Improving English Test Qualities

Chanika Gampper

Language Institute
Thammasat University

chanivad@yahoo.com

Abstract

English Tests are widely used in classrooms and outside. Many of them draw a lot of criticism for not being an accurate measurement tool while their results are used to make important decisions for stake holders. This article looks into the definition of a test, its five necessary qualities, namely reliability, validity, authenticity, backwash, and practicality and how to improve them so that English tests are good, effect, and useful.

[Thammasat Review, Special Issue, 2013]

Keyword: English Tests, Test Qualities: Reliability, Validity, Authenticity, Backwash, Practicality

Introduction

Although student's performance can be assessed by other means such as portfolio, observation, self/peer assessment, tests are still widely used as an instrument that measures the test-taker'sability or competence in a language. Since test scores are used to make many important education aldecisions such as to accept the student into a school or place him or her into the right class, to pass or fail the student, to evaluate the effectiveness of the teaching and curriculum, teachers need to be sure that the test they use serve as an accurate "yardstick" in order to make meaningful comparisons among the test-takers.

Brown (2004) acknowledges that constructing a good test is a complex task that involves both science and art. Since a test draws on a limited sample of observable behaviors, it sometimes fails to reflect the test-taker's true ability. Other factors may also cause the inaccuracy of tests scores. Developing "good" tests is, therefore, very crucial not only for teachers but also the educational system in general.

This article begins with a definition of a test, followed by five qualities that are necessary for any good language tests. How to improve these qualities in English tests is explained in the same order that the test qualities are presented. Examples are also given to illustrate the points.

What is a test?

In simple terms, a test is defined as a "method of measuring a person's ability, knowledge, or performance in a given domain"(Brown, 2004, p.3). In other words, a test can be a set of techniques, procedures, or items that requires performance on the part of the test-taker. Tests must be explicit and structured, are usually relatively time-constrained, and normally occur at identifiable times in a curriculum.

What are necessary test qualities?

Tests need to have certain qualities so that teachers can justify using test scores—numbers—as a basis for making inferences or decisions. The number of important qualities proposed by scholars is different. For example, Heaton (1988) and Hughes (2003) include reliability, validity and *backwash* in their books. While *reliability* and *validity* are sometimes referred to as "essential measurement qualities,"(Backman& Palmer, 1996, p.19), *backwash* is so important that Hughes (2003) chooses to mentioned it first before anything else. Brown (2004) adds *authenticity* and *practicality* to his list, making altogether five qualities the "cardinal" criteria for "testing a test" (P.19).They are *reliability*, validity, authenticity, *backwash*, and practicality. The order of presentation below does not imply the order of necessity that a good language test must have.

1. Reliability

Reliability is often defined as consistency of measurement (Carr, 2011). If you weigh yourself again without eating or exercising and you get the same number, your scale is reliable. Likewise, if the same test taken by the same student on different occasions without any further instruction yields the same result, the test is said to be reliable. Reliability can be divided into:

1. Test/Re-Test Reliability

1.1Test Administration Reliability: the conditions in which the test is administered such as loud noise, photocopying quality, etc can affect reliability.

1.2 Test Reliability: the nature of the test itself such as test length, ambiguous questions, and many possible answers can be the source of test unreliability.

2. Mark/Re-Mark or Scorer/Rater Reliability

2.1 Interrater Reliability: the extent to which the same marks are awarded if the same tests are marked by different raters or scorers

2.2 Intrarater Reliability: the extent to which the same marks or grades are awarded if the same tests are marked by the same rater on different occasions

3. Student-Related Reliability

"Observed" test scores may deviate from one's "true" score due to the student's physical or psychological conditions such as illness, fatigue, or anxiety (Brown, 2004). Teachers should be aware of these conditions although they cannot directly control them.

2. Validity

The validity of a test is the extent to which it measures what it is supposed to measure and "nothing else" (Heaton, 1988, p. 159).In other words, a test is said to be valid if it measures accurately what it is intended to measure (Hughes, 2003).While Bachman & Palmer (1996) specify construct validity as a quality of test usefulness, others such as Heaton (1988), Brown (2004), and Hughes (2003) also include content validity as a quality of a good test. In general, a test can be considered as valid or invalid only with respect to purpose(s).

1.1 Construct Validity

The word 'construct' refers to any underlying ability or trait which is hypothesized in a theory of language ability such as 'reading ability' and 'control of grammar'. Construct validity, then, refers to "the extent to which we can interpret a given test score as an indicator of the ability (ies), or construct(s), we want to measure" (Backman& Palmer, 1996, p.21).

1.2 Content validity

A test with content validity has a representative sample of the language skills, structures, etc. with which it is meant to be concerned (Hughes, 2003). A comparison of test specification and test content is the basis for judgments as to content validity. If a course has 10 objectives, but only two are covered in a test, content validity suffers.

