ISSN: 2278-4012, Volume:7, Issue:3, July 2018

A Study of Validity in Assessment for Learning

WU Zongyou

Shanghai International Studies University, Shanghai, China

Anhui Normal University, Wuhu, China

Abstract: *Assessment for learning is any assessment for which the first priority in its design and practice is to serve the purpose of promoting students' learning. This definition implies that validity is central to developing or practicing assessment for learning. By definition, the purpose of assessment for learning is to lead to further learning. It thus differs from assessment designed primarily to serve the purposes of accountability, ranking, or certifying competence. Firstly, the evolution of validity and validation in language assessment has been reviewed from three perspectives: categorized, unitary, and argumentative. Then, the definition of assessment for learning is briefly introduced and the validity of assessment for learning is discussed. Secondly, some key factors influencing the validity of assessment for learning are explored. Finally, two important validation frameworks of assessment for learning are introduced and analyzed.*

Key Words: *validity, assessment for learning, formative assessment*

Introduction

Assessment for learning is any assessment for which the first priority in its design and practice is to serve the purpose of promoting students' learning (Black & William, 1998). By definition, the purpose of assessment for learning is to lead to further learning. It thus differs from assessment designed primarily to serve the purposes of accountability, ranking, or certifying competence. This definition implies that validity is central to developing or practicing assessment for learning.

Development of Test Validity

Test Validity. In Lado's (1961) Language Testing, validity is defined as "essentially a matter of relevance. Is the test relevant to what it claims to measure?" A test is said to be valid if it measures accurately what it is intended to measure (Hughes, 1989). For many test users, validity is seen as an essential quality of a language test because to them "a valid test" means "a good test" (Fulcher & Davidson, 2012). Validity is an ominous word. The Oxford English Dictionary assigns it several meanings, deriving from Latin origins of "powerful, effective," "possessing legal authority or force," "technically perfect or efficacious," and "sound and to the point, against which no objections can fairly be brought." These senses are all relevant to current uses of the term validity in language testing (Cumming, 1995).

Over the past 60 years, the evolution of validity and validation in language assessment has advanced mainly along the following three perspectives: categorized, unitary, and argumentative (Han & Luo, 2013).

Validity and validation from categorized perspective. We create language tests in order to measure such essentially theoretical constructs as "reading ability," "fluency in speaking," "control of grammar," and so on. For this reason, the term "construct validity" has been increasingly used to refer to the general, overarching notion validity (see Figure 1). However, it is not enough to assert that a test had construct validity and empirical evidence is needed. Such evidence may take several forms, including the subordinate forms of validity, "content validity" and "criterion-related validity." (Lado, 1961; Oller, 1979)

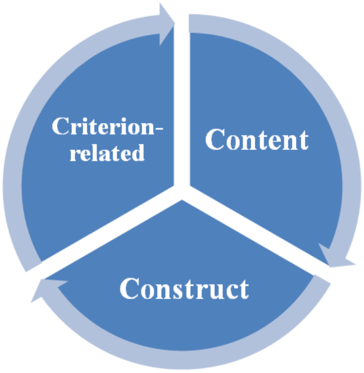

Figure 1 The earlier trio of validities (60s-70s).

In the 1980s, Henning (1987), Huges (1989), and Canale (1987) further expanded the earlier trio of validities. Henning (1987) made a distinction between empirical and non-empirical kinds of validity. Non-empirical validity does not require the collection of data or the use of formulae. Examples of this kind of validity include "face or content validity" and "response validity." Empirical kinds of validity usually involve recourse to mathematical formulae for the computation of validity coefficients. Common kinds of empirical validity include "concurrent and predictive validity," which are also termed criterion-related validities. Huges (1989) defined backwash or washback as the effect that tests have on learning and teaching and suggested testing the abilities whose development you want to encourage.

Validity and validation from unitary perspective. Messick (1989) defined validity as "an overall evaluative judgment of the degree to which evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores." Four characteristics are useful for summarizing Messick's conception of validity:

1. Validity is not a property of tests themselves. Instead, it is the interpretations and uses of tests that can be shown to be more or less valid;

2. Validity is best thought of as one unitary conception, with construct validity as central, rather than as multiple validities, such as “content validity,” “criterion-related validity,” or “face validity;”

3. Validity encompasses the relevance and utility, value implications, and social consequences of testing. This scope for validity contrasts with the view that validity refers only to technical considerations;

4. The complex view of validity means that validation as an ongoing process of inquiry. The focus on the process of investigation contrasts with a product-oriented perspective of a validated test—one for which the research has been completed.

