Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, Vol. 16, No. 1, pp. 54-59
The Forum

Test Fairness in Second Language Assessment

Heidi Liu Banerjee
Teachers College, Columbia University

ABSTRACT

Fairness, an essential quality of a test, has been broadly defined as equitable treatment of all test-takers during the testing process, absence of measurement bias, equitable access to the constructs being measured, and justifiable validity of test score interpretation for the intended purpose(s) (AREA, APA, & NCME, 2014). Given that test fairness is closely related to the interpretations and uses of test scores as well as the claims made from those interpretations and uses, it is critical to obtain and weigh validity evidence to support or refute the score interpretations, their uses, and the potential socio-political consequences in order to evaluate fairness (Chalhoub-Deville, 2015; Haertel & Herman, 2005; McNamara & Roever, 2006). The purpose of this article is to describe how test fairness has been conceptualized in second language assessment through the lens of validity theories. First, I will describe construct- and interpretive-argument-based validity theories and how they accommodate the integration of test fairness. Then, following Xi (2010), three major approaches to conceptualizing test fairness in relation to validity will be discussed. As observed by Xi (2010), all three major approaches share a common caveat in that they do not provide concrete steps to prioritize evidence in fairness investigations. In an attempt to build a more comprehensive fairness argument that allows for systematic investigation of test fairness, Xi (2010) proposes a new approach to conceptualizing fairness within a validity framework. Her contribution to the understanding of fairness issues in language testing will be presented as part of the conclusion of this article.

Theories of Test Validation and Test Fairness

Validity, in a general sense, has been defined as "the degree to which evidence and theory support the interpretations of test scores by proposed uses of tests" (AERA, APA, & NCME, 2014, p. 11), and it has been acknowledged as the core of test development and evaluation. Without the establishment of validity, it would be challenging to justify any interpretation of the test scores and the decisions made from the test (Bachman & Palmer, 2010; Chapelle, 2011; Purpura, 2011; Purpura, Brown, & Schoonen, 2015).

As pointed out by Xi (2008) and Chalhoub-Deville (2015), the development of validity theories in language testing, in general, has parallel paths with that of educational measurement (Cronbach & Meehl, 1955; Cureton, 1951; Kane, 1992, 2006; Messick, 1989). Arguably, the most influential validity theory in language testing, until recently, has been Messick's unitary notion of validity. According to Messick, construct validity reflects the extent to which test-takers' performance varies because of the construct being measured and nothing else. In order to provide a validation framework that positions construct validity as the core, Messick proposes a progressive matrix in which two sets of distinctions are made to describe the nature of test

validation. The first set of distinctions is between test score interpretation and test use, with interpretation focusing on the evidence internal to the construct being measured, and use focusing on the relevance between the utility of the test and the construct. The second set of distinctions is between evidence and consequences, with the former (i.e., evidence) referring to the range of sources that can or should be used to justify the construct being measured, and the latter (i.e., consequences) referring to the impact the test use has on individuals as well as society. Messick's progressive matrix provides test validation researchers with a rigorous framework to refer to, so that various aspects related to the construct can be taken into consideration during the validation process.

While Messick's progressive matrix has been credited for "making explicit the role of consequences in validity theory" (Chalhoub-Deville, 2015, p. 7), the view on evidence and consequences of language assessment practices is restricted in that it only focuses on the technical, or psychometric, aspects of validity, mirroring the position of the *1999 Standards* (AERA, APA, & NCME, 1999; as critiqued in McNamara & Roever, 2006). In other words, test fairness in Messick's validity theory is established through the claims made from empirical evidence in terms of test-takers' performance, echoing the individualistic and cognitive orientation to validity in educational psychology.

One of the main criticisms of Messick's unitary notion of validity is its questionable effectiveness and practicality, because "it does not provide a place to start, guidance on how to proceed, or criteria for gauging progress and deciding when to stop" (Kane, 2012, p. 8). Seeing the limitations of construct validity for addressing practical issues, Kane (1992, 2006) proposes an interpretive-argument (IA) approach to test validation, emphasizing "the logic, evidence, and rhetoric of arguments for the validity of an assessment" (Cumming, 2013, p. 3). Later, Kane (2013) expands his IA validation framework by incorporating a broadened role of test use and consequences to form an interpretation/use argument (IUA) approach to test validation. Through a logical chain of inferences—domain description, evaluation, generalization, explanation, extrapolation, and utilization—the IUA validation framework allows language testers and researchers to logically prioritize different types of evidence, evaluate the strength of a validity argument, and measure the overall validation outcomes. Kane's approach provides a transparent process for test developers to systematically locate "potential threats to the assumptions and the inferences" (Xi, 2008, pp. 180–181) and subsequently make adjustments to ensure the overall validity of a test.

