PSYCHOLOGY AND EDUCATION (2025) 62(02):75-85
ISSN: 1553-6939

Reconsidering Language Testing in English Language Classrooms

PhD. Malika Kouti

University of Ghardaia (Algeria), E-mail: kouti.malika@univ-ghardaia.dz, kouti.malika@gmail.com

Received: 12/2024, Published: 02/2025

Abstract:

Teaching a foreign language such as English requires more attention and involvement on the part of both teachers and learners; nonetheless, there are some important matters that affect English language teaching such as language testing. In fact, language testing which is part of language assessment (McNamara, 2004) has been less focused on and less given importance. This is due, in the main, to the lack of awareness of its importance in teaching in general and English language teaching in particular. Another factor is the lack of knowledge of what really testing is and its criteria and types. This paper casts light on the relationship between teaching and testing while highlighting the difference between testing and assessment and evaluation besides pointing out the difference between tests and examinations. It also sheds light on tests and their types and on the main criteria of testing while designing tests, namely validity and reliability and ends with stating some testing challenges.

Keywords: language testing, language assessment, tests, examinations, validity, reliability.

Introduction

Language testing is recognised within applied linguistics and has its own journals and its national and international conferences and has witnessed rapid developments in the past 70 years within this field (McNamara, ibid). It is obvious that teachers not only do they teach, but also assess; that is, they need from time to time to check how well their learners have mastered and digested the different aspects of the English language course. To this end, they select many ways to achieve their set objectives through different types of assessment. Testing is part of assessment and is most of the time the only type used by teachers. However, testing is not really understood properly by teachers. They lack either awareness of its importance in English language teaching or knowledge of what testing is exactly and its requirements. The following will focus on language testing as part of language assessment and related elements, namely tests and their types, validity and reliability and testing challenges. Nonetheless, we shall begin by establishing the relationship between teaching and testing and making distinctions between testing and evaluation.

1. Teaching, Testing and Evaluation

In this section, we should stress the relationship between teaching and testing. Moreover, we should highlight the difference between testing and evaluation. Actually, teaching and learning should be evaluated.

1.1.Teaching and testing

Although there exists an intimate relationship between teaching and testing, they are not considered as equivalents. However, according to Davies and Pearse (2000), some teachers

transform teaching into a continuous test. Davies and Pearse (ibid, p. 169) provide the following example from an English language classroom:

- **Teacher:** Where did you go in the holidays, Sofia?
- **Learner 1:** I didn't go anywhere.
- **Teacher:** Very good, very good. And you, Giovanni. Where did you go?
- **Learner 2:** I go to Scotland.
- **Teacher:** no, no, Giovanni, no.

From the above conversation, it seems that the teacher is not teaching but testing.

Teaching should be directed towards building up the learners' ability and confidence in using English for effective communication. Davies and Pearse (2000) stated that "Especially when you are trying to develop fluency, it is very important that learners should not feel that they are being tested all the time" (p.170), for this does not allow them to achieve fluency. Most teaching should not be seen by the learners as a test. However, teachers should evaluate performance and progress of the learners and their own teaching constantly. Therefore, evaluation is essential in teaching.

1.2. Testing and Evaluation

The two concepts of 'testing' and 'evaluation' are used interchangeably; however, they should be distinguished, and the distinction in English is important.

