Advances in Language and Literary Studies
ISSN: 2203-4714
Vol. 5 No. 4; August 2014

A Historical Overview on the Concept of Validity in Language Testing

Mehraban Hamavandy
English Department, Tarbiat Modares University, Tehran, Iran
E-mail: mehraban2544@gmail.com

Gholam Reza Kiany (corresponding author)
English Department, Tarbiat Modares University, Tehran, Iran
E-mail: rezakiany@yahoo.com

Doi:10.7575/aiac.alls.v.5n.4p.86 Received: 18/05/2014
URL: http://dx.doi.org/10.7575/aiac.alls.v.5n.4p.86 Accepted: 30/06/2014

Abstract

This article provides an overview on language test validation theories, especially the Messickian view on construct validity and the way it's been translated into practice. First, a brief historical synopsis will be set forth, followed by recent views on test validity as advanced by Messick and Kane. The review goes on to lay out the similarities and differences between various validity conceptions, and concludes with their critical evaluation.

Keywords: validation, language testing, validity theory

1. A Historical Review

In the realm of language testing and assessment, the main goal, predominately sought after, is to elicit the knowledge of the test takers in a particular linguistic domain through a test. The test must be capable of producing results deemed to be a true representation of the language ability the test taker can maintain in the real language situation (Weir, 2005, pp. 44-46).

It's a conviction that concerns about the notion of 'validity' of a test is as old as the practice of testing itself (Lissitz,, 2009, p. 12). The primordial instance of a high-stakes test whose validity should largely have been of concern is the biblical story of the Gileadites who developed a test to detect among them Ephraimites, as their foes. The test consisted of the correct utterance of the word "shibboleth" whose failure would result in beheading of the failed individuals.

Yet, Alfred Binet has been probably the most outstanding figure, in the modern era of educational assessment, who developed a test with consideration of its validity. Binet's intelligence test aimed to segregate and to predict the lower versus higher ability school children in the early 20th century of Paris. Binet's test displayed significant features of what is now esteemed as a valid test, i.e. children were tested under similar conditions with the same test and were rated across a similar procedure. Though a validity index of this test was not reported then, a quick look at some other later developed tests of the early 20th century (e.g. Wechsler's intelligence tests, or Stanford-Binet's intelligence test) indicates that the concept of 'validity' was mostly taken as a correlation between the test and a given predefined criterion. This correlational mindset towards validity is justifiable partly due to the fact that, back then, Pearson had already introduced his formula for the correlation coefficient (Pearson, 1896). Pearson's correlation was a great asset to investigate how well performance on a test is related to a well-accepted conceptualization of the attribute tested.

Important works of that era on educational measurement vividly define validity as such, namely Kelly's (1927) elaboration on validity or Bingham's (1937) definition of a valid instrument for aptitude measurement. This correlational approach to validity was the unrivaled and dominant form of validity investigation throughout the first half of the 20th century, as is summed up by Guilford (1946) that "a test is valid for anything with which it correlates" (p. 429). "Bowl empiricism" is the label Morphon (2007) attributes to that era emphasizing that by then "the empirical finding was considered sufficient to sustain the inferences made on the basis of the test scores" (p. 5). Similarly, Kane (2012), in a retrospective review of validity history, opined that:

> 'Up till about 1920, there did not seem to be a very clear distinction between reliability and validity. The connection between the assessment performances and what was being measured was generally

> taken for granted, and the main concern was with the precision and cohesion of the measurements' (p.5).

Yet it didn't take long before criterion-oriented way of validity investigation saw dissatisfaction among psyco-metricians. Rulon (1946), for instance, argued that having merely a criterion-based view on validity may ignore many other significant dimensions of a test and its use. In an achievement test, for instance, he believed there should be:

> ". . . two general ways of asking about the validity. The first is, are the materials on this test, and the processes called for on these materials, the same things and processes we are trying to teach children? The other is, if they are not, do we have evidence that scores on this test go hand in hand with those we would obtain with a test about which the answers to our first questions were in the affirmative? Thus we see that the direct observation of the things and processes which are the aims of instruction is the final proof of validity, as compared to the correlation coefficient of validity, which is at best secondary" (p. 292).

