Print Close

Valid, Reliable, and Appropriate Assessments for Adult English Language Learners

Dorry Kenyon, Center for Applied Linguistics
Carol Van Duzer, National Center for ESL Literacy Education
November 2003

Since 1998, federal guidelines have stated that assessment procedures to fulfill the accountability requirements of the Workforce Investment Act (WIA) must be valid, reliable, and appropriate (U.S. Department of Education, 2001). This provision is likely to remain in the reauthorization of the WIA in the coming year. The U.S. Department of Education's blueprint for this legislation continues to call for local programs to demonstrate learner gains in reading, math, language arts, and English language acquisition. It also requires states to implement content standards-defined as clear statements of what learners should know and be able to do-and to align assessments with these standards (U.S. Department of Education, 2003). The educational functioning levels, which form the basic framework and structure of the National Reporting System for Adult Education (NRS), remain. These levels may be refined; for example, test benchmarks may be linked to real-life outcomes (Keenan, 2003).

As the field of adult English as a second language (ESL) instruction moves towards content standards, program staff and state and national policy makers need to be able to make informed choices about appropriate assessments for adult English language learners. This Q&A examines the concepts of validity, reliability, and appropriateness from a language testing perspective as they apply to the following four assessment issues raised by the NRS:

1. What type of language assessment seems to be required by the NRS: proficiency or achievement? What type of assessment would be most appropriate for the NRS? What does validity entail for appropriate NRS assessment?
2. What does reliability mean for performance measures meeting the rigorous requirements of the NRS?

What type of language assessment seems to be required by the NRS: proficiency or achievement?

For adult English language learners in the United States, the basic reason for learning English is to use it. It is not to know about grammar or sophisticated details of English syntax or cultural aspects of the land where the language is spoken. All of these have their place, but knowing a language involves being able to put all of these pieces together in order to read for work or enjoyment, participate in conversations with others who speak English, or accomplish other tasks using the language.

Traditionally, achievement testing has been defined as assessing whether students have learned what they have been taught. Today, as the field of education institutes standards, assessment frameworks look not only at what students know about the language, but at what they can do with it. For adult language learners, that means using the language in everyday life. The goal of learning, then, is to develop proficiency. The American Council on the Teaching of Foreign Languages (ACTFL) defines language proficiency as "language performance in terms of the ability to use the language effectively and appropriately in real-life situations" (Buck, Byrnes, & Thompson, 1989, p. 11).

Proficiency distinguishes itself from achievement in that, when measuring language skills, proficiency is not necessarily confined to what is taught in the classroom. Language acquisition-learning new vocabulary and structures-also occurs outside the classroom as learners live, work, and interact with others in an English-speaking environment (Gass, 1997).

The NRS defines six educational functioning levels for English language learners. These levels describe what learners can actually do. For example, learners at the beginning ESL listening and speaking level can

- understand frequently used words in context and simple phrases spoken slowly with repetition, communicate basic survival needs with some help, and
- understand and participate effectively in face-to-face conversations on everyday subjects spoken at normal speed.

These aims are focused on what happens in real life outside the classroom. In language testing terms, the focus of the NRS is on proficiency.

The challenge, both for teaching and assessment, is determining the relationship among content standards, curriculum, instruction, and proficiency (versus achievement) outcomes. The field of foreign language education has been struggling for several decades to align these.Currently, only a few states have standards for adult ESL instruction. Among these are Arizona, California, Florida, Massachusetts, New York, and Washington (Florez, 2002). If content standards define what

learners can do in the real world (proficiency), then how do these standards influence what happens in the classroom, particularly how proficiency is assessed?

Adult learners come to the classroom with a variety of prior educational and life experiences. In acquiring English literacy, learners require different curricula and instructional strategies depending on whether they have ever acquired literacy in any language, have a high level of literacy in their own language, or are literate in a language that uses a Roman or a non-Roman alphabet (Burt, Peyton, & Adams, 2003). Learners also differ in their opportunities for language acquisition outside the classroom. For example, they may work in jobs where contact with native English speakers or speakers of other languages require them to use English, or they may work in jobs with very little contact with other workers, particularly English speakers. Some learners are able to attend class several times a week and others only once. A couple hours of instruction a week is a very limited amount of time for developing English language proficiency. What goes on inside the classroom needs to help learners take advantage of what goes on outside the classroom, so that learners can maximize opportunities to increase their language acquisition (Van Duzer, Moss, Burt, Peyton, & Ross-Feldman, 2003).

