Issues of Validity and Reliability in Second Language Performance Assessment

Yen-Fen Liao
Teachers College, Columbia University

With an increasing practical need for language tests that can provide predictive information about how successfully a candidate will perform in a non-testing setting, second language performance assessment has recently aroused immense interest in the field of language testing. Of the many issues involved in performance assessment, validity and reliability in particular have been of great concern to language testers and educators. In this regard, it is the intent of this commentary to briefly discuss the issues of validity and reliability in the context of second language performance assessment.

It was once believed that "by following established procedures, it is possible to design a format for administering and scoring a valid and reliable language performance test" (Jones, 1979, p. 50). However, this seems to be an overly simplified view of performance testing given the complexity of validity and reliability issues in performance assessment. Current inquiry into the issues of validity and reliability in second language performance assessment represents a broader field with multiple perspectives and a wider use of sophisticated research methodologies.

First of all, validity has been identified as the most important quality of test use, which concerns the extent to which meaningful inferences can be drawn from test scores (Bachman, 1990). In order to examine the validity of a test, it requires a validation process by which a test user presents evidence to support the inferences or decisions made on the basis of test scores (Cronbach, 1971, as cited in Crocker & Algina, 1986). Validation studies of language performance assessment are mainly concerned with three types of validity: construct validity, predictive validity, and content validity.

Construct validity is associated with two distinctive approaches to performance assessment development: the construct-centered approach and the task-centered approach (Bachman, 2002). The task-centered approach has been favored over the construct-centered approach by some proponents of performance assessment. For instance, a group of researchers at the University of Hawaii at Manoa argue that performance on a task-based test itself is the construct of interest, indicating that predictions to be made are about the test-takers' abilities to accomplish certain tasks (Brown, Hudson, Norris, & Bonk, in press, as cited in Bachman, 2002). Building on this definition, task-based language performance assessment (TBLPA) is regarded as one type of performance assessment where the construct of interest is task performance itself. One potential problem with this approach, however, is that inferences may not be made beyond a specific testing context, which thus severely weakens the interpretation and generalization of test results (Bachman, 1990). Bachman (2002) therefore argued that both task-centered and construct-centered approaches should be adopted in the performance-based test design.

The other pivotal validity considerations in second language performance assessment are predictive validity and content relevance and coverage. Since the major purpose of performance tests is to provide predictive information about how well the testee will use the second language under specific target conditions, predictive validity has been one of the primary concerns in performance assessment (Wesche, 1985). How accurately a prediction can be made relies on the degree of content validity. Content validity involves two crucial concepts: content relevance and content coverage (Bachman, 1990). Content relevance refers to the extent to which the aspects of ability to be assessed are actually tested by the task, indicating the requirement to specify the ability domain and the test method facets (Bachman, 1990). Content coverage concerns the extent to which the test tasks adequately demonstrate the performance in the target context, which may be achieved by randomly selecting representative samples (Bachman, 1990). The second aspect of content validity is similar to that of content representativeness, which also concerns the extent to which the test accurately samples the behavioral domain of interest (Bachman, 2002). Some problems in investigating content validity have been identified by language testers (e.g., Bachman, 2002). First, difficulties may arise in defining the TLU domain in a situation where examinees come from diverse backgrounds and have widely ranging needs in language use. Furthermore, even when the TLU domain can be well defined, selecting representative samples from that domain may be problematic (Bachman, 2002). As pointed out by Hughes (1981), it is quite difficult to sample representative language skills as a result of inadequate needs analyses and the lack of comprehensive and complete descriptions of language use. This sampling problem may complicate and lengthen the test design (Jones, 1979). Some attempts have so far been made to identify representative samples. For instance, Branden, Depauw, and Gysen (2002) highlighted the value of needs analysis for sampling tasks in the instructional and learning contexts. However, needs analysis has been challenged in cases where testees come from various backgrounds. These challenges may pose a serious extrapolation problem beyond a specific testing context.

