Reliability and Validity in Language Testing – A Real Conflict?

Cian O'Mahony, Asia University

Abstract

Language testers have been told that the qualities of reliability and validity are essentially in conflict, that it is, "not possible to design test tasks that are authentic and at the same time reliable" (Bachman & Palmer, 1996, p. 18). The is the view many educators hold, that reliability and validity are largely incompatible with each other, i.e. that increasing reliability of tests will necessarily reduce the validity, while increasing the validity will reduce the reliability. However, both Bachman and Palmer (1996) and Underhill (1987) refute this view as an unnecessarily uncompromising position. Although Bachman and Palmer (1996) note that a "tension" exists between reliability and validity, they argue the "need to recognize their complementarity" (p.18). They prefer to focus on the construct of "test usefulness" for judging tests, of which reliability and validity are a part, along with authenticity, interactiveness, impact and practicality: all six of these qualities need to be optimized in order to increase a test's usefulness (Bachman & Palmer, 1996). Although all of these qualities are clearly important, this author will focus only on threats to and conflicts between reliability and validity, except where the other qualities are directly related. This paper will argue that, although there are tensions between reliability and validity, increasing one does not necessarily decrease the other. In fact, since scholars claim that reliability is a necessary component of validity, it is clear that increasing reliability of a test can indeed increase its validity (Underhill, 1987; Bachman, 1990).

Introduction

The distinction between reliability and validity is essential in order to develop the discussion any further. Scholars state that reliability refers to how consistent a test's scores are in different (hypothetical or real) situations with the same test-takers, while validity is how well the test measures what it is supposed to measure (Bachman & Palmer, 1996; Henning, 1987). Thus, if students were to take a given test on a given day, if their scores were similar to the scores they would have attained on another day, the test would be seen to have high reliability. Conversely, if the scores were to differ wildly on different occasions, the reliability would be low. A test that aims to measure ability to write compositions would have high validity if the test involved actual writing of compositions, but lower validity if test-takers' writing skills were being inferred by their ability to choose the correct grammar form or vocabulary item in a multi-choice exercise. The score that a test-taker would hypothetically get if there was perfect reliability is known as the "true score" (Henning, 1987). Thus, the more reliable a test is, the closer the actual scores of the test-takers should be to the true score.

The validity of a test depends on the testing situation (Henning, 1987) – a test that is valid for one group (e.g. students on an academic programme) may not be valid for another (e.g. customs officials learning the target language specifically for their job). There are various types of validity. For example, if a test item looks right to other testers, teachers, moderators and testees, it can be described as having at least face validity. Thus, it is always useful to show a test to other colleagues and friends to check. It is very easy as the constructor of a test to be too subjective, miss "silly" mistakes or ambiguities (Heaton, 1988).

It is said that in order to be valid, a test also needs to be reliable, i.e. if the test lacks consistency in scores it lowers the validity (Bachman, 1990; Hughes, 1989). Underhill (1987) takes this a step further by including reliability as a "specific form of validity" (p. 105), although scholars list them as separate categories (Henning, 1987; Hughes, 1989). While reliability may be necessary for validity, it does not follow that a test requires validity in order to be reliable (Bachman, 1990; Hughes, 1989). For example, a multi-choice test on general knowledge may have high reliability, yet it will not necessarily be a good measure of a test-taker's ability to write about their general knowledge, and thus would not have a high validity if this was the purpose of the test. There are several types of validity: content validity, or how well the content of a test matches the content of the course (at least for tests that are taken following a course); criterion-related validity, which compares test scores with “some independent and highly dependable assessment of the candidate’s ability,” such as another test which has already been assessed as being highly valid (concurrent validity) or comparing a placement test with scores that students received in a test after the course of instruction (predictive validity); construct validity, which refers to the effectiveness of a test to measure a specific “construct” or hypothesised ability; and face validity, which refers to how appropriate on the surface the test appears to test-takers, teachers, employers and others (Hughes, 1989, p. 23).