Classroom assessments-such as reading, writing, or speaking logs; checklists of communication tasks; and oral or written reports-can show how learners have mastered curricular content or met their own goals. (See Van Duzer & Berdan, 1999, for a list and discussion of classroom assessments.) The assessments may reflect what the learners can do in the real world. However, without specific valid and reliable links to the NRS functioning levels, these tools and processes may not meet the current requirements to show level gain.

Knowing that the NRS focuses on what learners can do in the real world and knowing the challenges to classroom teaching, what type of assessment would be most appropriate?

A good language proficiency test is made up of language tasks that replicate what goes on in the real world (Bachman & Palmer, 1996). Performance assessments-which require test takers to demonstrate their skills and knowledge in ways that closely resemble real-life situations or settings (National Research Council, 2002)-seem appropriate. A performance assessment generally has more potential than a selected response test (e.g., true-false or multiple choice) to replicate language use in the real world. That potential is realized, however, only if the assessment itself is of high technical quality, not just because it is a performance assessment.

Performance assessments are not easy to develop, administer, score, and validate, because there are many variables involved. The Performance-Based Assessment Model (see Figure 1) illustrates the many variables that apply to the development of performance-based assessments.At the base of the model is the **student** (or examinee) whose **underlying competencies** (knowledge, skills, and abilities [K/S/A]) are to be assessed. To do this, the student is given tasks to perform. Several variables surround these tasks. What is the quality of the task? Is it a good task or a poor task? Are conditions provided so that it can be successfully completed? Will the student be given enough time to do it?

Next is the **test administrator**, who may interact with the examinee. The administrator may bring his or her own underlying competencies (knowledge, skills, and abilities) into the student's performance. Does the administrator know what to ask the student to do and how to ask it?

These three elements (student, task, and administrator) interact to produce a **performance**. The performance needs to be assessed by a **rater**. Sometimes, one person may act as both the administrator and rater (e.g., in an oral interview); at other times, the administrator and the rater will be two individuals (e.g., in a writing assessment). Raters bring additional variables. Are they well trained? Do they have the knowledge base needed to rate the performance?

In order to assess the student's performance, raters need **criteria**, often contained in a scale or a rubric. The rubric needs to be useful and easy to interpret, and it must address the aspects of the performance related to the examinee's underlying competencies that are to be assessed. For example, if writing is being assessed, do the rating criteria relate to characteristics of a good writer (e.g., ability to organize the writing, ability to use appropriate mechanics)? If speaking is being assessed, do the criteria relate to competencies of a good speaker (e.g., ability to make oneself understood)?

Finally, raters use the rubric or scale to assign a score to the performance. This score has meaning only in so far as it is a valid and reliable measure of what the learner can do. In other words, do the many variables depicted in the diagram work together to produce a score that is a valid indicator of an examinee's ability? Does the performance assessment allow the examinee to give a performance that reflects proficiency in the real world, can be adequately described and measured by the rubric, and can be scored reliably? Can the assessment be repeated, both in terms of the performance being elicited and the score applied?

Knowing that all these variables need to be attended to, what does validity entail for an appropriate NRS assessment?

Messick (1989) offers a technical definition of validity: "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationale support the adequacy and the appropriateness of inferences and actions based on test scores or other modes of assessments" (p. 11).

This view adjusts the focus of validity from the test itself to include the use of the test scores. One has to ask if the test is valid for this use, in this context, for this purpose. With regards to the NRS, the main questions that need to be answered seem to be the following: How well is the performance elicited by the test aligned with the NRS descriptors? How well can

the test assess yearly progress? How indicative of program quality are the performances on the assessment?

Any assessment used for NRS purposes will be valid only if evidence can be provided that the inferences about the learners made on the basis of the test scores can be related to the NRS descriptors, i.e., what the learners can do (proficiency). The assessment must also be sensitive enough to learner gains to be able to show progress, if that is the use to which it is put. In addition, if the quality of programs is to be judged by performances on the assessment, then it must be demonstrated that there is a relationship between the two.

Establishing validity for a particular use of a test is not a one-activity task or study. It is an accumulation of evidences that support the use of that test. It includes such things as examining the relationship between performance on the test and performance on similar assessments, examining test performances vis-ˆ-vis criteria inherent in the NRS descriptors, and examining the reasonableness and consequences of decisions made on the basis of test scores. Each of these examinations requires the collection and analysis of evidence (data).

What does reliability mean for performance assessments meeting the rigorous requirements of the NRS?

