for creation, imagination, intelligence and general knowledge, although in practice it is very difficult to separate these from writing itself). Although the context should be clearly given in instructions, these should not be too long because, otherwise, it also becomes a test of reading. Students should not be allowed to go too far astray and, in a way, they should be restricted as to what to write or the test can be unreliable and scoring difficult. In fact, some authors say it is not advisable to allow students a choice of composition items (at least for achievement tests).

With regard to the different types of exercises we have to distinguish three stages (cf. Madsen 1983: 101 ff):

a) Controlled writing: sentence-combining, sentence-expansion, sentence-reduction, copying, oral cloze, conversion exercises, easy dictation passages.
b) Guided writing: writing the previous/ following sentence, framework essay (the outline of the story is given), split dialogues, changing a narration into a dialogue or the other way round, changing a passage (grammatically, stylistically, changing the point of view, adding further information, using linkers), building from a paragraph outline, cloze, dictation.
c) Free writing: expositive writing, narration, description, argumentative writing, letter writing.

As far as scoring and correction are concerned, we can use the 'holistic' or the 'analytic' procedure and a new attitude towards written errors should be adopted,

avoiding over-correction and negative marking. We should concentrate on some areas and look for strengths as well as weaknesses.

6. TESTING THE LINGUISTIC COMPONENTS

6.1. *Testing grammar*

Despite the fact that with grammar tests we assess the ability to recognise or produce correct forms of language rather than the ability to use language to express meaning, we must agree that the testing of grammar is necessary. Grammar is, in fact, the skeleton of a language. W. M. Rivers (cf. Arnold 1991), in an interview in Seville, said: *"... many specialists have been saying that we don't need to teach grammar. But grammar is there. It is the framework within which the language is operating. It is like saying that you can have a chicken walking around without bones"*. The same can be said for testing, where grammar presents the advantage that large numbers of items can be administered and scored within a short period of time. In addition, the lack of grammatical ability is an obstacle to skills performance and, thus, also accounts for the necessity of its testing.

The format of grammar tests is familiar to many teachers, because most traditional tests only included grammar items. These formats are still valid, but items should be made to sound as natural as possible (avoid 'lab sentences') and contextualized. Among the possible types of exercises we find: multiple-choice, recognition, rearrangement, completion, transformation, items involving the changing of words, 'broken sentence' items, pairing and matching, combination and addition.

6.2. *Testing vocabulary*

Knowledge of vocabulary is essential to the development and demonstration of linguistic skills. Something we must agree on from the very beginning is the need for contextualization. As for the lexical items to be included in the test we should select from the vocabulary taught in class and with the help of frequency lists.

The first thing we have to decide is if we want to test active or passive vocabulary, that is to say, production or recognition. If we are testing vocabulary, this should be our primary purpose and nothing else: *"Tests of vocabulary should avoid grammatical structures which the student may find difficult to comprehend. Similarly, tests of grammar should contain only those lexical items which present no difficulty to the students"* (Heaton 1988[2]: 52).

One of the classic formats for testing vocabulary has been multiple-choice items. In this sense, it is worth remembering some guidelines: if the stem is difficult the

options should be easy and viceversa; each option should belong to the same class as the word in the stem; key and distractors should be the same level of difficulty, refer to the same area of meaning and have approximately the same length.

Apart from multiple-choice, other types of exercises are the following: making sets of associated words, matching items, objective items (word formation, items involving synonyms, rearrangement of letters to form a word, definitions) and completion.

To finish, we must refer back to the necessity of testing grammar and vocabulary but always considered as means and not as ends in themselves.

6.3. *Testing pronunciation*

Tests of phoneme discrimination and of sensitivity to stress and intonation were mentioned when dealing with listening. Likewise, tests of phoneme production were implicit in reading aloud. It is the case that pronunciation is rarely tested exclusively but with listening or speaking. Among the types of exercises we can mention phoneme discrimination tests (with or without the help of pictures), tests of stress and intonation and tests to understand statements and dialogues.

The language laboratory is a useful teaching aid for these tests. However, they have the disadvantage of being rather artificial and usually devoid of context.