Threats

Threats to reliability can come from four main areas: the learner (i.e. the test-taker), the scoring, the test administration and the test itself (Henning, 1987). The first three of these categories have little influence on the validity of the test, either positive or negative, except inasmuch as reliability is considered to be an essential component of validity (as claimed above). For example, sickness is an example of learner variation that can affect the true score (Henning, 1987), yet the effect on reliability bears little relation to the validity of the test since it is completely out of the test designers’ and administrators’ control. Other learner factors that may affect the test score include further learning (i.e. the difference in scores could be a result of extra study rather than the test itself), forgetting (the opposite of further learning), fatigue and emotional disturbance (Henning, 1987). Other relevant learner variables include response arbitrariness, or guessing, and test wiseness, which is an ability to answer test items correctly despite lack of ability or knowledge in the area being tested (Henning, 1987). Unlike other learner factors, these can impact strongly on validity, since the test-taker’s score may reflect an ability other than what is being measured. This is related to the concept of *response validity* – “the extent to which examinees responded in the manner expected by the test developers” – where validity (and reliability) can be adversely affected by factors such as lack of cooperation from the students (Henning, 1987, p. 96). Some problems discussed in this paragraph (e.g. sickness) have little to do with test designers and teachers, and thus little can be done to reduce their impact on reliability. Some solutions do exist, however, such as “correction-for-guessing procedures” and “strategy guidelines for all examinees” to reduce the effect of test-wiseness

(Henning, 1987, p. 96). Response validity can be improved by giving students incentive to respond appropriately, such as by assigning a portion of the final grade to a piece of assessment (Henning, 1987).

Further Sources of Unreliability

Inconsistent scoring can also contribute to unreliability, particularly where some kind of subjective judgement may be required (Hughes, 1989). Inconsistency in scoring can be caused by either intra-rater (where a single rater may assign a different grade to the same paper on different sittings) or inter-rater (where different raters assign different grades to the same paper) error (Henning, 1987). Each of these types of inconsistency can be influenced by such factors as quality of handwriting, attitude towards the test-taker (if the name is known) or the experience and skill of the rater(s) (Henning, 1987). The effect of the scorer's attitude towards the test-taker can be nullified by using numbers rather than names on test papers (Hughes, 1989). One obvious solution to most scoring problems is to reduce subjectivity, such as by using multi-choice format or using test items with clear right-or-wrong answers. However, in increasing reliability the validity may be reduced: if, for example, with a test aimed to gauge learners' conversational ability, a multi-choice paper test may assess something other than conversational ability. Nevertheless, Hughes (1987) notes that it is not necessary to sacrifice validity in such cases, since scoring can be made more reliable by using detailed scoring keys and by providing adequate training for scorers, including comparing ratings with other scorers to check for consistency.

Fluctuations in test administration are also a possible source of unreliability: the clarity of verbal instructions, the (in)effectiveness of a particular administrator in preventing cheating, timing (whether the test is morning or evening, during a long break, close to other tests, etc.), and environmental problems (such as noise interference or uncomfortable chairs) are examples of administrational factors that can affect students' scores (Henning, 1987). Care should be made to minimise these problems by ensuring the conditions are as similar as possible for all test-takers, and guidelines can be given so that there is consistency between different administrators in different exam rooms (Henning, 1987). Henning (1987) specifically mentions that "any irregularity in the testing situation" will reduce validity as well as reliability (p.93), although this reduction in validity appears no different from reduction caused by any lack of reliability, as mentioned above.