In a similar vein, in their seminal article Cronbach & Meehl (1955) inveighed overreliance on the criterion-related validity as the (then) major form of test validity investigation claiming that "the asymmetry between the test and the so-designated criterion arises only because the terminology of predictive validity has become a commonplace in test analysis" (p. 4).

In the beginning of the second half of the 20th century the American Psychological Association (APA) attempted to provide a convergence on the terminology of validity and bring forth consensus on the issue by providing the Technical Recommendations for Psychological Tests and Diagnostic Techniques: A Preliminary Proposal (APA, 1952). In the proposal, four "categories" of validity were introduced, namely predictive validity, status validity, content validity, and congruent validity (p. 24). In later definitions of validity the four "categories" of validity were transformed to "types" or "attributes" of validity. Also the terms "construct validity" and "concurrent validity" were substituted for "congruent validity" and "status validity", respectively. The two influential figures who later revised and tried to disambiguate the concept of "construct validity" were Lee Cronbach and Paul Meehl. They asserted that:

> "Construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined.' The problem faced by the investigator is what constructs account for variance in test performance?" (Cronbach and Meehl, 1955, p. 282)

In other words, they tried to redefine the earlier notion and application of construct validity by laying out their definition of the issue that "construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings" (ibid, p.9).

In later revisions of APA, predictive and concurrent validity were merged to form "criterion-related" validity, an effort to ease the terminological perplexity. Thus it was hoped that by providing a tripartite typology of validity, it would be more convenient for researchers to scientifically investigate the degree of validity of a test and be able to discuss the test scores' proposed interpretations more justifiably.

Through these terminological modifications, codification of new concepts and introduction to new trends in validation research it was hoped to come up with an all-inclusive, comprehensive framework which was capable of covering the requirements of a quality based research endeavor in psychometrics. Yet the scientific community felt dissatisfied and brought into question the earlier typologies. Angoff (1988), for example, opined that ". . . construct validity as conceived by Cronbach and Meehl cannot be expressed as a single coefficient. Construct validation is a process, not a procedure; and it requires many lines of evidence, not all of them quantitative" (p. 26). Recently, Davies et al (2003) in a retrospective book chapter maintained that early notions of validity lacked harmony and coherence and thus "each approach in the classical tradition had failings…also classical validity studies looked in different and unrelated directions" (p.799).

Therefore, the time seemed right for a paradigm shift in validity theory and validation studies. In 1989, Schmitt defined construct validity as "the degree to which certain explanatory psychological concepts or constructs account for performance on a test" (p. 332). Similarly Binning and Barrett (1989) put their emphasis on defining validity as a unitary concept. They articulated that all inferences made from a test should be conceived from a unitary perspective. But undoubtedly it was Messick (1989) who shifted the validation paradigm and laid the ground for a unitary vision of validity in his groundbreaking book chapter on validation.

2. Messick's Unitary Hypothesis of Validity

A fully psychometrical approach to validation was prevalent within the realm of educational measurement well up to the end of the 1970s. As stated earlier, validity was dissected into several sub-domains and was investigated accordingly. Though at times some voices defied this fashion of validity conceptualization, they were not received glamorously. For instance, Cureton (1951) criticized the earlier notion of validity on the grounds that a single score attached to a test as its validity index doesn't fully represent its nature. Instead he considered validity as a must-be-adapted concept which should embody attention to the way a test is used and interpreted. In the first edition of Educational Measurement, Cureton (1951) conceives of validity as an attribute of a test whose essential question is:

> "…how well a test does the job it is employed to do. The same test may be used for several different purposes, and its validity may be high for one, moderate for another, and low for a third. Hence, we cannot label the validity of a test as "high" "moderate" or "low except for some particular purpose". (Cureton, 1951, p. 621)

Yet it was not until 1989 when Messick lay the foundation for a drastically different approach towards validity. Even now many consider Messick's validity statements as the most significant ones throughout the historical development of validation theory. For instance McNamara and Roever (2006) believe that:

> "The most influential current theory of validity of relevance to language testing remains that developed by Samuel Messick in his years at Educational Testing Service, Princeton from the 1960s to the 1990s, most definitively set out in his much-cited 1989 article" (p. 12).