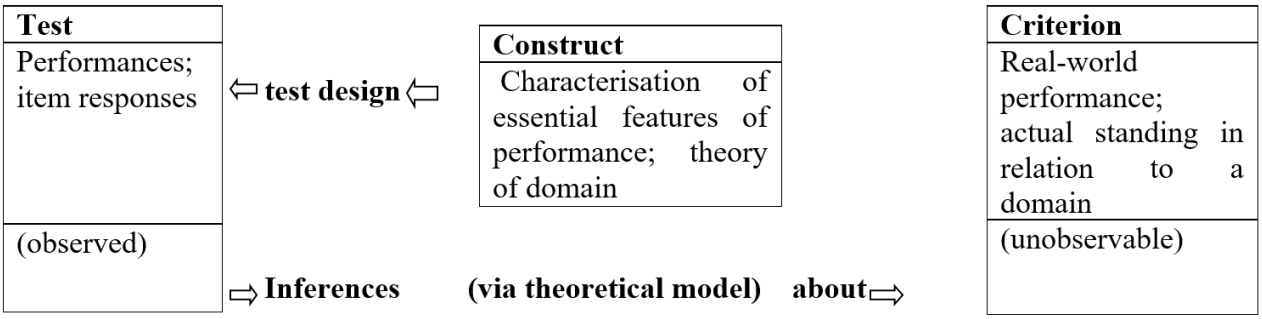
As a matter of fact, 'evaluation' is more general than 'testing'. In Genesee (2004), evaluation is "a process of collecting, analysing and interpreting information about teaching and learning in order to make informed decisions that enhance student achievement and the success of educational programmes" (p. 144). You can evaluate teaching, programmes, teaching materials, tests and learning (ibid). Furthermore, you can evaluate learning in several ways, not only with the formal tests given to the learners.

2. Tests and Testing

When we talk about language testing, it is necessary to define it and explain what a test is and make a distinction between tests and examinations.

2.1. Testing Definition

Testing is the use of tests, or the study of the theory and practice of their use, development, evaluation, etc. Hedge (2000) stated that testing is a term "that is not always used precisely" (p. 378). She defined it as "the specific procedures that teachers and examiners employ to try to measure ability in the language, using what learners show they know as an indicator of their ability" (ibid). Carter (1993) defined *language testing* as "a process by which a student's ability, knowledge, performance or progress in language use can be measured" (p. 68). For McNamara (2004), *language testing* is "a process of gathering information about test-takers from observed performance under test conditions" (p. 765). In the figure below, he explained that



2.2.Test Definition

According to Ingram (1974), a test is a measuring device that is used when comparison of an individual with other individuals who belong to the same group is needed. For example, you may use a language test for comparing them in terms of their mastery of a foreign language.

> Tests … invite candidates to display their knowledge or skills in a concentrated fashion, so that the results can be graded, and inference made from the standard of performance in the test about the general standard of performance that can be expected from the candidate, either at the time of the test or at some future time (ibid, p. 313).

- **Tests versus Examinations**

There is a difference between tests and examinations. This difference lies in the marking (Ingram, ibid); that is, in an examination, the marker must use his judgment, whereas in a test, the marking does not depend on the judgment of any individual. However, some of applied linguists such as Pilliner (1968 as cited in ibid) stated that the marking is the only objective thing in objective tests.

A test should be carefully designed for a specific purpose, whereas evaluation may be spontaneously and very flexibly handled. A test consists of one or more exercises or tasks, each with a clear objective, learning's evaluation "usually employs formal tests, but it may also include other options". When the evaluation of learning is based on class participation, progress tests, homework, and projects rather than final tests alone, this is referred to as 'assessment' or 'continuous assessment'.

Despite this fact, Hedge (2000) stated that

> tests are the main instruments for evaluation of learning in most teaching situations. "Good tests provide the opportunity for learners to show how much they know about language structure and vocabulary, as well as how they are able to use these formal linguistic features to convey meanings in classroom language activities through listening, speaking, reading, and writing. Tests of this type may be used as part of an integrated assessment system (p. 378).

2.3.Approaches to Testing

There are two main approaches to testing, according to Hedge (ibid). These are: the structuralist influence and the communicative influence.

2.3.1. The Structuralist Influence

This approach to testing appeared in the late 1950s and early 1960s. It was based on knowing a language by knowing its structures or forms and the linguistic elements of lexis and phonology. This approach was influenced by the works of Lado and Caroll (1961) that affected test design. The test was considered as a set of separate parts such as discrete-point items as the following example from Heaton (1989):

Underline the correct option.

He may not come, but we'll get ready in case he …

A. will B. does C. may (Cited in Hedge, ibid)

This type of tests focuses only one aspect of language at a time, and is decontextualised. In fact, this type is also called 'objective' test.