Validity and validation from argumentative perspective. Bachman (2005) had made an model of “assessment use argument (AUA)” which based on Toulmin’s (1958) argument structure model, the inferential links between the test scores and its interpretation about language ability of test taker are given evidence by the logical support from the model, which concerns on the scores and test use.

In the AUA model, the valid argument includes four important parts which are warrant, backing, data, and claim. Warrants are used to give evidence for the intended interpretations, meanwhile, the backing support the warrants. In the argument structure, the alternative explanation might support or weakened by rebuttal data. The elements and the relations among the elements are shown in the following Figure 2 (Bachman, 2005).

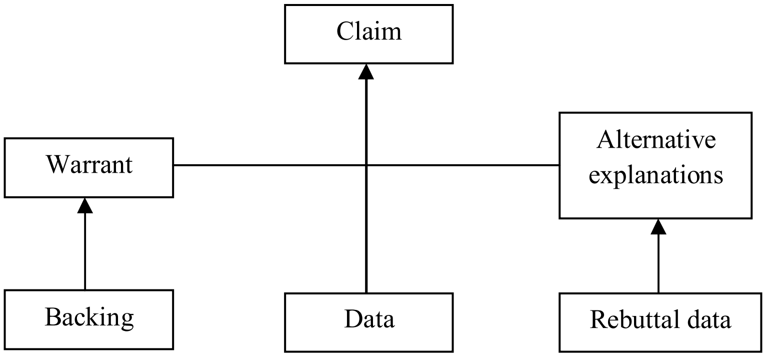

Figure 2 Diagram of the structure of arguments.

Claim: It means the intended interpretation, on the basis of the data, about what a test taker can do and knows;

Data: the responses of a test taker to assess tasks;

Inference: The link between the data and the claim, which is supported by the warrant;

Warrant: A proposition that we use to justify the inferences from data to claim;

Backing: Evidence that the warrants are legitimate. In language testing, theory, and previous findings generally consist of the backing for warrants;

Alternative explanations: A competing proposition that might provide a reasonable explanation for the data. In Messick’s (1989) term, this could be referred as “construct irrelevant variance” and “construct under representation;”

Rebuttal data: Evidence, which may either weaken or support the alternative explanation.

Assessment for Learning

Traditionally, assessment and instruction are "conceived as curiously separate in both time and purpose." Since Black and William's (1998) ground-breaking research that focuses on the substantive connection between assessment and learning, the notion of "assessment for learning" has been advocated as having a strong potential in improving learning. Assessment for learning is closely related with learning-oriented formative practices and it is defined as "any assessment for which the first priority in its design is to serve the purpose of promoting students' learning" (Black & William, 2009). Therefore, in this paper, "assessment for learning" and "formative assessment" is used interchangeably. The well-known and widely disseminated Assessment Reform Group (ARG) leaflet published in 2002 gave the following definition, "Assessment for learning is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there."

Validity of Assessment for Learning

As stated earlier, "assessment for learning" has been advocated as having a strong potential in improving learning, so if assessment for learning is valid, it must lead to further learning. The validity argument is therefore about the consequences of assessment. The assumption is that assessment for learning generates information that enables this further learning to take place—the "how to get there" of the definition of assessment for learning. One implication of this is that assessments may be formative in intention but are not so in practice because they do not generate further learning.

This "consequential" approach differs from how the validity of summative assessments is generally judged. Here, the emphasis is on the trustworthiness of the inferences drawn from the results. It is about the meaning attached to an assessment and will vary according to purpose.

Validity is no longer simply seen as a static property of an assessment, which is something a test has, but is based on the inferences drawn from the results of an assessment. This means that each time a test is given, the interpretation of the results is part of a "validity argument" (Stobart, 2006).

Key Factors Influencing the Validity of Assessment for Learning

Validity in assessment for learning hinges on how effectively this learning takes place. What gets in the way of this further learning can be treated as a threat to the validity of formative assessment. Crooks (2001) listed four types of factors influencing the validity of assessment for learning: affective factors, task factors, structural factors, and process factors (see Table 1).

Table 1 Key Factors Influencing the Validity of Assessment for Learning

Affective factors	Motivation	Teacher is devoted to helping student learn.