In his seismic article Messick (1989) openly attacks and challenges what he called the "disjunction between validity conception and validation practice" (p. 34). He maintained that "construct-related evidence undergirds not only construct-based inferences but content- and criterion-based inferences as well" (1989, p. 40). He also adds that the procedure involved in a validation study should be a fully systematic attempt which takes into account all scientific approaches:

> "Test validation in the construct framework is integrated with hypothesis testing and with all of the philosophical and empirical means by which scientific theories are evaluated. Thus, construct validation embraces all of the statistical, experimental, rational, and rhetorical methods of marshaling evidence to support the inference that observed consistency in test performance has circumscribed meaning" (Messick, 1989, p. 41)

He, thus, goes on to put forth his reconceptualization of validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment"(Messick, 1989, p.13). In this oft-cited definition of validity, Messick shifts the view of validation from a product of test to interpretation of test score. Furthermore, Messick presents 'construct validity' as the underlying and core feature of any validation study with content and criterion related validity as merely two evidences, which can bolster the interpretations made about the test. He portrays his highly-quoted progressive matrix (table 1) which entails a four-way classification of validity, defined by the source of justification, which can be evidence- or consequence-based or both and the derived outcome or function of the testing that can include test interpretation or use, or both.

Messick's legacy has been his emphasis on the social consequences of a test and its impact on the individuals and the society at large (later labeled as consequential validity). This ideological change about test validity has exerted powerful influences over later frameworks of validation (e.g. socio-cognitive model or the argument-based model of validation). For instance, Weir's socio-cognitive model assumed such notions as washback, ethics and test administration procedures as indispensable issues from a validity study or the assessment use argument approach takes social beneficence as its key point of departure in any validation study (which will be more elaborated in the later sections).

Table 1. Messick's Progressive Matrix of Construct Validation

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity +Relevance/ Utility
Consequential Basis	Construct Validity +Value Implications	Construct Validity +Relevance/ Utility +Value Implications +Social Consequences

Though Messick (1985, p.5) points out that there exists a unified, central query for test validation both in theory and practice, he assumes this issue to take on four various dimensions (i.e. construct validity, relevance or utility, value implications, and the social consequences) which relate to "evidence about, rationales for, interpretations of, and uses of a particular test in a specific social context" (Cumming, 2001, p. 6). Altogether, these four aspects constitute a 'progressive matrix' which is supposed to enable a systematic appraisal of a given test's construct validity.

2.1 Messick and Construct Validity

The principle dimension of Messick's outline of validity is the construct validity. Historically construct validity has largely been used interchangeably with validity.

One of Messick's major breakthroughs in his validity layout is probably the demarcation that is set between different aspects that should be taken into account within any validity study. Messick defines construct validity as "setting out the nature of the claims that we wish to make about test takers and providing arguments and evidence in support of them" (McNamara and Roever, 2005). The other three cells in Messick's progressive matrix represent necessary evidences and rationale that should be elucidated prior to any decision making about the individuals. Briefly, Messick is highly desirous of "social usefulness" along with individual beneficence and "fairness" in his validity framework. Yet, though, they might appear theoretically attractive, Messick doesn't provide a procedure to guarantee these ideals are executed in a real testing condition. Hence, some scholars later proposed frameworks to avail Messick's model. McNamara and Roever (2005), for instance, set forth a diagram in order to systematically explore fairness (Figure 1). In this figure the relationships among test, construct and target with regards to the social and policy context at large has been explicated.

