DEVELOPMENTS IN VALIDITY RESEARCH IN SECOND LANGUAGE PERFORMANCE TESTING

Hacer Hande Uysal[*]

ABSTRACT

The present paper aims to provide a short historical overview of the theoretical developments in validity research in second language performance testing. A comparative description and critical evaluation of different views such as the "Trinitarian approach" versus the construct validity model; "uniform approach," versus "unified approach" as well as alternative and critical approaches to validation in L2 performance testing are presented. These various theoretical approaches are introduced in terms of their definitions of the validity concept, their suggested requirements for the validity research, and their attitudes towards reliability and theory while making interpretations of test scores. The paper also focuses on the current problems with the applicability of these theoretical approaches, and discusses future directions in validity research.

Key words: Second language assessment, performance assessment, validity, reliability, validity research

EARLIER THEORIES OF VALIDITY

In earlier times, validity was described as whether the test measures what it is supposed to measure (Lado, 1961 in Chapelle, 1999). Subject matter experts used to decide about the quality of the test based on the test content examining whether test tasks cover a representative sample of the target domain (Chapelle, 1999). However, this approach was based on a subjective judgment focusing only on the test content without considering the test scores. Therefore, while this approach offered evidence to support the domain relevance and representativeness of the test, it did not provide evidence about the inferences that could be made from the test scores (Bachman, 1990; Messick, 1994).

Later, Oller (1979, in Chapelle, 1999)[1] put reliability of test scores at the center of validation. Validity was defined in terms of the degree of the correlation of test scores with an older or well-established test or criterion focusing on *criterion related validity*. According to the correlations between future or present performance and the criterion, criterion validity was later

[*] Gazi University, Turkey

divided into predictive or concurrent validities (Cronbach & Meehl, 1955). However, this approach was also problematic, because it was not easy to find a well-defined valid criterion measure all the time; and even if it was found, validity of this established criterion would also be questionable. Therefore, criterion based model was not useful in many contexts (Kane, 2001).

In 50's, the construct validity was introduced as an alternative to content and criterion validity, and became one of the several types of validities. Construct validity was tied to theoretical constructs and started to be investigated by testing hypotheses related to how well the scores satisfy the theory (Chapelle, 1999; Kane, 2001). Between 50's and 70's, there were many kinds of validities –The Trinitarian Model (Shepard, 1993), and while performing validity research, the type of the validity to be addressed was chosen according to the purpose of the assessment. (e.g. content validity for achievement tests; criterion validity for selection and placement decisions, and the construct validity for theory-based proficiency tests). However, at those times, validation was still seen as "*one time activity*" (Bachman, 1990).

With the development of the construct validity model, limitations of other validation efforts started to be more apparent (Kane, 2001). The Trinitarian model was criticized as being "fragmented and incomplete" excluding "score meaning and social values from test interpretation and use" (Messick, 1995, p. 741). Cronbach & Meehl (1955), for the first time, regarded validity as a unitary concept including content, criterion, and construct based evidence under the name of "construct validity." During 80's, the Trinitarian validity definition was replaced with a single unified view of validity in the testing standards (APA, 1985). The focus of interest changed from validating test or test scores to validating proposed *interpretation of the score*s (Kane, 2001). In addition, validation became an *on-going process* through which a variety of empirical evidence about test interpretation and use had to be collected (Bachman, 1990). In addition, the consequential aspects of validity including washback, ethics and social responsibility were introduced into validity discussions (Messick, 1989). Messick (1989) proposed a "progressive matrix," which suggested that to *justify* a test score; evidence for construct validity should be gathered with consideration of value implications of the interpretation. To *use* the test scores; however, the relevance of the particular use and social consequences should also be considered.

The uniform construct validity approach suggested that interpretations of all tests – including performance tests – should be validated in the same way in terms of the theoretical constructs. Messick (1994) stated that adjusting validity criteria for language performance assessments might cause de-emphasis on important validity aspects such as construct representativeness and relevance. Bachman (2002 a,b), although he distinguished between a construct-based and a task-based approach, suggested that a cognitively based

model of language ability and use should be established for all types of assessment; only then, the similarity between the TLU domain language use tasks and assessment tasks; adequateness of the domain sampling; and extrapolation would gain meaning in validation.