		Student cares about learning and wants to improve.
	Trust	Teacher is encouraging, constructive, sensitive to student's feelings. Class/peer relationships and attitudes support student's learning. Student feels safe to admit difficulties and uncertainties.
Task factors	Knowledge	Teacher understands the key aspects and difficulties of the task.
	Criteria	Teacher identifies and explains well the qualities sought. Student understands clearly what is needed.
	Standards	Teacher sets standards appropriate to student. Through descriptions and examples, the standards are explained. Student understands the standards and accepts them as appropriate.
Structural factors	Connections	Final version of task can benefit from the formative assessment. Work on subsequent tasks can benefit from the formative assessment.
	Purposes	Formative use of task is not undermined by parallel summative use.
Process factors	Self-assessment	Teacher helps student to develop self-assessment skills. Student takes increasing responsibility for his/her own learning.
	Peer involvement	Teacher encourages collaboration among students to improve work. Peers learn to be constructive and generous in offering feedback.
	Monitoring	Teacher monitors student's work to track both process and progress.
	Insight	Teacher detects misunderstandings or other obstacles to success. Teacher detects exciting possibilities in student's work.
	Timing	Feedback is given at times when student is most receptive to it.
	Balance	Feedback gives attention to strengths as well as weaknesses.
	Selectivity	Feedback addresses mainly the aspects likely to have biggest benefit.
	Wisdom	Feedback is convincing, appreciated, and useful to student.

Stobart (2006) summarized the key factors that may support or undermine formative assessment into two categories: the learning context and feedback. The learning context in which formative assessment takes place is seen as critical. This includes what goes on outside the classroom, the social and political environment, as well as expectations about what and how teachers teach and learners learn within the classroom. At a more individual level, feedback has a key role in formative assessment. What we know about successful feedback is discussed, along with why some feedback practices may undermine learning.

Validation Frameworks for Assessment for Learning

In assessment for learning, validity arguments go beyond the focus on the inferences drawn from the results to consider the consequences of an assessment—a contested approach in relation to summative assessment (Stobart, 2006).

After investigating into classroom assessment from three aspects: the classroom assessment environment, the integration of assessment and instruction, and the pervasive formative purpose of classroom assessment. Brookhart (2003) proposed a measurement theory for classroom assessment, called the "classroometric" theory, which is different from classic psycho-metric theory in three aspects: (1) the relationship between the assessor and the assessed; (2) construct-relevant and construct-irrelevant variance; and (3) reliability and errors (Li & Kong, 2015) (see

Table 2).

Table 2 Contrasting Large-Scale and Classroom Assessment Concepts

Concepts in Large-Scale Assessment	Concepts in Classroom Assessment
Validity: The measure is external to the inferences made and actions taken. Students are "subjects" upon whom observations are made. The validity goal is a meaningful inference about student performance and/or effective use of that information for a specified purpose.	Validity: Inferences made and actions taken are internal to the measurement process. Students are observers jointly with teachers; "those measured" make the inferences and take the actions in the formative assessment process. Students' awareness of and benefit from assessment information are part of the "information" itself. The validity goal is an understanding of how students' work compares to "ideal" work (as defined in the learning objectives) and/or effective use of that information for further learning.
Validity: The measurement context is construct-irrelevant. Content specifications describe a domain. Administration can be standardized. Scores can be equated or linked across contexts and forms of assessment.	Validity: The measurement context is construct-relevant. Assessment is part of instruction. A good assessment is an "episode of genuine learning." Content specifications refJect both the domain (learning objectives) and instruction (modes, activities). Teacher beliefs, teacher instructional practices, and teacher understanding of both the subject matter and students (including cultural and linguistic differences) are relevant validity concerns.
Reliability is consistency over irrelevant factors. Occasions, time, items and/or tasks are facets of error variance. The reliability goat is stable ranking of students on a score scale (NRT) or stable categorization of students along an achievement continuum (CRT).	Reliability is sufficiency of information. The reliability goal is stable information about the gap between students' work and "ideal" work (as defined in students' and teachers' learning objectives).

The primary use of any language assessment is to collect information for making decisions. Furthermore, the use of an assessment and the decisions made will have consequences for stakeholders, the individuals and programs in the educational and societal setting in which language assessment takes place (Bachman & Palmer, 2010). In assessment for learning, the intended uses of language assessments are promoting students' further learning. One way to think of the use of a language assessment is as a series of inferences from the test taker's performance to the consequences. These links are illustrated in Figure 3.

