A REVIEW ON VALIDATING LANGUAGE TESTS

Dinh Minh Thu*

Haiphong University, 171 Phan Dang Luu, Kien An, Hai Phong, Vietnam

Received 26 June 2018

Revised 22 January 2019; Accepted 25 January 2019

Abstract: Validity in language testing and assessment has its long fundamental role in research along with reliability (Bachman & Palmer, 1996). This paper analyses basic theories and empirical research on language test validity in order to provide the notion, the classification of language test validity, the validation working frames and the trends of empirical research. Four key findings come out from the analysis. Firstly, language test validity refers to an evaluative judgment of the language test quality on the ground of evidence of the integrated components of test content, criterion and consequences through the interpretation of the meaning and utility of test scores. Secondly, construct validity is a dominating term in modern validity classification. The chronic division of construct validity into prior and post ones can help researchers have a clearer validation option. Plus, test validation can be grounded in light of Messick (1989), Bachman (1996) and Weir (2005). Finally, almost all empirical research on test validity the researcher has addressed concerns international and national high-stakes proficiency tests. The research results open gaps in test validation research for the future.

Keywords: language assessment, test usefulness, construct validity, validation

1. Introduction

Testing and assessment, shortened as assessment, has become a mainstream in global language education for several decades (Bachman, 2000). Bachman & Palmer's (1996) framework of test usefulness has functioned as a fundamental basis for professional English language test development, implementation and evaluation all over the world. It is a combination of six components, namely reliability, validity, authenticity, interactiveness, practicality and impact. Amongst this integration, validity is argued to be the dominating factor to ensure that a test will measure what it claims to measure (Messick, 1989; Bachman, 1995; Hughes, 2003; Borsboom, Mellenbergh, & Van Heerden, 2004). Language test validity and validation has been investigated in the world largely in light of the validation theories proposed by Messick (1989), Bachman & Palmer (1996) and Weir (2005). They require test developers articulate their test validity. Albeith to its significance, the matter has merely become Vietnamese assessment researchers' and practictioners' concern recently (Trần, 2011; Nguyễn, 2017; Bùi, 2016; Vũ, 2016; Nguyễn, 2018; Nguyễn, 2018). This paper expects to raise Vietnamese English language teachers' awareness of language test validity, which can impact their testing practice positively.

Four research questions are raised:

1. What is the concept of language test validity?

* Tel.: 84-912362656
Email: minhthu.knn.dhhp@gmail.com

2. What are the types of language test validity?

3. How can a language test be validated?

4. What has previous empirical research on test validation revealed?

The research is initiated with the theoretical backgrounds of testing and assessment. Later on, through content analysis and practical experience, the author would present the concept, the classification, the validation framework and the results of research on validity.

2. Methodology

This secondary research is conducted analytically when the researcher bases on the available sources of information to evaluate the interested research problem critically (Kothari, 2004, p. 3). In order to reach the unified definition and classification of validity, the researcher browses the prevailing relevant documentation. Findings from test validation framework and empirical studies undergo the same method. The data in this study comes from both objective and subjective reflective sources.

3. Theoretical background

3.1. Language testing and assessment

Testing and assessment are two terms which are currently in common parlance. While tests are defined as "a method of measuring a person's ability, knowledge or performance in a given domain" (Brown, 2003, p.3), assessment is referred to as an "ongoing process" (Cizek, 1997) or "an ongoing strategy" (Brown, 2004). Cizek's (1997) definition of assessment is selected herein for its relative wholeness:

1. the planned process of gathering and synthesizing information relevant to the purposes of (a) discovering and documenting students' strengths and weaknesses, (b) planning and enhancing instruction, or (c) evaluating progress and making decisions about students;

2. the process, instrument or method used to gather the information. (p.10)

Assessment is an umbrella term including tests with diverse educational practices (Brown, 2003). Popham (2002, p. 4) adds that "Educational assessment is a formal attempt to determine students' status with respect to educational variables of interest." The process is "formal" because it takes place professionally and systematically in the classroom context. The phrase "educational variables of interests" suggests the acceptance of variations in degrees of knowledge, learning styles, and attitudes. Therefore, assessing learners' abilities demands teachers'open-mindedness to accept diversities but keep inclusion of learning goals as well as equity among learners. Echoing the view, McTighe's (2014, p. 2) claims that assessment should (1) serve learning, (2) use diverse measurement tools, (3) align with goals, (4) measure with matters, and (5) be fair. In order to reach the goals, there should be test quality harness.