2.3.2. The Communicative Influence

According to Hedge (2000), the most influential approach to language testing suggested by applied linguists was the one of Canale and Swain (1980). They built their work on Hymes's communicative competence and identified its types, namely grammatical competence, sociolinguistic competence and later, discourse competence (Canale, 1983). Another influential work was the one of Bachman's (1990) model of communicative language ability (McNamara, 2004). The tests of listening, speaking, reading and writing under this model identified a number of features that were not measured in conventional tests. These features characterised language use as interaction-based and stimulated by a specific communicative purpose. They incorporated important elements of authenticity of language context, situation and topic. The communicative purpose and language function was clearly specified in tests. Morrow (1977) illustrated the following:

- Prior to the test the candidate is given a card which outlines a situation. He must invent a number of questions he would want to ask in that situation. For example:

You arrive at Victoria Station in London to catch the train to Paris. You want to find out something from each of these people. What would you say?

- A passer-by in the station entrance
- The booking clerk
- Your friend who is travelling with you
- The ticket collector

Here is another suitable example for a reading test from Morrow (1977 Cited in Hedge, ibid, p. 380):

> Everything you see in the shops can be bought to take away, with the exception of certain items of furniture and display stocks. Where smaller items of furniture are held in stock, these items will be clearly marked 'Take Away'. To cut down handling charges and the risk of damage, the larger items will be delivered direct to your home from our warehouse. When you place an order for furniture you will be given a

delivery period. If, for any reason, we cannot keep to this delivery period, we will advise you … Enquiries about the furniture order should be made to the Furniture Manager of the shop where you placed the order.

1 What is this about?	a. Buying furniture
	b. Take-away furniture
	c. Large items of furniture
2 What is the purpose of the passage?	a. To give information
	b. To apologise
	c. To promise

The examples of questions above are for learners at lower levels and are expected to answer. The following will about

3. Why testing?

According to Burgess and Head (2005), all students in exam classes need to be tested for various purposes, at different stages of their course. Edge and Garton (2009) stated that "The overall purpose of testing is to provide information about ability and about the learning and teaching process" (p. 161). Therefore, the purpose of testing is related to the objectives of the course, and requires from teachers testing their learners.

3.1. Basic Aspects of Testing

As we are focusing on testing, it is important to have an idea about the different types of tests, the criteria of testing and testing challenges.

- **Types of Test**

The purpose of English language tests is to gather reliable evidence of what learners can do in English and what they know of English. This information may be required to different reasons, and these reasons govern the type of tests used. McNamara (2000) pointed out that not all language tests are similar. They differ in method and purpose.

Concerning the method, a distinction is between traditional papers-and-pencil language test and performance tests. In the former, there is assessment of separate components of language knowledge such as grammar and vocabulary or of receptive understanding such as listening and reading comprehension. The test items are in fixed response formats such as multiple choice formats. Performance tests are in the main tests of speaking and writing in which extended samples of speech and writing are elicited from test-takers and rated using an agreed rating procedure, taking into account that "these samples are elicited in the context of simulations of real-world tasks in realistic contexts" (p. 6).

In terms of purpose, the main distinction is between achievement and proficiency tests; in that, achievement tests are related to the process of instruction. That is, they aim at checking the progress of learners during or at the end of a course. Proficiency tests

- **The criterion**

According to Davies and Pearse (2000), Edge and Garton (ibid) and Harmer (2012) and other researchers, there are five common types of test, each with a specific purpose. The following table summarises them.

Type of test	**Purpose**
Placement test	To place new students in the appropriate course or level. These are essential in large institutions that frequently receive new students.
Diagnostic test	To find out learners' strengths and weaknesses at the start of a course. They allow the teacher to adjust his or her teaching to the needs of the group and individual learners. They are especially useful with mixed – level groups.
Progress tests (short-term achievement tests	To check how well learners are doing after each lesson or unit, and provide consolidation or remedial work if necessary. They usually focus on language that has recently been introduced and practised.
Course tests (longer-term achievement tests)	To check how well learners have done over a whole course. These are the commonest basis for the marks teachers give learners at the end of each course. They are also the main concern in testing for most classroom teachers.
Proficiency tests	To determine learners' levels in relation to generally accepted standards. These are useful for the objective evaluation of learning, and also for the indirect evaluation of course design and teaching. The two best known systems of internationals proficiency tests are the UCLES exams and the TOEFEL tests.

