Appreciating the efficacy of formative and summative assessment in tertiary stage examination regimes

Sebastian SAINOO-FULLER

要　約

本稿は、形成的評価方法に基づく、学内・学外でのテストを比較することで、第３段階の教育機関における形式的評価の特徴を考察するものである。ある文脈における特定のテストの有効性を評価するために、その妥当性と信頼性について分析を試みた。本稿はテストの作成におけるその目的と背景の役割、被験者の成績の解釈の役割に焦点を置いたものである。

Abstract

This paper examines the characteristics of formalized testing in the context of tertiary stage education, contrasting external and internal testing in terms of formative and summative assessment. Concepts of validation and reliability are deployed to evaluate the efficacy of a specific test in a specific context. The paper concludes by emphasizing the roles of purpose and context in test design and interpretation of the examinees' results.

Keywords

Summative, formative, external and internal test design, tertiary education

1. Introduction[1]

Language tests have been often criticized for having rather a detrimental effect on teaching and learning, since they fail to assess a learners' performance accurately, whether for teaching or employment purposes.

In discussing test efficacy or usefulness, Bachman and Palmer (1996) envisage a model with different qualities for the design and development of language testing. However, their proposal disagrees with the traditional approach to describing test properties, which emphasize the need to maximize every dimension to build a universal test, leading some language testers to the extreme reactionary position that maximizing one quality contributes to the virtual loss of others. In fact, "language testers have been told that the qualities of reliability and validity are essentially in conflict, or that it is not possible to design test tasks that are authentic and at the same time reliable" (Bachman & Palmer, 1996, p.18).

These issues are further magnified when considered through the various prisms of summative and formative tests conducted externally or internally for the purposes of teaching or employment in tertiary level education. Mid-term test and final tests held in the semester, and nationally/ internationally recognized certificate examinations respectively all fall under this umbrella.

2. Discussion

In examining whether reliability and validity are essentially in conflict, or whether they are correlated for assessing the overall efficacy of a given test, it is essential to define 'reliability' and 'validity.'

Firstly, reliability often refers to consistency of measurement, which can be regarded as a function of the scores consistency from one set of tests and test tasks to another. Hence, a reliable test score will be consistent across different features of testing situations (Bachman & Palmer 1996, pp. 19-20). For example, if the same test were administered to the same group of students on two different occasions in two different settings, the more similar the scores would be, the more reliable the test could be said to be. Conversely, if two forms of a test intended to be used interchangeably were administered to a particular examinee and the test scores were not very consistent, the scores would be regarded as unreliable indicators of the skill in focus. Considering test scores with relatively less consistency fail to indicate the examinee's control of the specified skill, reliability is clearly an essential quality of test scores. At the same time, one must bear in mind that eliminating inconsistencies altogether is impossible without always delivering the same test (which would then be open obviously to cheating). Therefore, it is crucial to minimize the effects of those potential inconsistencies that are under control, in designing test tasks and developing language tests (Bachman & Palmer, 1996). Formalized tests such as TOEIC and TOEFL go to extreme lengths to maintain public credibility by maintaining a structure and content which is basically the same between tests without being so predictable that cheats can memorize rote answers to skill highly without understanding the skills being tested. This however limits the range of skills and the flexibility of the tests.

Secondly, Gronlund defines validity as the "degree to which inferences made from assessment results are appropriate, meaningful and useful in terms of the purpose of the assessment" (cited in Brown, H.D., 2004, p. 22). For example, a test for measuring a reading ability only includes items related to the knowledge or control of grammar and vocabulary, the test is said to be invalid, because the test does not measure the whole but a certain scope of the examinees' reading ability. Summative tests, such as mid-term tests tend to limit themselves very few skills due to time and taught content constraints, and are often left open to this criticism.