3.2. The quality of a good language test

Bachman and Palmer (1996, p.18) released a framework of test usefulness or test qualities. It is a combination of six components as presented below:

Usefulness = Reliability + Construct validity + Authenticity + Interactiveness + Impact + Practicality

To put it simply, reliability is the consistency in test results across testing times. Prior to defining construct validity, the notion of construct should be presented first. Construct is the specific ability definition used as the basis for designing a test task

and interpreting scores gained from the task. Construct validity is the degree for a test score to be interpreted and generalised accurately to indicate the ability in measurement. Authenticity refers to the correlation between the test tasks and the target language use. Interactiveness pertains to the engagement of test takers when performing the test. Impact means the test effect on stakeholders like learners, teachers, authorities and parents. Lastly, a test is practical when resources for developing, implementing and conducting the test are available and applicable.

To reach the target of usefulness, the test developer must identify a certain test purpose, a certain test taker, and a target language use domain. These qualities are integrative although the degree can vary across the contexts. A high-stake test puts more emphasis on reliability and validity while a classroom test can have more elements of authenticity, interactiveness and impact. Reliability and validity are core measurement qualities of a test because they are closely reflected by the score interpretation, while the remaining components concern more the societal aspects of a test.

4. Findings and discussion

4.1. Concepts of validity

It is undeniable that validity is a key concept in the field of testing and assessment (Messick, 1989; Lado, 1961; Fulcher & Davidson, 2007; Bachman, 1995; Hughes, 1989; Borsboom, Mellenbergh, & Van Heerden, 2004; Shepard, 1993). Cronbach & Meehl (1955) who are credited fathers of construct validity define validity as *the consistence* between test score interpretations and a nomological network involving theoretical and observational terms. Validity in a language test was first mentioned by Lado (1961). He claims that if a test measures what it purports to measure, it is valid. This claim sounds general and hard to be evaluated. The American Psychological Association (1995, p.9) makes it clearer that validity is "the *appropriateness, meaningfulness, and usefulness of the specific inferences made from the test scores"* (cited in Bachman, 1995, p. 243). Here we see the role of the test score to evaluate the validity of the test. In 1989, Hughes considers validity *the test ability* to announce a test to be valid by measuring "accurately what it is intended to measure" (p. 22). Concurrently, Messick (1989, p. 245) designates validity "an overall *evaluative judgment* of the degree to which empirical and theoretical rationales support the **adequacy** and **appropriateness** of **interpretations** and **actions** based on test scores and other modes of assessment". He regards construct validity as social consequences of testing, which can impose positive or negative washback on the users because it can determine the meaningfulness, appropriateness and usefulness of the test through the interpretation of the test score. Another approach to test validity is to label it the *test property* being evaluated rather than the judgment of the test (Borsboom et al., 2004).

All in all, as a majority of language testing and assessment experts state it, validity in a language test refers to an evaluative judgment of the language test quality on the ground of evidence of the integrated components of test content, criterion and consequences through the interpretation of the meaning and utility of test scores. Test scores are usually rendered to provide evidence for validity (American Psychological Association, 1995; Cronbach & Meehl, 1955; Lado, 1961; Messick, 1989). Nonetheless, even in Messick (1998)'s definition, validity can be examined with

other means except for test scores. This new light will be elaborated in the coming part.

4.2. A combined framework of validity

Validity standards made its debut in 1954 by the American Psychological Association in four forms namely *predictive validity*, *concurrent validity*, *content validity* and *construct validity* (Shepard, 1993). *Predictive validity* can be observed after test administration to predict the future performance while *concurrent validity* refers to the concurrency of the test score and the criterion of an already-accepted test. *Content validity* concerns the comparison between test specifications and test contents. Among types of validity, *construct validity* is the most complicated, which gets its evidence from the comparison between the need-to-be-proved-valid item and the supposed-valid item. Herein, it is important to clarify one key concept as "construct" in the field of testing and assessment. It is the definition of a *specific ability* used as the basis for designing a test task and interpreting scores gained from the task (Bachman & Palmer, 1996). Hence, *construct validity* denotes the degree for a test score to be interpreted and generalised accurately to indicate the construct or ability in measurement (Bachman & Palmer, 1996). Construct validity is qualified both qualitatively and quantitatively (Bachman & Palmer, 1996; Messick, 1998; Weir, 2005).