3.2. Designing Achievement Tests

When designing tests, teachers should consider a number of parameters, namely validity and reliability (McNamara, 2000). Professional test development and management involve "validity" and "reliability", and the relationship between them.

Achievement tests are associated with the process of instruction at the end of a semester or a year (Harmer, 2012; McNamara, ibid). They aim at, as mentioned above, checking how well learners have done over a whole course. These tests can be considered to have validity if:

- they contain only forms and uses the learners have practised in the course.
- they employ only exercises and tasks that correspond to the general objectives and methodology of the course.

Harmer (ibid) illustrates as follows:

> Tests need to have validity. This means that if we tell the students that we are going to assess their writing, we shouldn't make it dependent on a lot of reading because they were not expecting a reading test. When we make achievement tests, we need to test things that the students have been learning (grammar, vocabulary, etc.), and we have to be sure that we use the same kinds of test items and tasks as the ones they have been using in their lessons (p. 195).

As mentioned above, a valid test, in short, should consist of:

- things that students have been learning and
- the same kinds of items and tasks they have been using in their lessons.

3.3.Validity and its Types

According to Ingram (1974) and other researchers, there are three main types of validity: content, construct and face, as follows.

- **Content Validity**

Ingram (ibid) states that "If a test samples adequately the 'content' of a subject, for instance as defined by a syllabus or a textbook, it has content validity" (p. 329-330). This means that the grammar, vocabulary, and functional content of a test should be carefully selected on the basic of the course syllabus. For example, if the students have not practised the present perfect test, they should not be tested on it. Only when it is not relevant to the exercise or goal can the language content of the test deviate from the syllabus. For instance, in a reading comprehension examination, students may have been urged to overlook extraneous information they are unfamiliar with or to infer its meaning from context.

- **Construct Validity**

Ingram (1974) pointed out that "If a test is constructed in accordance with a relevant theory, then it has construct validity" (p. 330). This means that the exercises and tasks in a test should be similar to those used in the course and correspond to the general approach of the course. If the learners have never practised translating on the course, they should not have to translate a passage in the test. If the main aim of the course has clearly been to use grammar in natural discourse such as conversations, the grammar should not be tested only through grammar manipulation tests.

- **Face Validity**

According to McNamara (2000), face validity refers to "its surface acceptability to those involved in its development or use" (p. 50). In fact, face validity is about the extent to which a test appears to measure what it is intended to measure. In this case, if a test meets those expectations, we could say that it has strong face validity.

In sum, if a test conforms to these principles, it will probably be seen as fair by the teachers and the learners. If it does not, it will probably be considered unfair, and justifiably so.

3.4.Reliability

Reliability refers to the consistency of the scores that teachers give to learners (McKay, 2006). The idea is that we need to check that students get the same scores whether we or other assessors would give them for the same activity or question. Reliability, actually, refers to how far we can believe or trust the results of a test. As a teacher, you may question the reliability of a test when two of your own groups that you consider very similar in ability and achievement get very different results in the same test, one group doing well and the other badly.

According to Ingram (1974) a test "should give the same results every time it is used on the same objects or individuals, regardless of who is giving and marking it" (p. 313). For the author, Ingram, results are not reliable if they are not stable. Assessors can measure the stability of a test by

giving the test to a group of people and giving it to them again after a short time and then correlating the scores. She mentioned that a statistical method for determining how closely two sets of scores match one another is correlation. We would have a "test re-test" or "stability correlation of reliability" in this situation. The "reliability of equivalence" is another type of reliability that statisticians use. When employing one type of measuring instrument, the findings should be very similar to what one would have obtained with another, comparable measuring device.

A specific test exercise or task is normally reliable when:

- the instructions are clear and unambiguous for all learners.
- the exercise or task controls to some extent how learners respond, for example, it should be clear in 'fill the gap' exercises whether a single word or phrase is required.
- there are no errors in the test, for example, if the learners have to 'select the best answer—a, b, c, d' , there should not actually be two or more acceptable answers.

