Issues of Validity in the Assessment of Writing Performance

Constance Hui Ling Tsai
Teachers College, Columbia University

According to Weigle (2002), any writing test that involves actual writing, as opposed to completing selected response or limited production items, can be considered performance assessment. McNamara (1996) proposed distinguishing between a strong sense and a weak sense of performance assessment. In a second language (L2) writing test in the strong sense, the test task will represent a real-world task like making an official request, and performance will primarily be judged on real-world criteria. The focus of performance assessment is on the successful fulfillment of the task, and not on the successful use of language in performing the writing task. The L2 is only a medium of the performance and an insufficient condition for success. In fact, if aspects of L2 writing ability are stressed at all, criteria reflecting L2 writing ability will only be part of a larger set of criteria used. Performance of the task itself is the target of the assessment (Messick, 1994). In the weak sense of performance assessment, the focus of the assessment is on the language used. Although the task used to elicit writing may resemble real-world tasks, the purpose is to elicit a display of writing ability. McNamara (1996) suggested that most language performance tests are weak in this sense. The distinction between the strong and weak form of performance assessment is an important conceptual consideration. For example, as an L2 writing test in the weak sense is designed to elicit a performance of the test-taker's L2 writing ability, the scoring criteria should clearly articulate the definitions of the construct of L2 writing ability, and raters should be trained to interpret these criteria in language-related terms. Otherwise, test scores may reflect construct-irrelevant variability (like creativity and neatness), and we would not be able to relate inferences from the performance to inferences related to performance in any non-test situation.

Validity, or judgments of the relevance of test tasks to the intended score representation, is thus at the heart of the performance test (Kenyon, 1998). Messick (1989) describes 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" (p. 13). To ensure the relevance and validity of a test, test designers should take into account all aspects of the testing situation that significantly affect test performance, including the specification of the construct domain in terms of topical knowledge, test specifications, administration conditions, and criteria for scoring. Such specifications indicate what aspects of the test procedure are likely to influence test scores and should be controlled (Messick, 1994).

According to Weigle (2002), one useful way to conceptualize the specification of test constructs, or construct definition, might be to link it to the three stages in Bachman and Palmer's (1996) framework of test development: the design stage, the operationalization stage, and the administration stage. To ensure validity of the test at the design stage, McNamara (1996) suggests sampling test content from the communicative tasks facing the test-takers in the target language use situation to ensure content validity. The steps recommended include consulting with expert informants, examining available literature on the communicative demands of the target language setting, analyzing and categorizing communicative tasks in the target language

setting, collecting and examining texts from the target language setting, and deciding on a broad test method (e.g., developing test specifications and scoring procedures, and writing materials).

In the operationalization stage, information from the design stage is used to create test specifications or detailed procedures for test writers to follow. According to Douglas (2000), test specifications should contain a description of the test content, the criteria for correctness (i.e., the scoring rubric), and sample tasks or items. To ensure validity at this stage, the task must elicit the components of writing that are included in the definition of the construct we want to test – no more and no less. To do the former would result in what Messick (1989) calls construct-irrelevant variance, and to do the latter would lead to underrepresentation of the construct. For an L2 writing task to elicit the components that accurately define the construct, the writing prompt should be sufficiently clear and specific so as to limit possible interpretations (Horowitz, 1991); if test-takers are given a choice of prompts, they should be as similar to one another as possible (Hoetker & Brossell, 1986), so that differences in task demands will not introduce a source of variability in test-takers' scores. In addition, scoring rubrics that do not accurately reflect the construct being measured may impact the validity of inferences made on the basis of test results (McNamara, 1996). Rubrics should thus contain an explicit and unambiguous statement of the components of the construct that are being assessed. All things being equal, rubrics with a greater number of subscales, like analytic rubrics, are generally seen as tending to lead to greater overall consistency of scoring (Brown & Bailey, 1984; Hamp-Lyons, 1991).

Construct validation at these two test development stages – design and operationalization – roughly corresponds to what Weir (1988, as cited in McNamara, 1990) referred to as *a priori* construct validation, in which test content is based upon an explicit theory of language and language use. *A posteriori* construct validation, in contrast, relates to the empirical and statistical validation of the constructs posited at the earlier stages using test performance data. This second type of construct validation corresponds with the third and final stage in Bachman and Palmer's (1996) test development process, the administration stage. At this stage, we are interested in obtaining evidence that the test scores indeed reflect the components of the construct embodied in the design of the test by identifying and accounting for construct-irrelevant facets in the test situation.

Performance assessment thus imposes its unique set of challenges. However, in the context of L2 writing assessment, it is usually preferred over indirect methods of assessment. This is because performance tests require a test-taker to demonstrate not only language knowledge, but also skill in the use of that knowledge by actually performing it in communicative situations (Shohamy, 1983). It thus has greater predictive validity; it allows inferences to be made about test-takers' future performance in real-world contexts based on test performance. This makes it a much more powerful form of assessment than tests that sample aspects of language knowledge discretely and provide no information about actual performance, either in a specific test situation or in similar situations in the future or the real world.

REFERENCES

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. New York: Oxford University Press.

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*, 21-42.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-278). Norwood, NJ: Ablex.

Hoetker, J., & Brossell, G. (1986). A procedure for writing content-fair essay examination topics for large-scale writing assignments. *College Composition and Communication, 37*, 328-335.

Horowitz, D. (1991). ESL writing assessments: Contradictions and resolutions. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 71-85). Norwood, NJ: Ablex.

Kenyon, D. (1998). Approaches to validation in language assessment. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 1-16). Mahwah, NJ: Erlbaum.

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*, 52-75.

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

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*, 13-23.

Shohamy, E. (1983). The stability of oral proficiency assessment on the oral interview testing procedures. *Language Learning 33*, 527-540.

Weigle, S. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.

Constance Tsai received her Ed.D. in TESOL from Teachers College, Columbia University. Her research interests include the assessment of ESL writing and strategy use, rater behavior in essay rating, and TESOL teacher education. She is currently an adjunct lecturer of Supervised Student Teaching (Secondary level) in the TESOL masters program at Teachers College, a mentor for M.A. students (Returned Peace Corps Fellows) at Teachers College who are teaching in New York City public schools, an ESL instructor for the State University of New York Educational Opportunity Center in Manhattan, and a consultant for Educational Testing Services in Princeton.