1.3 Face Validity

Also called perceived validity, it refers to whether the test *looks* as if it is measuring what it is supposed to measure. It is hardly a scientific concept but important because when a test appears to be irrelevant, it may not be accepted by students and others.

1.4 Validity in scoring

Not only the test items but also the ways in which the responses are marked contribute to the validity of a test (Hughes, 2003).If a test aims to test how well the student can write an essay in English, deducting the score just because the handwriting is not neat will reduce the validity of this "writing" test.

3. Authenticity

Backman and Palmer (1996) define authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a target language use (TLU) task" (p.23). In other words, authenticity is "the degree to which test materials and test conditions succeed in replicating those in the target situation" (McNamara, 2000, p.131). Authenticity is necessary for a language test because it allows the teacher to predict more accurately how the candidate will cope, at least linguistically, with real-life activities (Ingram, 2003).

4. Backwash or Washback

Backwash is the effect that tests have on learning and teaching. It generally refers to the effects the tests have on instruction in terms of how students prepare for the test. (Brown, 2004) Backwash is now seen as a part of the impact a test may have on learners and teachers, on educational systems in general, and on society at large (Hughes, 2003). Since it can be harmful or beneficial, teachers should avoid the former and try to achieve the latter.

5. Practicality

Bachman & Palmer (1996) define practicality as the relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities. If the resources required for implementing the test exceed the resources (human/materials/time) available, the test will be impractical.Practicality, then, can be determined for a specific testing situation. A test may be practical in one situation but may be not in another.

In conclusion, a good test is supposed to have all the qualities discussed above. Bachman & Palmer (1996) argue that teachers need to find an appropriate balance among these qualities, which will vary from one testing situation to another. A good test in one situation, therefore, may not be a good test in another. Having only a few qualities very high without having the rest will not yield a good test, either.

How to improve an English test

The awareness of what constitutes a good test is a starting point for every teacher. The guidelines below are how to improve English tests.

1. How to make tests more reliable

Hughes (2004) recommends many ways to obtain both test/re-test and scorer reliability. Student-related reliability will not be discussed because it is not directly controlled by teachers.

1.1 Test/Re-test Reliability

- Take enough samples of behavior.

The test must have enough items for a teacher to tell whether the students know the materials. To increase the test reliability, however, additional items should represent a fresh start or be independent of each other and of existing items. In other word, the ability to answer the next question must not depend on the ability to answer the previous question. Otherwise, there is practically no additional question for the student, which means the teachers do not get an additional sample of the students' behavior, so the reliability is not increased. On the other hand, a test should not be made so long that the students become so bored or tired that the behavior they exhibit becomes unrepresentative of their ability.

- Exclude items which do not discriminate well between weaker and strong students.

Items on which strong students and weak students perform with similar degrees of success contribute little to the reliability of a test. Multiple-choice tests allow the calculation of the discrimination index (D), which ranges from 1 to 0 to -1. The higher the D, the better the item discriminates. Items with minus D should be excluded from the test.

- Do not allow the test-takers too much freedom.

It is more difficult to compare essays on different topics than those on the same topic with specific conditions such as audience, purpose and length. Requiring every test-taker to do the same well-defined tasks will help the teachers to mark them more reliably.

- Write unambiguous items.

An item that can be interpreted in different ways on different occasions means that the item is not contributing fully to the test reliability. Poor English may also cause ambiguity and is a bad model for the students. Having other teachers and native speakers scrutinize the draft will reduce this problem.

- Make sure there is only one correct answer for a multiple-choice test.

A multiple-choice item that can be answered in different ways on different occasions makes a test less reliable. The key to a great multiple-choice question, however, is a set of terrific distracters. They must be attractive but have less merit than the correct answer (Salkind, 2013).

- Ensure that tests are well laid out and perfectly legible.

Poor photocopying can dramatically reduce the quality of a well-written test.

A colorful graph or picture can simply be different shades of grey that cannot be understood by the students.

- Make candidates familiar with format and testing techniques.

When the formats or testing techniques are new, explain them in class before the test. Some students may not know about penalty for wrong guesses in certain multiple-choice tests. They would do better if they knew it.

- Provide uniform and non-distracting conditions of administration.

When students with the same listening ability take a listening test in a quiet room, they normally do better than taking the same test is a noisy place. Non-distracting conditions such as well-lit, cool and quiet are important for every kind of tests. Whenever there are more than one test rooms, the room conditions must be the same.

1.2 Scorer Reliability

- Use items that permits scoring which is as objective as possible.