As claimed by Bachman (2002), "ill-defined or indeterminate relationships between assessment tasks and TLU tasks affect extrapolation"(pp. 458-459). Some empirical attempts have been made to investigate the extrapolation issue in relation to generalizations across test tasks. For example, the findings in Brindley and Slatyer's (2002) study demonstrated generalizability problems in performance assessment and indicated an urgent need for a detailed exploration of sources of variation that may affect testees' performance. The comparability of various tasks has often been questioned and thus the generalization of performance on a certain test task to the broader universe of test tasks has been called into question (e.g., Bachman, 2002).

Reliability is in fact a prerequisite to validity in performance assessment in the sense that the test must provide consistent, replicable information about candidates' language performance (Clark, 1975). That is, no test can achieve its intended purpose if the test results are unreliable. Reliability in a performance test depends on two significant variables: (1) the simulation of the test tasks, and (2) the consistency of the ratings (Jones, 1979). Four types of reliability have drawn serious attention: (1) inter-examiner reliability, (2) intra-examiner reliability, (3) inter-rater reliability, and (4) intra-rater reliability (Jones, 1979).

Since the administration of performance tests may vary in different contexts at different times, it may result in inconsistent ratings for the same examinee on different performance tests. Attention, therefore, should be devoted to inter-examiner and intra-examiner reliability, which concern consistency in eliciting test performance from the testee (Jones, 1979).

In addition, performance tests require human or mechanical raters' judgments. The reliability issue is generally more complicated when tests involve human raters because human judgments involve subjective interpretation on the part of the rater and may thus lead to disagreement (McNamara, 1996). Inter-rater and intra-rater reliability are the main considerations when investigating the issue of rater disagreement. Inter-rater reliability has to do with the consistency between two or more raters who evaluate the same test performance (Jones, 1979). For inter-rater reliability, it is of primary interest to examine if the observations over raters are consistent or not, which may be estimated through the application of generalizability (Crocker & Algina, 1986). Intra-rater reliability concerns the consistency of one rater for the same test performance at different times (Jones, 1979). Both inter- and intra-rater reliability deserve close attention in that test scores are likely to vary from rater to rater or even from the same rater (Clark, 1979). For instance, the halo effect has been recognized as a serious problem when raters are required to score all test sections of a given tape and continually shift their scoring criteria (Starr, 1962). More studies on the issues of the scoring reliability in second language performance assessment seem very much in order.

Although a performance-based testing approach has been widely challenged especially in the issues of validity and reliability, there has been a consensus that performance assessment is valuable for measuring job applicants' language proficiency in vocational situations as well as for motivating language learners to make a greater effort to develop communicative language ability (Jones, 1979). Continued attention to the issues of validity and reliability in second language performance assessment is a challenging but necessary endeavor that will advance the development and use of performance tests.

REFERENCES

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

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

Branden, K., Depauw, V., & Gysen, S. (2002). A computerized task-based test of second language Dutch for vocational training purposes. *Language testing, 19*, 438-52.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19*, 369-94.

Clark, J. (1975). Theoretical and technical considerations in oral proficiency testing. In S. Jones & B. Spolsky (Eds.), *Language testing proficiency* (pp. 10-24). Arlington, VA: Center for Applied Linguistics.

Clark, J. (1979). Direct vs. semi-direct tests of speaking ability. In E. Briere & F. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp. 35-49). Washington, DC: TESOL.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth Group/Thomson Learning.

Hughes, A. (1981). Conversational cloze as a measure of oral ability. *English Language Teaching Journal, 35*, 161-168.

Jones, R. (1979). Performance testing of second language proficiency. In E. Briere & F. Hinofotis (Eds.), *Concepts in language testing* (pp. 50-57). Washington, DC: TESOL.

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

Starr, W. (1962). MLA foreign language proficiency tests for teachers and advanced students. *PMLA, 77*, 1-12.

Wesche, M. (1985). Introduction. In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 1-12). Ottawa: University of Ottawa Press.

Yen-Fen Liao is a doctoral student in Applied Linguistics at Teachers College, Columbia University. Her primary research interest is in second language assessment, especially issues in assessing second language learners' listening ability. She is currently working on her pilot study on the construct validation of a second language listening comprehension test in preparation for her dissertation research.