Considering the relationship between reliability and validity, Underhill (1987, pp. 105-108) goes further arguing that "validity" has many different dimensions based on the interpretation of the suitability of the test. Underhill underlines different kinds of validity, including personal judgments of validity, face validity, statistical measures of validity, predictive validity, concurrent validity, construct validity, and content validity. He additionally discusses that reliability can also be viewed as a specific type of general validity, stating that a test without any reliability cannot be generally valid (Underhill, 1987, p. 105). In these cases testing is conducted for appearances, for testing's sake. The ethics of such a test are at best highly questionable, drawing not just the practitioner's integrity into doubt but drawing the whole framework of language testing under fire for having a detrimental effect on teaching and learning, since they fail to assess a learners' performance accurately, whether for teaching or employment purposes. At the summative level, educators cannot use the results to enable the examinees to grow. From the formative point of view, employers and examinees are left in doubt to what skills the examinee possesses, what further training may be required, and in general prevents the examinee from demonstrating their proven potential at interview.

On the other hand, Underhill's paradox is that (cited in Bachman & Palmer, 1996, p.18) the qualities of reliability and validity are essentially in conflict. Underhill (1987, p. 105) demonstrates that reliability is usually regarded as an entirely different idea from validity and the two terms are often presented as mutually exclusive, namely that highly reliable tests are less valid, and vice versa. Hughes (2003, p. 50) also argues that in order to be valid, a test is to be measured with consistent accuracy, yet a reliable test may not be valid at all, citing the following anecdotal composition examination. Candidates were made to write down the exact equivalents of 500 pieces of discrete vocabulary listed in their mother language, which may well have been be a reliable vocabulary test; still it was unlikely to be a valid test of composition. It is true that restricting the scope of what candidates are permitted to speak or write in a test often helps to increase reliability, since the variety of skills being measured in the test (a factor which causes scores to be less consistent) is reduced. However, it also tends to diminish the validity of the test, as each skill of each candidate is restricted by the narrower scope. This problem is particularly seen in measuring the productive skills namely speaking and writing. The TOFL written paper encourages the examinees to produce coherent academic texts, but for reliability and practical issues of marking, the examinee is denied a choice of question and the time limit forces the writer to be concise. This favors a specific style of academic writing, privileging bold, argumentative compositions over more speculative discursive styles. This point has been a potential source of criticism since this style favors certain cultures and genders that are more familiar with dialectic debate, to those peoples who are more consensual in their approach to academic composition. Furthermore the recent introduction of computer aided testing has led to even more uniformity, making the marking less subjective, but also more artificial.

Bachman and Palmer (1996)、on the other hand, claim that reliability and validity are to be considered in tandem rather than as independent characteristics, considering the significance of finding an appropriate balance among the six qualities discussed above which includes reliability and validity. They additionally state that both reliability and validity are critical for tests and are sometimes viewed as essential measurement qualities, which provide the prime justification for using test scores as a basis for making inferences or decisions. It should be noted that TOIEC is used in this way by human resource departments, TOEFL by university admissions departments, and mid-term/ finals by individual educators when delivering credits for graduation. Nevertheless the blanket uncritical acceptance of a testing regime's reputation is most usually the basis of the final decisions.

Also, Underhill's former statement (1987) insists that the ambiguity behind the concepts of reliability and validity invokes a lack of clarity in definition as well as incorrect assessment and calculation. Despite this ambiguity, he still suggests that each type of validity is important and that as much information as possible about the different types of validity should be gathered and analyzed carefully and critically according to the objectives of the test tasks, so that a balanced answer to the question as to the efficacy of a test can be gleaned.

Therefore, in the view that an appropriate balance amongst the several qualities, including both reliability and validity, for development of the test tasks, are essential to the efficacy of any language test, Underhill as well as Bachman and Palmer rather seem in agreement, though Underhill claims that the concepts of reliability and validity are rather vague and difficult to be assessed and calculated.