According to Kane (2001) however, insistence on the necessity of a theory base for all types of assessments was meaningless especially in the areas where there is little theory, and the uniform approach to validation that is too theoretical and ambiguous caused confusion understanding what construct validity and validation study meant. Although Cronbach (1988) had made an attempt by suggesting a strong program (necessitated theory, but inapplicable) and a weak program (abstract, practical, and allowing the use of all kinds of relevant evidence without any criteria) for validation to reduce this ambiguity, the "strong" and "weak" arguments were still found to be far from being definitive and adequate to support the constructs (McNamara, 1996).

ALTERNATIVE AND CRITICAL APPROACHES: REJECTION OF THE THEORY

While the problems with regard to the inapplicability of the strong approach and the lack of criteria in the weak approach continued to cause ambiguities in validation attempts, new perspectives such as alternative paradigm and critical theory were included in the validity discussions. If we look at the validity discussions as a continuum, the "construct validity as a uniform approach" that requires a strong theory for all assessment types represents one end and the alternative approaches to validation that are skeptical of any kind of theory explaining human performance represent the other. This alternative view demanded for different criteria for validity judgments in performance assessments claiming that the complex constructs in human performance cannot be captured by any traditional theories (Lynch, 2001; Moss, 1994).

For example, Moss (1994) suggested an alternative hermeneutics approach to reliability and validity of interpretations. This view argued that validity was possible without reliability, and did not see inconsistencies in performances across tasks and among raters as a problem. According to Moss, it was possible to make generalizations across tasks by developing holistic, integrative and coherent interpretations based on a *collection of performances.* Generalization across raters, on the other hand, could be achieved through a critical dialogue and debate among raters in which initial disagreements would be resolved, and more refined interpretations would be formed by considering multiple perspectives and justifying the decisions.

Another alternative view was critical language testing, which put consequential validity at the center of the validity argument. It was suggested

that constructs are indefinable as there are multiple perspectives and no truth. Besides, all tests are subjective, relative, dependent on context, and power related; thus, there is no true score to be estimated (Shohamy, 2001; Lyncy, 2001). Validity framework according to this view was based merely on consequences; therefore, information about fairness, ontological authenticity, cross-referential authenticity, consequential validity, and evolved power relationships had to be collected (Lynch, 2001).

A MIDDLE WAY: OBSERVABLE TRAITS VS. THEORETICAL CONSTRUCTS

The uniform construct validity approach was too vague and inapplicable, whereas the alternative-critical approaches focused too much on consequences and completely rejected constructs and other important validity requirements. Kane (2001), on the other hand, suggested an alternative approach that was *unified*, but flexible in which, different kinds of validity arguments to support different kinds of inferences could be made according to the context. While the details of the validity argument for each interpretive argument would be unique, the general approach to specify and evaluate the inferences would be consistent or unified. Kane's validity definition did not require a theory; yet, it was still reflecting on the general principles in the construct model. Kane suggested adopting an argument-based approach – an *interpretative argument* – rather than validation research.

In the interpretative argument, there were several chains of inferences to be followed: 1) evaluation of performance on each task and giving a score; 2) generalization of the score beyond the observed to a universe of possible performances on similar tasks under similar circumstances; 3) extrapolation of the results beyond the testing context to various other contexts and task formats; 4) explanation and decision-making based on the theory. Kane suggested that all evidence relevant to each inference should be collected, alternative interpretations should be eliminated, and the most problematic assumptions – the weakest link—should be evaluated (Kane, et. al, 1999; Kane, 2001). Therefore, generalizibility link in performance assessments should be handled carefully because it is the weakest link due to the use of small number of tasks representing a narrow range of TLU domain and due to the variability associated to raters, task, and especially person-task interaction (Mc Namara, 1997). According to Kane, if generalizibility link fails, it is not possible to talk about extrapolation, and failure of any of the inferences fails the argument as a whole (Kane, 2001).

Kane (2001) also suggested that a distinction should be made between *theoretical constructs* and *observable attributes*. According to Kane, theoretical constructs and observable attributes are different both in terms of

validity definitions and interpretations, and the distinction is context dependent. Therefore, it is possible to limit the argument to a certain set of inferences such as evaluation of task accomplishment, or generalization to a specific universe of observation without the necessity of theory. For example, if the target is the piano performance, then scores can be interpreted as observable attributes without a need to generalize beyond test. Therefore, for observable attributes, interpretive argument involves only the inferences of evaluation, generalization, and extrapolation. However, for theoretical constructs, one more inference is needed to explain scores in terms of a construct and to interpret them as indicators of specific abilities (Kane, et al., 1999).