In Kane's (2013) view, consequences are integral to validity theory, and both the interpretation and the use of test scores need to be examined when evaluating the consequences of a test. While test fairness is not explicitly discussed in Kane's validation framework, the last inference in the IUA framework, *utilization*, hints at Kane's view on the issue. A utilization argument supports the inference that the use of test scores is appropriate and that scores provided to test users are useful and meaningful. Evidence gathered for the utilization inference is usually dependent on score reports, judgments of stakeholders, and decision-making processes of institutes (e.g., Sawaki & Xi, 2005; Stansfield & Hewitt, 2005). Commenting on Kane's inclusion of test uses and consequences as part of the validation process, Brennan (2013) contends that "any use of test scores" should "fall within the purview of validation" (p. 80), so that the responsibility of test uses can be taken into account.

Kane's IUA validation framework provides a rigorous, and yet flexible chain of inferences. It allows researchers and practitioners to apply specific inferences to their contexts of interest but at the same time follow a concrete sequence of steps (i.e., domain definition,

evaluation, generalization, explanation, extrapolation, and utilization). The fact that the validation framework needs to be established through both the *interpretation* and the *use* of test scores allows us to believe that the concept of fairness can be, and should be, embedded in all of the inferences. However, some researchers seem to claim otherwise. For example, Chalhoub-Deville (2015) argues that "Kane does not address how research needs to be framed to accommodate issues beyond score interpretability, construct-related consequences and individual scores" (p. 9), a view also shared by McNamara (2008). Perhaps not explicitly addressing test fairness in either the construct-based or the interpretation/use-argument-based validity theory has opened up room for discussion in terms of where fairness fits in the validation framework, and how researchers should go about it. Xi (2010) notes that fairness has been conceptualized in a number of different ways, and outlines three major approaches to conceptualizing test fairness in language testing by differentiating their perceptions of how fairness relates to validity.

Conceptualizations of Test Fairness

Test fairness, in the context of a validity argument, has been conceptualized in three major ways. The first type of conceptualization views fairness as an independent facet of test quality. The representatives of this approach, as Xi (2010) observes, are the Code of Fair Testing Practices in Education (1988; hereafter referred to as the *1988 Code*) and the ETS Standards for Quality and Fairness by Educational Testing Service (ETS, 2002, 2014).

The *1988 Code*, largely informed by the *1985 Standards* (AERA, APA, & NCME, 1985), states that test developers and users share joint responsibilities in ensuring fairness in assessment practices in test development and selection, test administration and scoring, and score interpretation and reporting. It also maintains that test-takers should be well-informed in the test-taking process, and fairness should always be a goal to strive for. As can be seen, the *1988 Code* treats fairness "as a test quality that permeates the whole assessment process" (Xi, 2010, p. 149). On the other hand, the ETS Standards for Quality and Fairness is largely influenced by the *1999 Standards* (AERA, APA, & NCME, 1999). According to ETS, fairness is established through minimizing "construct-irrelevant personal characteristics [that] have no appreciable effect on test results or their interpretation" (ETS, 2002, p. 17). In order to address test fairness, ETS provides a comprehensive list of fairness standards for assessment products and services to abide by, including fairness in the design, development, and administration processes, and fairness in the language and content of test materials. While the ETS Standards for Quality and Fairness provides some ground rules to ensure test fairness, it has its caveats in that it does not provide a systematic way to prioritize the standards or to weigh one piece of fairness evidence against another (Xi, 2010).

The second type of conceptualization perceives fairness as an overarching test quality. That is, researchers holding this view believe that fairness "subsumes and goes beyond validity" (Xi, 2010, p. 150). The most known representative of this view is Kunnan (2000, 2004), who, drawing from the *1988 Code*, the *1999 Standards*, and Willingham and Cole's (1997) notion of comparable validity, proposes a framework of test fairness in language assessment. According to Kunnan, fairness is a test quality that encompasses validity, absence of bias, accessibility to the test, conditions of administration, and social consequences. While Kunnan's test fairness framework has contributed to a broadened understanding of the expanded scope of fairness (e.g., McNamara & Roever, 2006), Xi (2010) argues that Kunnan's proposed qualities of fairness have all been addressed coherently in previously established validity frameworks (Bachman, 2005;

Kane, 1992, 2006, 2013; Messick, 1989). Therefore, she does not see the necessity to separate fairness, validity, and other related facets. In addition, similar to the limitation of how the *1988 Code* and the ETS Standards address fairness, Kunnan's framework does not offer a systematic guideline for prioritizing the qualities of fairness in validation research.

The third type of conceptualization views fairness as a fundamental test quality, and links fairness directly to validity. Such a view is represented by the *1999 Standards*, Willingham and Cole (1997), and Willingham (1999). As one of the most influential references in educational and psychological testing, the *1999 Standards* not only elaborates on the fairness issues related to testing procedures as well as test-taker rights and responsibilities, but also advocates "the gathering of multiple types of evidence to support test fairness" (Xi, 2010, p. 152), including evidence related to content validity, construct validity, and concurrent validity.

Willingham and Cole (1997) and Willingham (1999) view test fairness as an important aspect of validity, arguing that fairness should be viewed as "comparability in assessment; more specifically, comparable validity for all individuals and groups" (p. 7). To elaborate, Willingham and Cole's conceptualization of fairness can be characterized as: comparable treatment in the testing process (including test interpretation and use), comparable learning opportunities and outcomes of learning, and comparable test material selection (Kunnan, 2000; Xi, 2010). They propose that fairness issues should be addressed throughout the entire assessment process, from design to development, administration, and use.