One more key principle to underline is the degree to which it is not possible to design test tasks that are authentic and at the same time reliable (Bachman and Palmer, 1996, p. 18). According to Underhill's definition of authenticity (1987), "an authentic task is one which resembles very closely something we actually do in everyday life", while reliability is defined as "a measure of the degree to which a test gives consistent results" (Richards, J.C. & Schmidt, R., 2002). Based upon this, the dilemma remains whether or not it is possible to design authentic test tasks with reliability. Hughes (2003, pp. 163-164) describes the two cases of testing listening, in which utilizing authentic materials may lower reliability. Firstly, using authentic speech drawn from the radio, television or a recording of native speakers' discourse to test listening skills may affect reliability, if the quality of recording causes added difficulties to examinees. Secondly, if native speakers' failure in speech necessitates re-recording, the performance of the individuals may be affected by recording faults in different degrees from one occasion to another. However, as for the cases mentioned above, Hughes suggests that reliability of tests utilizing authentic materials will not deteriorate, if great care is taken by test developers in the recording process. TOEIC and TOEFL both use professional actors to re-create authentic scenes, though the speed and intonation is often altered to make the dialogues more understandable. One factor damaging authenticity is the privileging of one specific regional dialect over all others, albeit making the test more understandable but diminishing authenticity in the global corporate workplace.

Another concern about the relationship between authenticity and reliability is that a relatively objective test, such as multiple-choice testing, is preferred to a subjective one if reliability is over-emphasized, because the former, presenting one single correct answer with no personal judgment to be exercised, makes test scores more consistent. One the other hand, an objective test, which does not include any language production or performance, such as speaking or writing, is far from authentic, indicated by the degree to which language materials have the qualities of natural speech or writing (Richards and Schmidt, 2002). Bachman and Palmer (1996, p. 18), however, seem to suggest some solutions to this problem: "both authenticity and reliability are critical qualities of test scores, yet, in the development of testing, it is the degree to which they complement each other, rather than the tension between the different qualities that is to be recognized".

Moving on though, validity and reliability as defined above are often under threat from the marking regime deployed after the test has been delivered. Reliability, whose performance is measured by consistency, is often to be threatened by inconsistent rating seen in subjective composition tests. Bachman and Palmer (1996, p.20) argue that given that a vast number of compositions are assigned to different markers, some markers might mark more strictly than others, damaging the consistency of the overall test. Nevertheless, Hughes (2003, pp.94-95) views this reliability problem as surmountable, by making several suggestions for facilitating reliable scores of testing writing, such as setting as many tasks as possible, restricting candidates, offering no choice of tasks, ensuring long enough samples, creating appropriate scales for scoring and using holistic scoring, in which a single score is assigned to writing or speaking tasks based on an overall impression of examinees' performance (Richards and Schmidt, 2002). In terms of practicality, his proposal on testing of writing skills seems to be of great help to the solution of this problem. This is seen specifically in TOEFL writing, and to some degree automates the task of marking, but limits the range of expression available to the examinee as mentioned above.

In addition to the issue of rating and evaluation, the following also are considered to be threats to reliability: time allotted for the tests administered, the conditions of the facilities where tests are conducted, health of the examinee, psychological pressure on the examinee, and so forth, affecting the consistency of test scores. Hence, these factors must be carefully examined in the design and development of language testing, so that reliability may not be diminished.

Saying this, reliability can interfere with validity. It is undeniable that the primary purpose of language testing is to provide objective measures that can be used as an indicator of an individual's language ability against a standard. Therefore, reliability and validity are the two essential measurements crucial to the efficacy of any language test. Nevertheless, too much emphasis on reliability, which can refer to consistency of measurement in designing and developing tests, often leads into diminishing validity of test tasks. Eventually, there is the danger that the consistency of test scores becomes the overarching concern when designing test tasks. For example, a multiple-choice type of grammar test, which is expected to provide more highly reliability, is not sufficient for assessing examinees' composition ability, because ones` grammatical knowledge is only one scope of linguistic knowledge. Hence, this test would be regarded as highly reliable, but as invalid. Given the case indicated, finding an appropriate balance of these qualities in accordance with the aims of the test, rather than stressing one quality or the other is to be considered by test developers for the usefulness of testing (Bachman and Palmer, 1996, p. 18).