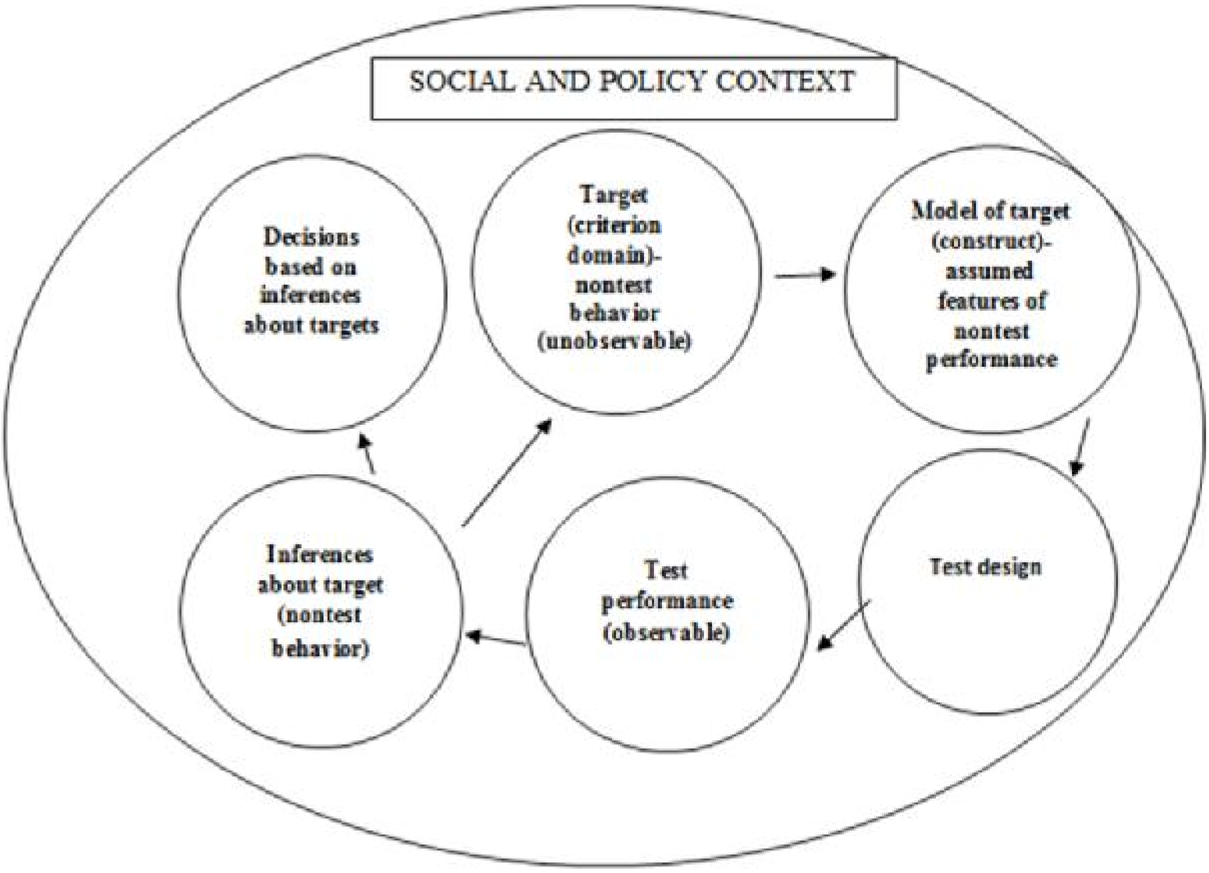
SOCIAL AND POLICY CONTEXT
Decisions based on inferences about targets
Target (criterion domain)- nontest behavior (unobservable)
Model of target (construct)- assumed features of nontest performance
Test design
Test performance (observable)
Inferences about target (nontest behavior)

Figure 1. Inferential links for a fair decision making

For example, considering the two internationally renowned tests of English proficiency, namely TOEFL iBT and IELTS, it can be claimed that the decisions made about the test scores can be highly consequential for the stakeholders. Thus if we assume that one of these high-stakes tests is used for admitting applicants into a postgraduate program, we need to articulate our expectations from the successful test takers in the form of a construct (e.g. those who pass will be able to cope with the demands of communicating in a native language setting). Thence, the iterative path starts from designing the test (based on the proposed construct) towards obtaining the sample behavior (aka test performance) which by itself can be an indicator of the probable behavior (aka language proficiency) of the individual in the actual setting. Accordingly, a fairer decision can be made about the real language proficiency of the individual test taker. This ongoing iterative process can further continue to refine the test construct assumed initially for a fairer decision making process. It is to be noted that all these procedures should be conducted with regards to the socio-political context of the test and its social consequences.

2.2 Threats to Messick's Construct Validity

Messick warns against two major sources of test invalidity, namely construct under-representation and construct irrelevant variance. Construct underrepresentation refers to the failure of a test to capture all the required features of a construct. Messick (1989) argues that "the breadth of content specifications for a test should reflect the breadth of the construct invoked in score interpretation" (p. 35). Therefore, a construct under-representative test is unlikely to measure the true abilities of the individuals according to the indicated construct.