While the *1999 Standards* and Willingham and Cole's work both have had a tremendous impact on fairness in assessment practices, they again fall short as a framework for investigating fairness and validity systematically. Seeing that all prevailing test fairness frameworks cannot sufficiently account for prioritizing and weighing the evidence of fairness for a systematic investigation, Xi (2010) proposes that an argument-based approach to validation may overcome the shared limitation.

In Xi's (2010) conceptualization, fairness is treated as "an aspect of validity" and is conceptualized as "comparable validity for all *relevant* groups" (p. 147; emphasis in original). She argues that in order to systematically address fairness issues in different stages of the assessment process, the fairness argument should be built within the IUA-based validation framework (Kane, 1992, 2006). To illustrate, she outlines the major fairness concerns in each of the inferences (i.e., domain description, evaluation, generalization, explanation, extrapolation, and utilization), and through warrants and rebuttals, she describes the counter-arguments that may challenge or weaken the comparable validity essential to fairness. For instance, a rebuttal that could weaken the fairness argument in the evaluation (i.e., scoring) inference would be group differences in item or test scores caused by construct-irrelevant knowledge, skills, or abilities measured by the test items. Xi maintains that, in order to articulate a coherent fairness argument, a series of rebuttals related to each inference must be specified, and only when there is evidence to refute or reduce the rebuttals can fairness be properly claimed.

Conclusion

Whether fairness is conceptualized as an independent test quality, as an all-encompassing test quality, or as a quality directly linked to validity, its close relation with validity is evident. While some researchers criticize that Kane's IUA-based validation framework does not adequately integrate fairness, Xi's (2010) approach to investigating test fairness has shown otherwise. Through the example of the TOEFL iBT™, Xi shows that, by taking advantage of

Kane's (2006, 2012, 2013) well-established validation framework, test fairness can be more thoroughly investigated and clarified. Through the argument-based chain of inferences, the framework also allows for prioritizing the evidence of validity and fairness, and for tracking how fairness issues permeate the entire process of assessment practice.

REFERENCES

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

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

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

Bachman, L. F., & Palmer, A. S. (2010). *Language Assessment in Practice.* Oxford, UK: Oxford University Press.

Brennan, R. L. (2013). Commentary on "validating the interpretations and uses of test scores". *Journal of Educational Measurement, 50*(1), 74–83.

Chaloub-Deville, M. (2015). Validity theory: Reform, policies, accountability testing, and consequences. *Language Testing*, 1-20. DOI:10.1177/0265532215593312

Chapelle, C. (2011). Validation in language assessment. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning, Vol. (II)* (pp., 717–730). New York, NY: Routledge.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In Wainer, H. & Braun, H. I. (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

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

Cumming, A. (2013). Validation of language assessments. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics.* New York, NY: John Wiley and Sons.

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

Educational Testing Service (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Educational Testing Service (2014). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Haertel, E., & Herman, J. (2005). *Historical perspective on validity arguments for accountability testing, CSE Report 654*. Los Angeles, CA: University of California, Los Angeles.

Joint Committee on Testing Practices (1988). *Code of fair testing practices in education*. Washington, DC: Author.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–35.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–

64). Washington, DC: National Council on Measurement in Education and the American Council on Education.

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

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–14). Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In Milanovic, M. & Weir C., (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference* (pp. 27–48). Cambridge, UK: Cambridge University Press.

McNamara, T. (2008). The social-political and power dimensions of tests. In E. Shohamy and N.H. Hornberger (Eds.), *Encyclopedia of Language and Education, Vol. 7: Language testing and assessment* (2nd ed., pp. 415–427). Dordrecht, The Netherlands: Springer.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.

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.

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.

Sawaki, Y., & Xi, X. (2005). *Standard setting for the next generation TOEFL*. Paper presented at the 2005 TESOL Convention, San Antonio, Texas.

Stansfield, C.W., & Hewitt, W.E. (2005). Examining the predictive validity of a screening test for court interpreters. *Language Testing, 22*(4), 438–462.

Willingham, W. W. (1999). A systemic view of test fairness. In Messick S. (Ed.), *Assessment in higher education: Issues in access, quality, student development, and public policy* (pp. 213–242). Mahwah, NJ: Lawrence Erlbaum.

Willingham, W. W. & Cole, N. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum.

Xi, X. (2008). Methods of Test Validation. In Shohamy, E., and Hornberger, N. (Eds.) *Encyclopedia of Language and Education, 2nd Edition, Volume 7: Language Testing and Assessment* (pp. 177–196). New York, NY: Springer.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*, 147–170.

Heidi Liu Banerjee is a doctoral student in the TESOL & Applied Linguistics program at Teachers College, Columbia University. Her research focuses on second language assessment, with specific interests in game-/scenario-based assessment, learning-oriented assessment, assessing integrated skills, learner cognition, and psychometric measurement. Correspondence should be sent to Heidi Han-Ting Liu, Teachers College, Columbia University, Box 66, 525 West 120th Street, New York, NY, 10027. Email: hl2641@tc.columbia.edu