In 1966, the Association revised the validity structures to make it a "Trinitarian doctrine", including *construct validity, concurrent validity* and *criterion-related validity* (combined by *predictive validity* and *concurrent validity*) (Shepard, 1993). Lado (1961) and Davies (1968) add the element of *face validity*, which is decided by the look at the test appearance, to the *content validity*. From another aspect, Campbell and Standley (1966) introduce *internal validity* and *external validity*. The former is a vital quality, shown through the analysis of the test content, whilst the latter finds out the test generability for a test to be applied to different contexts based on the test score. External validity is stated to belong to criterion validity. Alderson, Clapham and Wall (1995) echo the classification of validity into internal and external classes and label external validity as criterion-oriented validity. D'Este (2012) reviews Messick (1989)'s new contribution to the validity framework on the ground of the test score. The new unified framework is composed of two facets, one being the source of test justification from "either evidence or consequence", and another one concerning the function of the test outcome through "interpretation or use" (Messick, 1989, p. 20).

	Test interpretation	**Test use**
Evidential basis	Construct validity	Construct validity + Relevance/ utility
Consequential basis	Value implications	Social consequences

Figure 1. Facets of validity (Messick, 1989, p. 20)

In 1996, Messick went further to discuss consequential validity along with the term *washback*. He argues that "washback is a consequence of testing that bears on validity only if it can be evidentially shown to be an effect of the test and not of other forces operative on the educational scene" (p.2). In this way, it is not easy to measure the washback of the test. To make it possible, in case test specifications are widely known to the test

developers, test users and test takers, it can be said that content validity leads to washback on test preparation through making test takers familiar with the test and reduce their anxiety (Messick, 1996, p. 6). Positive washback can be enhanced by a valid test (Morrow, 1986; Anderson & Wall, 1993; Frederiksen & Collins, 1989; cited in Messick, 1996). That is why it is important to find out the evidence of validity in a test. Through Messick's (1989) lenses, general validity consists of six aspects: *the content aspect*, *the substantial aspect, the structural aspect, the generalizability aspect, the external aspect* and *the consequential aspect*. Except for the content and structural aspects, the four remaining criteria pertain to the interpretation of the test score. The content aspect is shown the relationship between the content relevance and representativeness of technical quality like the appropriate reading level. The substantive aspect includes both the theoretical ground and empirical evidence

The consequential validity of the test refers to the evaluation of both intended and unintended consequences of score interpretation and use concurrently and in the future, with evidence of bias in scoring and interpretation, the positive or negative influence on class instructions and knowledge acquisition (Messick, 1996, p. 13). It is interesting when Messick (1996, p.14) claims that validity of a test should be investigated as an assumed basis for washback.

According to Bachman (1995, p. 244-256), a framework of validity comprises *content validity* (actualized by the content relevance and content coverage), *criterion validity* (shown through concurrent validity and predictive validity), *construct validity* (revealed by the meaningfulness of construct). The consequential basis of validity is discussed in Bachman (1995) in that a test is not designed for the sake of the test but for a proposed consequence. The consequential validity signals the shift from technical, empirical and logical focus to the test use or policy focus (Bachman, 1995).

Weir (2005), one more significant theorist of validity, categories validity as construct validity in accordance with the temporal consequence; therefore two major types of validity are *priori validity* and *posterior validity*. The former can be investigated before the test event, embracing *theory-based validity* and *context validity*. By comparison, the latter accumulates evidence during and after the test event and is divided into *scoring validity, criterion-related validity* and *consequential validity.* Theory-based validity emphasizes the test developers' knowledge of theories pertaining to the underlying language processes for real life application (Weir, 2005, p.18). Context validity is traditionally referred to as content validity but Weir uses this modern term with an intention to cover both the test contents and the test administrative setting (Weir, 2005, p.19). Scoring validity measures the stability of the test results over time "in terms of the content sampling and free from bias" (Weir, 2005, p.23). In this sense, scoring validity is popularly known as reliability. The two sub-types namely criterion-related validity and consequential validity echo Bachman (1990) and Messick (1989).

