Commentaries on Validity Issues in Foreign and Second Language Assessment

Fred S. Tsutagawa and Yuna Seong
Teachers College, Columbia University

Introduction

In empirical applied linguistics research, the primary goal and concern is to operationalize key variables (i.e., measured constructs) in a valid and reliable way, generate scores for the measured variables through quantitative and/or qualitative means (e.g., various kinds of pre- or posttests, surveys, or coded observations), treat those scores appropriately, and allow for proper hypothesis testing of the research questions under investigation (Purpura, Brown, & Schoonen, 2015, p. 37). If the consequences of the research are "low stakes" in that the participants in the study are generally not directly impacted by the results (i.e., decisions are not made on the results to either advance or demote them in some way), the research can be published, our knowledge and understanding of the phenomenon in question deepened, and the story can essentially end there. But if there are important "high stakes" decisions to be made about the participants based on the results, decisions that can potentially impact their lives directly, it becomes imperative that our procedures and theoretical constructs have been thoroughly examined and are valid. That is why in the subfield of second and foreign language assessment, where high stakes decisions such as university admission or classification as an English language learner (ELL) in the U.S. K-12 public school system do take place based on the various test results, a higher standard needs to be adhered to in the development and implementation of the test instruments, potential interpretations of the results, and any possible subsequent uses of the results. Consequently, in second and foreign language testing, *validation frameworks* have been thoroughly developed and discussed to ensure that best measurement practices and high professional standards are followed (American Educational Research Association [AERA], American Psychological Association [APA], and the National Council on Measurement in Education [NCME], 1985, 2014), and that is why second/foreign language testers subject test scores to rigorous validity evaluation so that claims made about the measured constructs can be deemed meaningful and appropriate for their intended purpose(s), and their intended use and interpretation in decision making can also be justified (Purpura et al., 2015).

Two main schools of thought have developed regarding validation theory that emphasize different priorities. The traditional way of viewing validation has been from a more factual and practical orientation and it primarily focuses on properties of the test itself (Cronbach, 1971; Cronbach & Meehl, 1955; Kane, 1992, 2006, 2012; Messick, 1986, 1989). It views validation in terms of measured variables and constructs and "how a validation framework can be useful in helping researchers think about the instruments they use and the assumptions implicit in the scores they generate" (Purpura, Brown, & Schoonen, 2015, p. 40). The other school of thought is more philosophical in nature and emphasizes the potential social, ethical, and justified usages of tests, mainly for high-stakes situations, discussing important issues of test fairness and justice (Kunnan, 2000, 2005, 2014; McNamara & Roever, 2006; Shohamy, 2007).

But perhaps the most influential validation framework has been Kane's (1992, 2006, 2012) interpretive argument framework for validation that consists of two types of arguments, interpretive and validity arguments. Beginning with a validation argument that explicitly lays out a "network of inferences and assumptions leading from the test performances to the conclusions to be drawn and to any decisions based on these conclusions (Crooks, Kane, & Cohen, 1996; Kane, 1992; Shepard, 1993)" (Kane, 2006, p. 22) so that the same framework can be consistently applied across many different kinds of applications, he later proposed a more thought-out interpretive argument framework (Kane, 2006, 2012). In its most current form (and after some revision by others), the chain of inferences is comprised of: 1) Domain Description, 2) Scoring/Evaluation, 3) Generalization, 4) Explanation, 5) Extrapolation, and 6) Utilization (Bachman, 2005; Chapelle, Enright, & Jamieson, 2008; Kane, 2006) (see Durkis, Appendix A, in this issue for an illustrative diagram of the framework). Once all of "the chain of inferences leading from performance to claims of trait interpretation and use has been laid out along with its supporting assumptions" (Purpura, 2011, p. 739), the validity argument, then, is used to evaluate the argument as a series of claims, counterclaims, warrants, and empirical backing (i.e., through various quantitative measurement methods) (Bachman, 2005; Chapelle et al., 2008). According to Kane (2006), "validation" has two important usages. The first "involves the development of evidence to support the proposed interpretations and uses," and in the second it is "associated with an evaluation of the extent to which the proposed interpretations and uses are plausible and appropriate" (p. 17). In sum, this strand of validation research has been a major driver in the field to find appropriate, meaningful, and useful score interpretations for second and foreign language assessments.

For this issue of *Teachers College, Columbia University Working Papers in TESOL & Applied Linguists*, we invited three contributors to extensively comment on validity issues in second or foreign language assessment. Jorge Beltran looks at the English Language Arts Regents Exam, a test widely used in the K-12 system in New York. He takes an analytical approach in evaluating the exam in light of current validity frameworks in order to discuss the possible threats to validity and its impact on English language learners taking the exam. Next, Andrea Durkis looks at language assessments used as part of the citizenship process in France. While reviewing its history over the past twenty years, she examines validity issues through the lens of Kane's framework. Finally, Heidi Liu Banerjee further explores fairness, an important test quality essential to the discussion of validity and its conceptualization in relation to Kane's framework.

REFERENCES

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In

C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1-25). New York, NY: Routledge.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

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

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th Ed.) (pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kane, M. T. (2012). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing, 29*(1), 3-17.

Kunnan, A. J. (2000). Test fairness (ch. 2) (pp. 27-48). Cambridge, UK: CUP.

Kunnan, A. J. (2005). Language assessment from a wider context. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning, Vol. II)* (pp., 779-794. New York, NY: Routledge.

Kunnan, A. J. (2014). Fairness and Justice in Language Assessment. In A. J. Kunnan (Ed.), *Companion to Language Assessment, Vol.3, Ch. 66: Evaluation, methodology, and interdisciplinary themes.* Oxford, UK: Wiley-Blackwell.

Messick, S. (1986). *The once and future issues of validity: Assessing the meaning and consequences of measurement*. ETS Research Report (No. RR-86-30). Princeton, NJ: Educational Testing Service.

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

McNamara, T., & Roever, C. (2006). Validity and the social dimension of language testing (Ch. 2). In *language Learning, 56(Supplement 2),* 9-42.

Purpura, J. E. (2011). Quantitative research methods in assessment and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning, Vol. II)* (pp., 730-751. New York, NY: Routledge.

Purpura, J. E., Brown, J. D., Schoonen, R. (2015). Improving the Validity of Quantitative Measures in Applied Linguistics Research. *Language Learning, 65(Supplement 1)*, 36-73.

Shohamy, E. (2007). Tests as power tools: Looking back, looking forward. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 141-152). Ottawa, Canada: University of Ottawa Press.

COMMENTARIES