On the other hand, construct irrelevant variance occurs when unrelated tasks, facets or dimensions to the construct creep into the test. The contamination (that could even be included inevitably) causes variance in test scores due to the inclusion of unwanted variables to the test. Construct irrelevant variance can take two forms, labeled as, "construct-irrelevant easiness" and "construct--irrelevant difficulty". As the terms imply, in construct irrelevant easiness extraneous variables cause the test taker to score higher compared with the normal condition, while in construct irrelevant difficulty the individual may obtain invalidly low scores due to inclusion of irrelevant tasks.

2.3 Criticisms Leveled at Messick's Unitary Validity

Though Messic's unitary validity theory has been much influential in shaping the way educational assessment theoreticians and practitioners view the inquiry, it has also confronted a number of outcries. First, as Shepard (1997), Chapelle (1998), and Kunnan (2004) contend, Messick's theory ignores a clear approach to implementation of the concept. Messick (1989) admits that he considers the boundaries between different categories as not watertight, but rather "fuzzy". This fuzziness between different evidences required for interpretation of test score can be delusive in the real world context. A related concern, mentioned by McNamra and Roever (2006), is the natural question that may arise about "the relationship of the fairness oriented dimensions of the top line of the matrix to the more overtly social dimensions of the bottom line, a question it could be argued that Messick never resolved and remains a fundamental issue facing our field" (p. 12). In other words, Messick leaves us empty-handed for "a means of combining the various elements of his framework in a principled and meaningful way" (Davies et al, 1990, p. 67)

Another criticism levelled at Messick's unitary theory is what Fulcher and Davidson (2007) name as problem of "validity-as-interpretation mantra" (p. 279). They raise the question that "if a test is typically used for the same inferential decisions, over and over again, and if there is no evidence that it is being used for the wrong decisions, could we not speak to the validity of that particular test-as a characteristic of it?" (p. 279). Likewise, it's been asked if it wouldn't be more instinctive to consider validity as the feature of a test (which is also generalizable) than constraining the notion of validity only to the interpretation of a test in a given context and under certain circumstances.

3. Conclusion

Finally, as Kane (2012) points out "the uniform model based on construct validity is elegant and conceptually rich and suggestive, but it is not easy to implement effectively, because it does not provide a place to start, guidance on how to proceed, or criteria for gauging progress and deciding when to stop" (p. 7) He further goes on to propose an argument-based approach to validation (aka interpretive argument framework) which, he maintains, is capable of addressing such limitations. Argument-based methods of validation (Cronbach, 1988) resort to the inferential reasoning model of Toulmin, a scholar in the field of logics in order to enable the practical implementation of the unitary view of validity in the real world of language assessment and testing.

References

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*(3), 478.

Chapelle, C.A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Second language acquisition and language testing interfaces* (pp. 32–70). Cambridge: Cambridge University Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

Cronbach, L. J. (1988). Five perspectives on validity argument. *Test validity*, 3-17.

Cumming, A. (2001). *Validation in language testing.* Multilingual Matters.

Cureton, E. F. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 621-694). Washington, DC: American Council on Education.

Davies, A. (1990). *Principles of language testing*. Oxford: Basil Blackwell.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge

Guilford, J. P. (1946). New standards for test evaluation. *Educational and psychological measurement*, *6*(4), 427-438.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3-17.

Kelly, A. L. (Ed.). (1927). *Kelly's Directory of Hampshire, Wiltshire, Dorsetshire, the Isle of Wight and the Channel Islands*. Kelly's.

Kunnan, A. J. (2004). Language assessment from a wider context. Second Language Testing and Assessment, Part VI of E. Hinkel (ed.) *Handbook of Research in Second Language Teaching and Learning.* Mahwah, NJ: Erlbaum.

Lissitz, R. W. (Ed.). (2009). *The concept of validity: Revisions, new directions, and applications*. IAP.

McNamara, T., & Roever, C. (2006). *The social dimension of language testing.* Malden, MA: Blackwell.

Messick, S. (1980). Test validity and the ethics of assessment. *American psychologist*, *35*(11), 1012-27.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11.

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review*. 16(4), 290-296.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement*: Issues and Practice, 16(2), 5-24.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength & Conditioning Research*, 19(1), 231-240.