(Davies & Pearse, p. 173)

The reliability of a test also depends partly on how far it can be marked objectively. Multiple choice exercises, where the learners have to select the best answer from a choice of three or four, are purely objective by nature. One-word fill-in exercise—completion of a text with one word in each space---are purely objective when only one word is possible. But when many different words are possible, they are fairly subjective, requiring teachers to use their personal judgment. Composition marking is by nature highly subjective.

Furthermore, the reliability of a test depends on its length and on how it is administered. A long test is usually more reliable than a short one. Any test provides a sample of learners, and a small sample of something is less reliable than a large one. The following should be considered:

- One group is given much more time than another.
- One group is helped by the teacher and another is not.
- Invigilation is strict in one group and not in another, so that there is a lot of copying or other types of cheating in the second group.

3.5.Balancing Validity and Reliability

A valid test for a course with communicative objectives should consist of exercises and tasks in which the learners use language in realistic contexts. For example, they could complete a dialogue, write a letter, and role-play an interview. These tasks would test (1) their ability to use specific grammar and vocabulary (the dialogue completion), (2) to use written English effectively (the letter-writing), and (3) understand and produce effective spoken English (the interview).

However, there is often a conflict between validity and reliability. The most reliable types of questions are multiple-choice. The learners produce no English themselves, but only recognize correct language. Their answers can actually be marked by a computer, with no need for any subjective human judgments. The least reliable types of task include precisely the letter-writing and the interview role-play proposed above. These have to be marked subjectively by human beings.

The solution reached by many teachers and institutions is a compromise. Some exercises in the tests should be of an objective, recognition type such as multiple-choice whose answers are not recorded linguistically, just by a tick or a cross in a box, a circle round a number or letter, etc. and

which can be marked by a machine (Broughton et al., 1980). These can cover a range of grammar and vocabulary as well as listening and reading comprehension. Other exercises and tasks should be of a more subjective type, involving production and the communicative use of include the possible answers for fill-in and completion exercise, and criteria for marking composition and interview. This compromise also makes tests more practical. Multiple-choice exercise can usually be answered faster by learners and marked faster by teachers than production exercise and tasks.

4. Writing and Evaluating Achievement Tests

As a teacher, you may have to use course tests provided by your institution, or you may produce your own course tests. If the course tests are provided by the institution, you may still have opportunities to comment on them and make suggestions for modification. In addition, you may want to produce a number of short progress tests. The following ideas should help you write, modify, or give opinions on tests.

Tests should normally be designed for specific teaching-learning situations:

- Some situations may call for more objective language exercises such as 'true'/'false'.
- Others may call for more communicative tasks.
- Some situations may permit quite long tests,
- Other ones may require short, easily administered tests because of a lack of time.

Nonetheless, as Harmer (2012) pointed out, before you start writing a test, you need to list the following:

- what it is you want to measure and
- how to do it

For example, for testing syntax, we may use 'reordering sentences' items, and 'putting pictures in order' for testing comprehension. However, 'reordering sentences' does not test comprehension, and 'putting pictures in order' does not test syntax.

Another point is the balance of items (Harmer, 2012). That is, you need to think about whether you want to include *discrete items* that test only one thing at a time such as a verb tense in all the questions, or you want to include more *integrative* tests such as 'using a variety of items' or where students should read and write.

One of the major points to be taken into consideration is rubrics. Rubrics or instructions should be written carefully and easily understood by the students (ibid). That is, each question should be accompanied by an example that will help students in answering the questions. For example:

Rewrite each of the following sentences using the word in parentheses so that they have the same meaning.

Example:

Adam was late, so he took a taxi. (because)

Adam took a taxi because he was late.

Another point of great importance and which does not happen except for research purposes is piloting. That is, tests can be piloted or tried out by giving them to colleagues or to students who are not going to do the tests afterwards (ibid). This process is going to show any change that could be made before the tests that take place.