From the above discussion, it can be concluded that validity is a very complicated concept and labels of validity types can be overlapped from different authors' perspectives. Nonetheless, the term "construct validity" plays the key role from the beginning of the classification of validity and continues taking its prioritized place (Messick, 1989; 1996; Shepard, 1993; Weir, 2005; Bachman, 1995). So far, the framework of validity can be visualized as follows:

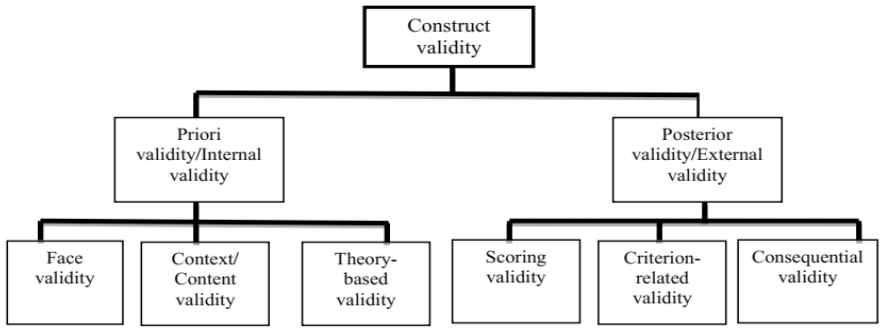

Figure 2. Classification of validity aspects

In accordance with the latest views on validity, construct validity even functions as a superordinate term (Messick, 1989; Messick, 1996; Alderson, Clapham, Wall, 1995; Weir, 2005), containing content validity and consequential validity which will be investigated further in the main parts of this study to find out the link between them in the priori test event. It is important to note that although Messick (1989, 1995) views washback as an integral part of validity, Bachman (1990) still puts washback as an equal independent unit besides validity in his framework of test usefulness. Two radical threats to validity as stated by Messick (1989) include "construct-under-representation" and "construct irrelevance". If these take place in the test, class instructions will be alleviated, causing negative washback. Messick (1996, p.16) claims that a test can be validated by reducing evidence of "construct underrepresentation and construct irrelevancies", from which the potentially positive washback can be intensified and good educational practices can be enhanced.

4.3. Language test validation

In order to figure out the validity of a test, the procedures of validation are conducted. Test validation is defined as "the process of generating evidence to support the well-foundedness of inferences concerning traits from test scores" (Weir, 2005, p.1). It is a form of evaluation accumulating evidence both quantitatively and qualitatively (Messick, 1989; Messick, 1996; Weir, 2005; Bachman, 2000). "Test validation is the empirical evaluation of the meaning and consequences of measurement, taking into account extraneous factors in the applied setting that might erode or promote the validity of local score interpretation and use" (Messick 1996, p. 246).

Messick's (1989) unitary framework of validity which merges both construct validation and consequential validity has been widely used as a model in a large volume of educational and psychological research (Bachman, 2000). Evidence of validity, according to him, can be generated in unlimited methods, for example, an investigation of the correlation between the test content and the content of the domain identified as sources of inferences, a study of the correspondence among internal factors of the test, or an examination of the connection between the test scores and the test external structures like the background variables. Another famous validation framework is proposed by Bachman and Palmer (1996). They suggest accumulating evidence of language knowledge, metacognitive strategies and topical knowledge by answering nine questions:

1. Is the language ability construct for the test clearly and unambiguously defined?

2. Is the language ability construct for test relevant to the purpose of the tests?

3. To what extent does the test task reflect the construct definition?

4. To what extent do the scoring procedures reflect the construct definition?

5. Will the scores obtained from the test help make the desire interpretations about test takers' language ability?

6. What characteristics of the SETTING are likely to cause different test takers to perform differently?

7. What characteristics of the test RUBRIC are likely to cause different test takers to perform differently?

8. What characteristics of the TEST INPUT are likely to cause different test takers to perform differently?

9. What characteristics of the EXPETED RESPONSE are likely to cause different test takers to perform differently?

10. What characteristics of the RELATIONSHIP BETWEEN INPUT AND RESPONSE are likely to cause different test takers to perform differently?