CURRENT TRENDS AND FUTURE DIRECTIONS

Validity is currently defined in Standards parallel to Messick's construct validity model as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, & NCME, 1999, p.9). However, Borsboom et al., (2004) claim that despite all the evolutions in the concept of validity over the years, when asked, most researchers in the field of psychology still define validity as "whether a test measures what one intends to measure" probably due to the failure of the construct validity model in providing a clear and workable conceptual scheme for practitioners.

In language performance assessment however, Kane's interpretative argument seems to be acknowledged to offer a feasible plan for validation, and has already been put into practice by Chapelle et al., (2004, in McNamara & Roever, 2006) in validating TOEFL. Recently elaborating on Kane's model, Bachman (2005) has developed the "assessment use argument" as a conceptual framework based on Toulmin's structure of reasoning involving claims, warrants, evidence, and rebuttals to achieve validation. Bachman's argument consists of two parts: 1) an assessment utilization argument that links assessment performance to a decision; and 2) an assessment validity argument that links the assessment performance to an interpretation. According to this model, since the aim is to justify a specific assessment, a "local theory" is sufficient to make claims about the decisions and interpretations based on the assessment, and to determine the types of evidence that needs to be collected to support these claims.

Although Bachman's model seems to be comprehensive and practical at first glance, given the context-dependent intricate interactions inherent in L2 construct during a performance, it may still be a problem to require a theory base for all assessment types. For example, the social interactive view states that constructs in performances are co-constructed through social interactions,

socially and culturally embedded, and context- dependent; therefore, ability, ability in language user, and context are inseparable making it impossible to measure the underlying abilities (Chalhoub-Deville, 2003).

Therefore, the distinction made by Kane between theoretical constructs and observable traits without requiring a theory all the time is a sensible approach. As Chalhoub-Deville & Deville (2005) suggest, according to the purpose and context, it is possible to seek different validity arguments and prioritize evidence for a particular use or decision-making. If we are interested just in the performance/task fulfillment, replicability and generalizibility would not be the issue; however, if we are interested in performance with relation to the construct definition/language ability, generalizibility would be necessary because the consistency or variability of performances contributes to the score meaning.

In conclusion, linking the performances to a theory to be able to interpret the results in terms of abilities and accordingly to be able to generalize is the most problematic area in validation. Therefore, further research is needed primarily to determine constructs which are very complex and elusive in performance assessment. Chalhoub-Deville (2003) suggests that it might be possible to determine any stable constructs that are accessed in similar ways across contexts by analyzing tasks and interacting factors in performance assessments in different contexts especially through ethno-methodological research so that the association networks used in varied situations to transfer knowledge and skill can be understood, and generalizibility across contexts can be achieved. In addition, as social consequences of tests are not adequately integrated in validation models, development of a social theory regarding the social and political context in which assessment takes place should also be considered to understand the potential sources of unfairness and the meaning of test use in context (McNamara, 2006).

BIBLIOGRAPHY

American Psychological Association (APA) (1985). *Standards for educational and psychological testing.* Washington, DC.

AERA, APA, & NCME (1999). *Standards for educational and psychological testing.* Washington, D.C.

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

Bachman, L. F. (2002, a) Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice*, *2*(3), 5–18.

Bachman, L. F. (2002b). Some reflections on task-based language performance assessment. *Language Testing, 19*, 453-476.

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

Borsboom, D., Mellenbergh, G. J. & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111* (4), 1061-1071.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing, 20* (4), 369-383.

Chalhoub-Deville, M. & Deville, C. (2005). A look back at and forward to what language testers measure. In Hinkel, E. (ed.). *Handbook of research in second language teaching and learning* (pp. 815-831).Mahwah, NJ: Lawrence Erlbaum.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254-272.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cronbach, L.J. (1988). Five perspectives on validation argument. In H. Wainer and H. Braun (eds.) *Test validity* (pp. 3-17). Hillsdale, NJ: L.Erlbaum.

Kane, M., Crooks, T., Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18* (2), 5-17

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38* (4), 319-342

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18* (4), 351-372

McNamara, T. (1996). *Measuring second language performance*. London: Longman

McNamara, T. (1997). “Interaction” in second language performance assessment: Whose performance? *Applied Linguistics, 18* (4), 446-466.

McNamara, T. & Roever, C. (2006). *Language Testing: Social dimension*. Oxford: Blackwell.

McNamara, T. (2007). Language assessment in foreign language education: the struggle over constructs. *The Modern Language Journal, 91* (2), 280-282.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validity of performance assessment. *Educational Researcher, 23*: 2, 13-23

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 2*3, 5-12.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405-450.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests.* London: Longman