5. Testing Challenges

Despite all what has been aforementioned, there exist some testing challenges, as follows :

- Many teachers are unaware of the importance of language testing.
- Many teachers lack knowledge of why to test, what to test and how to test.
- Some teachers misuse some testing methods. For example, in multiple choice questions, they give more than four choices, which is not appropriate.
- Some teachers use only one question for the whole test while neglecting so many points of the syllabus uncovered. It should be noted that the more a test includes many and a variety of questions, the more it is reliable. It is unfair to concentrate on only one lesson or part of a lesson. Teachers should give the opportunity to students to demonstrate their language ability and content comprehension.
- Some teachers argue for composition writing in examinations while they neglect the time constraints especially in the COVID-19 era. Moreover, according to many studies, composition writing is subjective on the part of the corrector. Furthermore, when some teachers favour composition writing, what action are they going to take? Are they going to give feedback to their students, and then help them reduce mistakes or do they just correct and submit the papers to the administration?
- Another challenge is online testing mainly in the pandemic era as the criteria of testing such as fairness, authenticity will be doubted. The educational system should reconsider the electronic assessment hindrances teachers are struggling with.
- The allotted time for examinations has also had a negative impact on the results' expectations. A one-hour exam is never sufficient if we consider the majority of students in terms of individual differences.

Those are some of the challenges that should be thought of and reconsidered by all parties, the administration and teachers.

Conclusion

Because of its great importance, language testing needs to be reconsidered mainly by English language teachers. In reality, there is a lack of awareness and lack of knowledge on the part of many English language teachers. In fact, they are required to gain knowledge and get benefited from the great available body of research about language assessment in general and language testing in particular. Many researchers have investigated the issue of language assessment in general and language testing in particular from different perspectives. Therefore, teachers should select what it suits them and their classes. Validity and reliability are the main criteria of language testing and must be taken into consideration. Moreover, English language teachers should consider the testing challenges stated above.

References

1. Brindley, G. (2001). Assessment. In Ronald Carter and David Nunan. The Cambridge guide to teaching English to speakers of other languages. Cambridge. Cambridge University Press.
2. Broughton, G., Brumfit, C., Flavell, R., Hill, P. & Pincas, A. (1980). *Teaching English as a foreign language 2nd Ed*. London. Routledge & Kegan Paul Ltd.
3. Burgess, S. & Head, K. (2005). *How to teach for exams*. England. Pearson Education Limited.
4. Corder, S.P. (1973). *Introducing applied linguistics*. England. Penguin Books Limited.
5. Davies, P. & Pearse, E. (2000). Success in English teaching. Oxford. Oxford University Press.
6. Edge, J. & Garton, S. (2009). From experience toknowledge in ELT. Oxford. Oxford University Press.
7. Genesee, F. (2001). In Ronald Carter and David Nunan. The Cambridge guide to teaching English to speakers of other languages. Cambridge. Cambridge University Press.
8. Harmer, J. (2012). *Essential teacher knowledge: core concepts in English language teaching*. England. Pearson Education Limited.
9. Hedge, T. (2000). *Teaching and learning in the language classroom*. Oxford. Oxford University Press.
10. Huerta-Marcías, A. (2002). Alternative assessment: Responses to commonly asked questions. In Jack C. Richards and Willy A. Renandya. *Methodology in language teaching: An anthology of current practice* (pp. 295-305). New York. Cambridge University Press.
11. Ingram, E. (1974). Language testing. In J. P. B. Allen and S. Pit. Corder. Techniques in applied linguistics. Oxford. Oxford University Press.
12. Luoma, S. (2004). *Assessing speaking*. Cambridge. Cambridge University Press.
13. McKay, P. (2006). *Assessing young language learners*. Cambridge. Cambridge University Press.
14. McNamara, T. (2000). *Language testing*. Oxford. Oxford University Press.
15. McNamara, T. (2004). Language testing. In Alan Davies and Catherine Elder. The handbook of applied linguistics. USA. Blackwell Publishing Ltd.
16. Nunan, D. (1992). *Research methods in language teaching*. Cambridge. Cambridge University Press.
17. Richards, J. C., Platt, J. & Platt, H. (1992). *Longman dictionary of language teaching and applied linguistics 2nd Ed*. England. Longman Group UK Limited.
18. Rivers, W. M. (1981). *Teaching foreign-language skills* (2nd Ed).Chicago. University of Chicago Press.
19. Spada, N. & Lightbrown, P. M. (2006). How languages are learned 3rd Ed. Oxford. Oxford University Press.