6.4. *Testing functional language*

It is a practical attempt to communicative testing, dealing with the functional aspects of a language. In an excellent article, Mary Spratt (1985) speaks about some methods to test functional language and their implications. The aspects tested are the following: the concept of functions, the form, meaning and degree of formality of exponents of functions, social meaning and appropriateness to different situations.

These are some of the methods she proposes: reading functions and matching with degrees of formality, expanding a discourse chain into a dialogue, odd man out (eliminating one exponent from a list of exponents: different degree of formality or different function), writing parallel texts with different degrees of formality, appropriate responses to given situations, rewriting a conversation from the description of it, written role-play, multiple choice (to choose the best exponent of a function or the appropriate degree of formality), and split dialogues.

WORKS CITED

Alcaraz, E. & J. Ramón. 1980. *La evaluación del inglés: teoría y práctica.* Madrid: SGEL.

Alderson, J. C. 1981a. "Introduction" in Alderson, J. C. & A. Hughes (eds.). 1981: 5-8.

Alderson, J. C. 1981b. "Reaction to Morrow paper (3)" in Alderson, J. C. & A. Hughes (eds.). 1981: 45-54.

Alderson, J. C. 1981c. "Report on the discussion on communicative language testing" in Alderson, J. C. & A. Hughes (eds.). 1981: 55-65.

Alderson, J. C. & A. Hughes (eds.). 1981. *Issues in Language Testing. ELT Document 111.* London: The British Council.

Allen, J. & A. Davies (eds.). 1977. *Testing and Experimental Methods. The Edinburgh Course in Applied Linguistics. Vol. 4.* London: O.U.P.

Arnold, J. 1991. "Reflections on Language Learning and Teaching: an interview with Wilga Rivers". *English Teaching Forum* 24/1: 2-5.

Bachman, L. F. 1990. *Fundamental Considerations in Language Testing.* Oxford: O.U.P.

Bachman, L. F. & A. S. Palmer. 1981. "The construct validation of the FSI oral interview". *Language Learning* 31/1: 67-86.

Baker, D. 1989. *Language Testing. A Critical Survey and Practical Guide.* London: Edward Arnold.

Bell, R. T. 1981. *An Introduction to Applied Linguistics.* London: Batsford Academic and Educational Ltd. (Appendix C, "Language testing").

Bestard Monroig, J. & Mª. C. Pérez Martín. 1992. *La didáctica de la lengua inglesa. Fundamentos lingüísticos y metodológicos.* Madrid: Síntesis. (Capítulo 21, "La evaluación").

Brown, H. D. 1987[2]. *Principles of Language Learning and Teaching.* New Jersey: Prentice-Hall Regents. (Chapter 11 "Language Testing").

Canale, M. & M. Swain. 1980. "Theoretical bases of communicative approaches to second language teaching and testing". *Journal of Applied Linguistics* 1/1: 1-47.

Carroll, J. B. 1961. "Fundamental considerations in testing for English language proficiency of foreign students". *Testing.* Washington, D. C.: Center for Applied Linguistics.

Carroll, J. B. 1980. *Testing Communicative Performance. An Interim Study.* Oxford: Pergamon Institute of English.

Carroll, B. J. & P. J. Hall. 1985. *Make Your Own Language Tests. A Practical Guide to Writing Language Performance Tests.* Oxford: Pergamon Press.

Dangerfield, L. 1985a. "Writing achievement tests: practical tips" in Matthews, A. *et al.* (eds.). 1985.

Dangerfield, L. 1985b. "Three types of objective tests" in Matthews, A. *et al.* (eds.). 1985.

Dangerfield, L. 1.985c. "Making extended writing tests less subjective" in Matthews, A. *et al.* (eds.). 1985.

Darnell, D. K. 1970. "Clozentropy: a procedure for testing English language proficiency of foreign students". *Speech Monographs* 37: 36-46.

Davies, A. 1978. "Language Testing: survey article". *Linguistics and Language Teaching Abstracts* 11/ 3-4.

De Jong, J. H. A. L. (ed.). 1990. *Standardization in Language Testing. AILA Review, No. 7.* Amsterdam: Free University Press.

Doff, A. 1988. *Teach English. A Training Course for Teachers.* Cambridge: C.U.P. (Unit 22 "Clasroom tests").