Test techniques such as multiple-choice, matching, and true-false do not have a problem with scorer reliability as they do not require judgment from the scorers, making them popular because they are easy and fast to mark. However, they not only encourage guessing, but also do not require the test-takers to *produce* the language. An alternative is the open-ended item which has a unique, possibly one-word, correct response which the test-takers produce themselves. Having the students write the correct form of the given verbs will certainly better demonstrate their ability to conjugate those verbs than having them do a multiple-choice test.

- Provide a detailed scoring key.

A test cannot be a good test unless its correct and complete key is provided. Specify all acceptable answers and assign points for partially correct responses for short-answer items. Compile a banding system for a particular group of students for a writing or speaking task.

- Employ multiple, independent scoring

When possible, subjective tests should be scored by at least two independen ttrained scorers.

- Do everything possible to make the test reliable so that it can be valid.

If a test is not reliable, it cannot be valid. However, a reliable test may not be valid. For example, a multiple-choice “writing” test may be reliable, but it cannot be a valid test on composition writing.

2. How to make tests more valid

1.1 Construct Validity

- Define the construct or underlying trait(s) that you want to test before writing a test.

A "vocabulary" test could mean a test of word meanings to one teacher, but it may include the knowledge of word parts and the part of speech to another. The latter, however, may be a part of a "grammar" test for others. A shared view based on a theory of language learning can help increase the construct validity of the test.

- When feasible, use direct testing.

A test is said to be direct when it requires the candidate to perform precisely the skill which we wish to measure (Hughes, 2003). The so-called "speaking" test that requires the test-takers to choose or write the correct answers instead of saying them has lower construct validity than the one that require sactual speaking. However, scorer reliability may be a problem in many direct tests. Therefore, it is essential to devise a valid test first and then to establish ways of increasing its reliability (Heaton, 1988).

1.2 Content Validity

- Write explicit specifications for the tests.

All course objectives and the contents that need to be tested should be listed in the specification. Compare the actual test against the list to see how representative the test items are. For example, if a news writing course aims to teach how to write headline, lead and news body, and the test includes only the headline and the lead, the content validity of this test is affected.

1.3 Validity in Scoring

- Make sure that the scoring of answers relates directly to what is being tested.

If an interview is given to see how well the students can speak, their outfits or hair styles should not be judged because they have nothing to do with English.

3. How to make tests more authentic

Teachers need to think of what language use their students will be likely to encounter. Giving directions, for example, may be more common than making a conversation with a bank teller in English for many Thai students. Choosing the materials (reading passages, scripts for a listening test, etc) from the real-world sources will make the test more authentic because the language can be more natural and contextualized than the language found in many stems written for tests by non-native English teachers.

4. How to achieve beneficial backwash and avoid harmful backwash

Since there is a tendency to test what is easiest to test rather than what is most important to test (Hughes, 2003), teachers should remember to test the abilities whose development they want to encourage. Students will prepare for the final exam differently if they know they will be interviewed instead of taking a multiple-choice test. Preparation for the interview may help them improve their speaking abilities.

One way to avoid harmful backwash is not to overuse multiple-choice items. Although this test format has many advantages and seems to have no major disadvantages in certain areas such as vocabulary (Nation, 2001) or reading comprehension, objective tests can never test the ability to communicate in the target language (Heaton, 1988). If the teachers always give multiple-choice English tests, which only require the students to recognize the correct answers rather than producing the language, the they should not be surprised at all that the students cannot communicate in English.

Conclusion

In conclusion, teachers should do their best to make sure that their tests are both reliable and valid so that their decisions based on the test scores can be justified. At the same time, tests should not only be as similar to the real-world tasks as possible but also gives beneficial backwash. All the necessary test qualities should be balanced and maximized to ensure good, useful and effective tests.

References

Bachman, L.F. & Palmer, A.S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Brown, H.D. (2004) *Language assessment: Principles and classroom practices.* White Plains, NY: Pearson.

Carr, N.T. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.

Heaton, J.B. (1988). *Writing English language test.*(new ed.). Essex: Longman.

Hughes, A. (2003) *Testing for language teachers*. (2nd ed.). Cambridge: CambridgeUniversity Press.

Towards More Authenticity in Language Testing Paper to the AFMLTA National Conference 2003, Languages Babble, Babel and Beyond, Hilton Hotel, Brisbane, Retrieved 10-12 July, 2003 from http://www.islpr.org/PDF/Towards_More_Authenticity_in_Language_Testing.pdf

Mcnamara, T. (2000). Language Testing.Oxford: Oxford University Press.

Nation, I.S.P. (2001) *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Salkind, N.J. (2013) *Test & measurement for People who (think they) hate tests &measurement.*(2nd ed.).Thousand Oaks, CA: Sage Publications.