The final source of unreliability is the test itself. Unlike most of the factors described in the previous three paragraphs, many of the following issues also relate directly to the validity of a test. Tests that are too easy or too difficult may result in bunched scores, and as Henning (1987) points out, greater variance in scores generally gives greater reliability – the person separability (how well the test distinguishes different students) is reduced when the test scores are close together, and even a slight change in score could greatly alter a student's ranking. Items with low discriminability, i.e. those which do not separate high achievers from low achievers, affect person separability in a similar way (Henning, 1987). Too few test items may also compromise reliability, since even good students make mistakes and poor students may get some correct answers through luck (Henning, 1987). Ambiguous or unclear instructions can cause students who know the correct answer to lose marks, thereby reducing reliability (Hughes, 1989). Henning (1987) claims that increasing person separability, such as increasing the number of test items and ensuring that individual items can distinguish between stronger and weaker students, improves reliability. Adding an element of speed to a test, i.e. forcing students to complete the test within a certain time-frame, can also contribute to person separability. Although in this instance, validity could be adversely affected if, for example, the test is aimed to gauge students' knowledge rather than their ability to perform tasks quickly (Henning, 1987). In order to measure reliability, it is useful for "items of similar format and content" to be used, yet this too could reduce validity if a number of different skills need to be tested (Henning, 1987). Hughes (1989) argues against a test format that allows students too much freedom, such as choosing from a selection of topics to write a composition, since this makes reliable comparisons between students more difficult. He does, however acknowledge that such restrictiveness could potentially damage validity, as this may "distort too much the task that we really want to see them perform" Hughes (1989, p. 38). Conversely, measures to increase reliability such as reducing ambiguity and making instructions clearer cannot in any way reduce validity (unless the aim of the test is to find out how well students can interpret confusing instructions).

Validity Threats

There are various additional threats to validity. Henning (1987) labels "invalid application of tests" as the most obvious (p. 91): tests which may well be valid for one purpose or one particular group of students but may not be valid for another. For example, a test with proven validity for testing grammar structures learnt in a certain course may not be valid for testing a student's ability to write business letters. It is important to ensure that the test is relevant to its purpose and to the particular students who take it (Henning, 1987).

Another common threat to validity is inappropriate content selection, "when items do not match the objectives or the content of instruction" or where the test covers these objectives or content insufficiently (Henning, 1987, p. 91). This also presents problems for face validity, since students are often unhappy when they feel the test has covered areas they have not studied in class (Henning, 1987). As both a student and a teacher I have witnessed many such instances, which often lead to complaints about an exam. Clearly care should be taken to make the test items relevant to the course of instruction. Henning (1987) suggests using specification schemes to prevent inappropriate content from being included while ensuring that the important points are covered.

Neither of the previous examples of threats to validity are necessarily affected by reliability. A test with inappropriate content or a test that is applied invalidly could be either reliable or unreliable, thereby not implying any tension between reliability and validity. One area where tension does exist between reliability and validity involves authenticity. Although Bachman and Palmer (1996) list authenticity as a separate category from validity, it is clear from the quote (see Abstract) that there is a close link between them.

Finally, criterion-related validity can be threatened if the criterion a test is being correlated against has low validity or reliability, while construct validity can be adversely affected if the hypothesised constructs are not valid (Henning, 1987). Clearly, then, it is important to select an appropriate criterion or construct when assessing validity.

Conclusion

In conclusion, I would have to disagree, in part, with the initial statement (see Abstract) as this paper has discussed, there are occasions where there is no conflict between reliability and validity and that reliability is seen as a necessary component of validity. Increased reliability can improve validity or threaten it. To have more of one does not necessarily lead to less of the other. As has been discussed, there are occasions when a conflict does become apparent especially when trying to make authentic tests that require more subjective marking. Although it is impossible to achieve perfect validity and reliability, it is possible when conflict occurs to achieve a balance between the two by careful planning and consideration, and by implementing detailed marking criteria for the more subjective items.

References

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

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

Brown, H.D. (2004). Principles of Language Assessment. New York: Pearson Education.

Heaton, J.B. (1988). *Writing English language tests*. London: Longman.

Henning, G. (1987). *A guide to language testing: development, evaluation, research.* New York: Newbury House.

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

Mangubhai, F. (2004). *Language testing*. Toowoomba: University of Southern Queensland.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques.* Cambridge: Cambridge University Press.

Weir, C. (1990). *Communicative language testing.* Hemel Hempstead: Prentice Hall International.