Also, when discussing the threats to validity, one must consider the scoring of test tasks designed for measuring language ability, since they also reflect validity, one of the essential basis for total quality control for developing test tasks. Hughes (2003, pp. 32-33) provides a reading test as an example to illustrate the problem of scoring. If the scoring of a reading test, which may require short written responses, takes spelling and grammar into account, it is not valid, because the test is meant to measure reading ability, not spelling or knowledge of grammar. This case demonstrates that measuring more than one skill makes the skill in question less valid, for example, when spelling mistakes or inaccurate grammar interferes with the accurate assessment of reading ability. Thus, if a test, the primary purpose of which is measuring speaking or writing ability, does not concord with the purpose of the test due to the scoring of the speech or writing, the validity of the test is diminished.

Furthermore, Underhill's argument about the ambiguity of validity (1987) mentioned above indicates the issue of interpretation, which is another threat to validity. He argues that the different kinds of answers to the question of validity on the basis of the different interpretations are often in conflict, thus, striking a balance between these conflicting answers is a key to decision making for the improvement of tests as well as evaluation. However, according to Underhill, the problem is that there has been tendency among language testing experts to rely on statistics measures for test evaluation excessively, which makes evaluation quite scientific and objective, but also endangers one's intuitive judgment. He criticizes this stance stating that the view that subjective test is always bad, while an objective one is necessarily good, is likely to be correct, if statistical validity is a major concern. Yet the reverse is true, if communicative validity is prioritized. Especially in an oral test, which is a personal encounter between two human beings, designed by humans, administered by humans, taken by humans and marked by humans, (Underhill, 1987, p. 105), if the statistical and objective validity is favored, the result may be a strong bias towards mechanical tests and against the human aspect of oral tests. Therefore, Underhill suggests that the best way to interpret the ambiguous validity is to find a well-balanced design of test program in conformity with the aims of the test tasks. He also stresses that validation, which refers to the process of establishing the general validity in a test regime, is not an absolute but a relative process, thus the degree of validity of a test relates only to the particular social context in which it was established (Underhill, 1987, p. 104). Consequently, validity in one test program cannot be expected to have the same validity when applied to a totally different context, which is to be kept in mind in drawing up test regimes.

3. Conclusion

At the tertiary level, the main concern should be the context of the test within the course and curriculum. This defines the efficacy of the test, since other factors such as reliability, validity and authenticity become apparent.

Learner centered tests will provide the best summative regimes enabling students to visualize, monitor and attain their learning goals. Mid- term tests conducted internally fall into this category. More automated, reliable tests based on an external repeatable framework however are proposed for the formative tests used in decision making, such as graduation, and employment.

This dichotomy leaves a dilemma for the educator. Firstly the educator must make it explicit to the examinee during the enacted curriculum what the content and objective of the test is. Namely the context must be explained, otherwise the examinee suffers test fatigue, and loses interest in the learning and testing process to the detriment of their performance in the language.

Secondly, the educator and the whole academic community has a responsibility not only to design batteries of authentic tests but also to make the public, examinees, and bureaucrats, educators and employers aware of the context and limits of each testing regime so that they can make educated choices based upon more than the general reputation of the test.

Notes

[1]TOEIC and TOEFL are the registered trademark of Educational Testing Services.
For the purposes of this thesis familiarity with both tests is assumed.
More information can be found at: http://www.ets.org (last accessed 2007/8/9)

References

Bachman, L. & Palmer, A. (1996). LIN 8007 Language testing. University of Southern Queensland.

Brown, H.D. (2004). LIN 8007 Language testing. University of Southern Queensland.

Hughes, A. (2003). Testing for language teachers. Cambridge: Cambridge University Press.

Richards, J.C. & Schmidt, R. (2002). Longman dictionary of language teaching & applied linguistics. (3rd ed.). Harlow, Essex: Pearson Education Limited.

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