(pp. 140-142)

As required in the questions, the test construct is the primary concern in its relevance to its obvious clarification, test purpose, test tasks and score interpretation. Regarding the test performance, the test setting, test rubrics, test input, oriented response, as well as the correspondence between the test input and test response are also worth consideration. The variations of these factors are likely to swing the test results, which will make it hard to reach the appropriate conclusion of test takers' language ability. Test takers are likely to perform a listening test better if they are in a sufficiently small room with sufficiently loud recording, for example. These theories sound reasonable and can be applied for validating tests of various educational areas. However, the details need more discussions. Or else, a language test will require a more specific set of questions to be answered.

Weir (2005) proposes four socio-cognitive validation frameworks corresponding to four language skills at two phases before and after the test event. As previously mentioned, Weir's (2005) classification of validity embraces five types as content validity, theory-based validity, scoring validity, criterion-related validity and consequential validity. The same structure of validating the assessment of each skill is depicted, initiating from *test taker characteristics* to the two first types of *context validity* and *theory-based validity*. From theory-based validity, *responses* are collected for *scoring validity* which bases on the score of the test, then followed by *consequential validity* and *criterion-related validity.*

Four language skills are operated in the same validation procedure as follows:

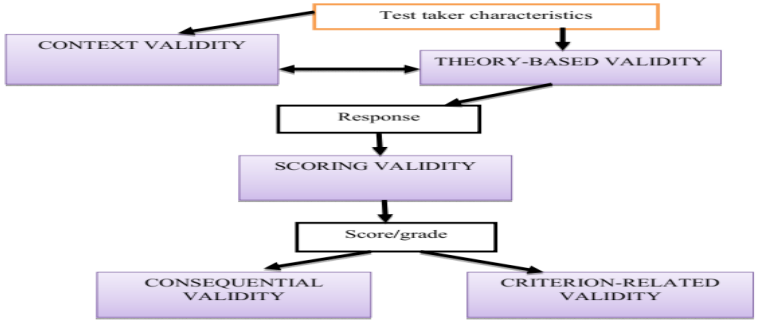

Figure 3. A socio-cognitive validation framework of language skills (adapted from Weir, 2005)

In all the above validation suggestions, the quality of test construct, test input and test characteristics are studied first. The correlation between the test input, response and scoring is investigated to validate the test. Messick (1975; cited in Sheparrd, 1993, p.414-415) did not consider the role of content validity because of the traditional thought of validity coming from the test score. Nonetheless, this view has been changed (Yalow & Popam, 1983; Messick, 1989, 1996; Shepard, 1993; Bachman, 1990; Weir, 2005) on the ground that content validity functions as a precursor to reach appropriate score interpretations. Therefore, content validity deserves a serious investigation prior to the implementation of the test.

As presented, concerning language tests, while validity is largely discussed in terms of its definitions and aspects, validation has its limited procedures, despite its complicateness.

4.4. A review of large-scale test validation studies

High-stake language tests have been validated by international and local researchers, exploiting both qualitative and quantitative approaches. High-stakes tests like entrance/ placement university tests or IELTS, TOEFL are widely investigated (Fulcher, 1997; Tran et al., 2010; Ito, 2001; Bui, 2016; Zahedkazemi, 2015), besides the tests measuring tertiary students' achievement or proficiency language tests at individual universities (Rethinasamy & Nong, n.d.; Choi, 1993; Zahedkazemi, 2015; Hiser & Ho, 2016; Graves, 1999; Sims, 2015; Choi, 1999; Zahedkazemi, 2015; Trần, 2011). Both positive and negative findings have been found from the research. Bachman's (1990) and Messick's (1989) validation frameworks are popularly exploited.

Take a look at the first stream of validating entrance tests. According to Hitotuzi (2002), the entrance examination of the Federal University of Amazonas lacks both face and content validity. Spelling and grammar mistakes are found. The test is blamed to have complex syntax and lexis regardless of normal language education at high school. Bui (2016) investigated the test usefulness of the Vietnam's College English Entrance Exams (VCEEE) between two tests in 2014 and 2015. She also uses Bachman and Palmer (1996)'s model of language knowledge to validate the test. It is reported that validity is supported by the test methods of gap filling and closes, but multiple item test methods, error detection and synonym/antonym selection cause problems of interpreting correct test takers' ability. In addition, multiple choice questions is the sole test method in the old version, which is mended by the subjective writing parts of sentence rewriting and paragraph rewriting. Zahedkazemi (2015) conducts construct validation of two global sub-tests IELTS and TOEFL basing on the test scores. The results show that both tests share differences and similarities in gauging test takers' language proficiency. In 2010, Tran et al. (2010) built up the conceptual framework and the methodology for the validation of the interpretation and use of the 2008 University Entrance Examination English test scores, exploiting Messick (1989)'s unified validation framework. Content analysis, Rasch modelling and path analysis contribute to the methodology in details.