Els, T. van *et al.* 1984. *Applied Linguistics and the Learning and Teaching of Foreign Languages.* London: Edward Arnold. (Chapter 15 "Language Testing").

García-Zamor, M. & D. Birdsong. 1977. *Testing in English as a Second Language: a Selected, Annotated Bibliography.* Washington, D. C.: TESOL.

Harrison, A. 1982. *Testing in Language Teaching.* London: Macmillan.

Harrison, A. 1983. *A Language Testing Handbook.* London: Macmillan.

Haycraft, J. 1978. *An Introduction to English Language Teaching.* London: Longman.

Heaton, J. B. 1975. *Writing English Language Tests.* London: Longman.

Heaton, J. B. 1988[2]. *Writing English Language Tests.* London: Longman.

Hubbard, P. *et al.* 1983. *A Training Course for TEFL.* Oxford: O.U.P. (Chapter 9 "Testing").

Hughes, A. 1988. *Testing English for University Study. ELT Document 127.* London: Modern English Publications.

Hughes, A. 1989. *Testing for Language Teachers.* Cambridge: C.U.P.

Jones, R. L. 1985. "Second language performance testing: An overview" in Hapuptman, P. C., R. Leblanc & M. B. Wesche (eds.). 1985. *Second Language Performance Testing.* Ottawa: University of Ottawa Press.

Lado, R. 1961. *Language Testing. The Construction and Use of Foreign Language Tests.* London: Longman.

Lloyd-Jones & E. Bray. 1986. *Assessment: from Principles to Action.* London: Macmillan.

Madsen, H. S. 1983. *Techniques in Testing.* London: Longman.

Martínez Haro, R. 1984. *Evaluación de conocimientos en inglés. Teoría y Práctica.* Jaén: edición del autor.

Martínez López, M. 1989. *Análisis de errores. Reevaluación del dictado en la enseñanza del inglés como lengua extranjera.* Granada: Servicio de Publicaciones de la Universidad.

Matthews, A. *et al.* (eds.). 1985. *At the Chalkface. Practical Techniques in Language Teaching.* London: Edward Arnold. (Section C: "Achievement testing").

Moller, A. 1981. "Reaction to Morrow paper (2)" in Alderson, J. C. & A. Hughes (eds.). 1981: 38-44.

Morrow, K. 1977. *Techniques of Evaluation for a Notional Syllabus.* London: Royal Society of Arts.

Morrow, K. 1979. "Communicative language testing: revolution or evolution?" in Brumfit, C. J. & K. Johnson (eds.). 1979. *The Communicative Approach to Language Teaching.* Oxford: O.U.P.

Oller, J. W. Jr. 1976. "Evidence of a general language proficiency factor: an expectancy grammar". *Die Neuren Sprachen* 76: 165-174.

Oller, J. W. Jr. 1979. *Language Tests at School.* London: Longman.

Oller, J. W. Jr. & K. Perkins (eds.). 1978. *Language in Education. Testing the Tests.* Rowley, Mass: Newbury House.

Pilliner, A. E. G. 1968. "Subjective and objective testing". *Language Teaching Symposium.* Oxford: O.U.P.

Rea-Dickins, P. & K. Germaine. 1992. *Evaluation.* Oxford: O.U.P.

Spolsky, B. 1975. "Language testing: art or science?". Paper presented at the Fourth AILA International Congress. Stuttgart.

Spolsky, B. 1978. *Approaches to Language Testing. Advances in Language Testing Series 2.* Arlington, Va.: Center for Applied Linguistics.

Spolsky, B. 1985. "The limits of authenticity in language testing". *Language Testing* 2/1: 31-40.

Spratt, M. 1985a. "Achievement tests: aims, content and techniques" in Matthews, A. *et al.* (eds.). 1985.

Spratt, M. 1985b. "Testing funcional language" in Matthews, A. *et al.* (eds.). 1985.

Weir, C. J. 1981. "Reaction to Morrow paper (1)" in Alderson, J. C. & A. Hughes (eds.). 1981: 26-37.

Weir, C. J. 1988. *Communicative language testing.* Exeter Linguistic Studies, vol. 11. Exeter: University of Exeter.

Williams, E. 1985. "Coming to terms with testing" in Matthews, A. *et al.* (eds.). 1985.