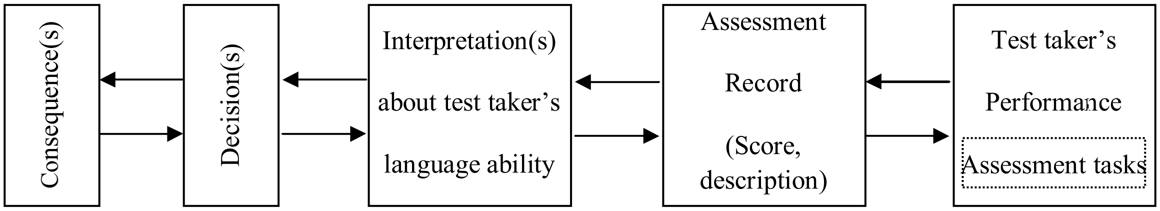
Consequence(s)
Decision(s)
Interpretation(s) about test taker's language ability
Assessment Record (Score, description)
Test taker's Performance
Assessment tasks

Figure 3 Links from test taker's performance to intended uses (decisions, consequences).

Conclusions

Firstly, the evolution of validity and validation in language assessment has been reviewed from three perspectives: categorized, unitary and argumentative. Then, the definition of assessment for learning is briefly introduced and the validity of assessment for learning is discussed. Secondly, some key factors influencing the validity of assessment for learning are explored. Finally, two important validation frameworks for assessment for learning are introduced and analyzed.

Acknowledgement

This work was supported by Project Cultivation Foundation of Anhui Normal University (No. 2017XJJ54); Tutor's Academic Guidance Program of Shanghai International Studies University (No. 2017010) and Teaching Research Project of Higher Education of Anhui Province (2017jyxm0089).

References

Assessment Reform Group (ARG). (1999). *Assessment for learning: Beyond the black box*. Cambridge: University of Cambridge.

Assessment Reform Group (ARG). (2002). *Assessment for learning: 10 principles*. Cambridge: University of Cambridge.

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

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Bachman, L. (2005). Building and supporting a case for test use . *Language Assessment Quarterly,* 2(1), 1-34.

Black, P. & Wiliam D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31.

Brian, L., & Peter, S. (2012). Portfolios, power, and ethics. *Tesol Quarterly, 39*(2), 263-297.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement Issues & Practice, 22*(4), 5-12.

Canale, M. (1987). The measurement of communicative competence. *Annual Review of Applied Linguistics, 8*(8), 67-84.

Crooks, T. (2001). The validity of formative assessments. In *British Educational Research Association Annual Conference* (pp. 1-9). Leeds: University of Leeds. Retrieved from http://www.leeds.ac.uk/educol/documents/00001862.htm

Cumming, A. H. (1995). *Validation in language testing*. Bristol: Channel View Publications Ltd.

Fulcher, G., & Davidson, F. (2012). *The routledge handbook of language testing*. London and New York, NY: Routledge.

Han B. & Luo K. (2013). The evolution of validity and validation in language assessment. *Foreign Language Teaching and Research*, 45(03), 411-425.

Henning, G. (1987). *A guide to language testing: development, evaluation and research*. Rowley, Massachusetts: Newbury House.

Hughes, A. (1989). *Testing for language teachers*. Cambridge and New York, NY: Cambridge University Press.

Lado, R. (1961). *Language testing*. London: Longman.

Li, Q. H. & Kong, W. (2005). A validation framework for formative assessment in EFL teaching. *Foreign Language Learning Theory and Practice*, 37(01), 24-31.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment. *Expectations and validation criteria. Educational Researcher, 20*(8), 15-21.

Messick, S. (1989). Validity . In R. L. Linn (Ed.), *Educational Measurement*. New York, NY: Macmillan.

Oller, J. (1979). *Language tests at school*. London: Longman.

Shepard, L. A. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement Issues & Practice, 28*(3), 32–37.

Stobart, G. (2006). The validity of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 133-146). London: Sage Publications.

Toulmin, S. (1958). *The Uses of Argument*. Cambridge: CUP.

Weir, J. C. (2005). *Language testing and validation: An evidence-based approach.* New York, NY: Palgrave Macmillan.