The second stream also records interesting cases of validation. Choi (1994) measures the content and construct validation of a criterion-referenced English proficiency test in order to come to a valid standardized test labelled Seoul National University Criterion-Referenced English Proficiency Test (SNUCREPT). Bachman's (1990) framework of communicative language

ability is exploited. The qualitative and quantitative approaches with native speakers and computable tools respectively are mixed. He claims that systematic development of the test can satisfy the validity and reliability of the test. Choi (1999) validates the Test of English Proficiency (TEPS), developed and utilized in Seoul National University by collecting both qualitative and quantitative feedback from the test takers on the pilot test and the first administrative test to see the validity of the test and the test fairness. He makes the comparison between the test in study TEPS and the valid test TOP (Test of Oral Proficiency). The analysis of the test score is made, along with an interview of respondents who got higher TEPS scores and available TOEIC scores and who are teachers of English. In terms of the test score analysis, high correlation is found between the data from the two tests, illustrated by the correlation coefficients of over .63. Regarding the interview result, 42.7 of respondents strongly agree on the test method/ fairness of the test. Ito (2001) validates the Join First Achievement Test (JFSAT) – Japanese nationwide university entrance examination by investigating the reliability, concurrent validity, criterion validity and construct validity of the test which is divided into five components, including pronunciation, grammar, spoken English, written English and reading comprehension. The finding reveals that instead of the low reliability coefficient of the paper-pencil pronunciation test (r =.208), other figures proves JFSAT a relatively valid test of English ability. In terms of the construct validity of the test, low correlation coefficients remain in the pronunciation (r =.238, n.s) and spoken English (r =.600, compared to the demanded criterion of r >.7). The pronunciation score has very little contribution to the overall score.

In Vietnam, Trần (2011) finds out the evidence of the content validity of an English achievement test for second year non-English major university students by using survey questionnaires for both teachers and students to see the degree of unsatisfactory level in some parts of the test due to the insufficient preparation in designing test specifications and writing part instruction. Hoang (2009) also supplies the same results in terms of test specifications. Rethinasamy & Nong (n.d) study the validity of the Advanced Educational Program English Test (AEPET) at a university in Vietnam on three aspects, including concurrent validity, predictive validity and content validity. IELTS scores are exploited to validate concurrent validity. Scores of AEPET in four components: listening, reading, speaking and writing are used to validate the test content, revealing high validity degree in the speaking and reading tests and moderate degree in the two remaining tests. The overall mean scores is also moderate at 3.35. Test preparation is included into the content validation, which shows an insufficient amount of instructions. Although the problem identified in the paper is interesting, the authors have not provided details in the validation method. Consequently, the result discussion is merely on the surface. In 2017, Nguyễn studies the cut-score validity of the VSTEP.3-5 listening test using Kane's (2006) current argument-based validation approach focusing on test tasks, accuracy and precision and cut scores. Findings show that the test tasks follow the test specification strictly, the language input relatively meets the demand. In terms of precision and accuracy, on the whole, the test can discriminate test takers to a reasonable extent. The Angoff method and Bookmark method are used to gauge the cut scores. By comparison with the expected reliability of

at least 0.88, VSTEP listening test reliability index is 0.815, which is rather low.

All in all, an insight into the experimental research of language test validity points out three pivotal matters. Firstly, in terms of methodology, both quantitative and qualitative approaches are exploited. Scoring validity, for example, suits the former while the latter applies to content validity. Secondly, high-stakes international and national language tests are the subjects in studies. Last but not least, validation mainly occurs to posterior or external validity.

5. Conclusion and pedagogical implication

So far, four research questions have been answered. A language test can claim its validity when it can measure exactly the test taker's language ability actualized by the test construct. In the past, construct validity is distinguished from content validity and criterion validity, but the modern view puts construct validity the umbrella concept and classifies validity into more types. The idea of prior validity and posterior validity proposed by Weir (2005) is worth considering. Weir (2005)'s validation model is also very interesting and specific for a language test, covering four language sub-skills. In Vietnam, test validation also largely pertains to high-stakes tests, especially a newly designed national test VSTEP (Vietnamese Standardised Test of English Proficiency) at the University of Languages of International Studies, Vietnam National University, Hanoi. Testing has never lost its society's concern. However almost all important tests have not been validated. English gate-keeping tests at universities or English entrance university exams in 2017 and 2018, for example, all deserve validation. In addition, there leaves a gap in the documentation of using Weir (2005)'s models to gauge the validity of an overall internally-developed achievement test. More importantly, the result of validity will serve as evidence for washback, as Messick (1996, p.252) claims: "rather than seeking washback as a sign of test validity, seek validity by design as a likely basis for washback". He also adds that all tests are in danger of construct irrelevance and construct under-representation. Compared to the theories in validity, research has not covered all. It is impossible to reach full validation, but recommendations to increase the degree of validity can be (1) making test specifications explicit, (2) maximizing direct testing, (3) closely linking the scoring of response to the test purpose, and (4) ensuring test reliability (Hughes, 1983).

References

Bachman, L. F. (1995). *Fundamental Considerations in Language Testing* (Third Edition). Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1–42. Retrieved from https://doi.org/10.1177/026553220001700101

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. Retrieved from https://doi.org/10.1037/0033-295X.111.4.1061

Bui, T. S. (2016). *The Test Usefulness of the Vietnam's college English Entrance Exam* (Master's Thesis). Korea University, Seoul.

Choi, I. (1993). Construct Validation Study on SNUCREPT (Seoul National University Criterion-Referenced English Proficiency Test)*. *Language Research, 29*(2), 243–275.

Choi, I. (1999). Test Fairness and Validity of the TEPS. *Language Research, 35*(4).

D'Este, C. (2012). New views of validity in language testing. *EL.LE, 1*(1), 61–76.

Fulcher, G. & Davidson, F. (2007). *Language Testing and Assessment - an advanced resource book.* London and New York: Routledge.

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language Testing, 14*(2), 113–139. Retrieved from https://doi.org/10.1177/026553229701400201

Graves, K. (1999). *Validity of the secondary level English Proficicency test at Temple University - Japan.* Princetton, NJ: Educational Testing Sevice.

Hiser, E. A. & Ho, K. S. T. (2016). C-Tests in Vietnam : An Exploratory Study of English Proficiency. *Electronic Journal of Language Teaching, 13*(2), 184–202. Retrieved from http://e-flt.nus.edu.sg/

Hughes, A. (2003). Testing for Language Teachers. *Australian Review of Applied Linguistics, 27*. Retrieved from https://doi.org/10.1017/CBO9780511732980

Ito, A. (2001). A Validation Study on the English language test in a Japanese Nationwide University Entrance Examination. *Asian EFL Journal, 7*(2), 11–33.

Kothari, C. R. (2004). *Research methodology: methods and techniques* (Second revision). New Age International Publishers. Retrieved from https://doi.org/http://196.29.172.66:8080/jspui/bitstream/123456789/2574/1/Research%20Methodology.pdf

Messick, S. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher, 18*(2), 5–11. Retrieved from https://doi.org/10.3102/0013189X018002005

Messick, S. (1996). *Validity and washback in langauge testing. Language Testing.* Retrieved from http://ltj.sagepub.com/content/13/3/241.short

Nguyen, T. N. Q. (2018). A study on the validty of VSTEP writing tests for the sake of national and international integration. *VNU Journal of Foreign Studies*, *34*(4), 115–129.

Nguyen, T. P. T. (2018). An investigation into the content validity of a Vietnamese standardised test of English Proficiency (VSTEP.3-5) Reading Test. *VNU Journal of Foreign Studies*, *34*(4), 129–143.

Nguyen, T. Q. Y. (2017). *Summary of doctor dissertation and investigation into the cut-score validity of the VSTEP. 3-5 listening test*. University of Languages and International Studies, Vietnam National University, Hanoi. Retrieved from http://saudaihoc.ulis.vnu.edu.vn/files/uploads/2017/12/Tom-tat-TA.pdf

Rethinasamy, S. & Nong, T. H. H. (n.d.). *Investigating the validity of the advanced educational program English test of Vietnam with IELTS: Implications for quality management of in-house test.* Universiti Malaisia.

Shepard, L. A. (1993). Chapter 9: Evaluating Test Validity. In L. Darling-Hammon (Ed.), *Review of Research in Education, 19*(1), 405–450. Retrieved from https://doi.org/10.3102/0091732X019001405

Sims, J. M. (2015). A Valid and Reliable English Proficiency Exam: A Model from a University Language Program in Taiwan. *English as a Global Language Education (EaGLE) Journal EaGLE Journal, 1*(12), 91–125. https://doi.org/10.6294/EaGLE.2015.0102.04

Tran, H. P., Griffin, P., & Nguyễn, C. (2010). Validating the university entrance English test to the Vietnam National University: A conceptual framework and methodology. *Procedia - Social and Behavioral Sciences*, *2*(2), 1295–1304. Retrieved from https://doi.org/10.1016/j.sbspro.2010.03.190.

Tran, Q. T. (2011). *The Content Validity of the Current English Achievement Test for Second Year Non Major Students at Phuong Dong University* (Master's thesis). University of Languages and International Studies, Hanoi.

Vu, T. P. A. (2016). 25 years of language assessment in Vietnam : Looking back and looking forward. In *New Directions in English Language Assessment in Vietnam*. Retrieved from https://www.britishcouncil.vn/.../new_directions_2016_dr_vu_thi_phu...

Weir, C. J. (2005). *Language Testing and Validation. An Evidence-based approach.* New York: Palgrave-Macmillan.

Zahedkazemi, E. (2015). Construct Validation of TOEFL-iBT (as a Conventional Test) and IELTS (as a Task-based Test) among Iranian EFL Test-takers ' Performance on Speaking Modules, *Theory and Practice in Language Studies, 5*(7), 1513–1519.

KHẢO CỨU VỀ XÁC TRỊ BÀI THI NGÔN NGỮ

Đinh Minh Thu

Trường Đại học Hải Phòng, Số 171 Phan Đăng Lưu, Kiến An, Hải Phòng, Việt Nam

Tóm tắt: Song song với độ tin cậy, độ xác trị trong kiểm tra đánh giá ngôn ngữ từ lâu đã giữ vai trò quan trọng trong các nghiên cứu (Bachman & Palmer, 1996). Bài báo này phân tích các lý thuyết cơ bản và các nghiên cứu thực nghiệm về độ xác trị để cung cấp khái niệm tính xác trị trong kiểm tra đánh giá ngôn ngữ, các tiểu loại xác trị, các khung lý thuyết đo độ xác trị và các khuynh hướng nghiên cứu thực nghiệm tính xác trị. Có bốn kết quả chính thu được qua phân tích. Thứ nhất, tính xác trị trong bài kiểm tra ngôn ngữ đánh giá chất lượng bài kiểm tra ngôn ngữ dựa trên nội dung bài thi, tiêu chí bài thi, hệ quả bài thi thông qua việc xác định ý nghĩa và việc sử dụng điểm số. Thứ hai, độ xác trị của năng lực ngôn ngữ là một thuật ngữ chủ chốt khi phân loại các độ xác trị. Thêm vào đó, khung phân loại tiền xác trị và hậu xác trị sẽ giúp nhà nghiên cứu lựa chọn hướng xác trị rõ ràng hơn. Thứ ba, khung lý thuyết xác trị dựa trên ba mô hình chính của Messick (1989), Bachman (1996) và Weir (2005). Một kết luận nữa trong nghiên cứu này là phần lớn các nghiên cứu về độ xác trị mà tác giả đã tiếp cận đều dựa trên các bài thi có tầm quan trọng lớn, ở quy mô quốc tế hoặc quốc gia. Kết quả nghiên cứu cho thấy mảnh đất nghiên cứu độ xác trị trong bài thi ngôn ngữ còn rất rộng.

Từ khóa: đánh giá ngôn ngữ, dụng tính của bài thi, độ xác trị về năng lực, việc xác trị