In the field of assessment, the concept of reliability is related to the consistency of the measurement when the testing procedure is repeated on a population of individuals or groups (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). For example, if a learner takes a test once, then takes it again an hour later and maybe another hour after that, the learner should get about the same score each time, provided nothing else has changed.

As the diagram indicates, a performance assessment has a number of potential sources for inconsistency. These include the assessment task itself, the administrator, the rater, the procedure, the conditions under which it is administered, or even the examinee. For example, an examinee might be feeling great the day of the pre-test but facing a family crisis on the day of the post-test.

The job of assessment developers is to demonstrate that reliability can be achieved even for a complex performance assessment. Accordingly, program staff using the test have a responsibility as well. They have an obligation to administer the assessment in the ways they have been trained to administer it, thus replicating the conditions under which reliability can be attained (American Educational Research Association et al., 1999). Programs need to plan for time to train individuals to administer the test, time to administer it, and time to monitor its administration. This may mean an additional expenditure of resources and time for staff training so that the test will be administered appropriately each time it is used. Finally, before post-testing, programs must ensure that enough time (or hours of instruction) has passed for learners to show gains.

Conclusion

Ensuring that language tests for adult English language learners are appropriate, valid, and reliable is a challenge. Performance-based assessments are inherently complex to develop and implement. Yet, because the focus of assessment-both in the NRS descriptors and in the Department of Education's definition of content standards-is on what learners can do with the language, performance assessments are worth developing and validating.

Meanwhile, as program staff choose assessments that meet current accountability requirements, they can take the following steps to ensure that valid, reliable, and appropriate assessments are chosen for their learners:

- Review the assessment and technical information provided by the test developer to determine that what the assessment purports to measure reflects real-life tasks. Review the technical manual to ascertain that the test developers have demonstrated that reliability can be achieved. Provide adequate resources to train test administrators and raters to maintain reliability of test administration and scoring.
- Post-test only after an adequate amount of instructional time has taken place to demonstrate level gain. Presently, assessment of learner gains is based on the NRS descriptors. Over the next few years, content standards will be implemented as well. If we cannot assess learners' performances in light of these standards in valid, reliable, and appropriate ways, the standards will have no practical value.

References

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

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

Buck, K., Byrnes, H., & Thompson, I. (Eds.). (1989). *The ACTFL oral proficiency interview tester training manual*. Yonkers, NY: American Council on the Teaching of Foreign Languages.

Burt, M., Peyton, J. K., & Adams, R. (2003). *Reading and adult English language learners: A review of the research*. Washington DC: Center for Applied Linguistics.

Florez, M. (2002). *An annotated bibliography of content standards for adult ESL*. Washington, DC: National Center for ESL Literacy

Education.

Gass, S. M. (1997). *Input, interaction, and the second language learner*. Mahwah, NJ: Erlbaum.

Keenan, C. (2003, May).*What is our vision?* Presentation at the National Center for ESL Literacy Education symposium on Assessment and Accountability in Programs for Adult English Language Learners. Washington, DC.

Messick, S. (1989). Validity. In R. Linn, (Ed.), *Educational measurement* (3rd ed., pp. 11-103). New York: Macmillan.

National Research Council. (2002). *Performance assessments for adult education: Exploring the measurement issues: Report of a workshop*. (R. J. Mislevy & K. T. Knowles, Eds.). Washington, DC: National Academy Press.

U.S. Department of Education, Office of Vocational and Adult Education, Division of Adult Education and Literacy. (2001, March). *Measures and methods for the National Reporting System for Adult Education: Implementation guidelines*. Washington, DC: Author.

U.S. Department of Education, Office of Vocational and Adult Education. (2003, June). *A blueprint for preparing America's future. The Adult Basic and Literacy Education Act of 2003: Summary of major provisions*. Washington, DC: Author. Available at www.ed.gov/policy/adulted/leg/aeblueprint2.doc

Van Duzer, C. H., & Berdan, R. (1999, December). Perspectives on assessment in adult ESOL instruction. *The Annual Review of Adult Learning and Literacy*, 1. Retrieved from http://gseweb.harvard.edu/~ncsall/ann_rev/index.html

Van Duzer, C., Moss, D., Burt, M., Peyton, J. K., & Ross-Feldman, L. (2003). *OECD review of adult ESL education in the United States: Background report*. Washington, DC: National Center for ESL Literacy Education & Center for Applied Linguistics.

This document was produced at the Center for Applied Linguistics (4646 40th Street, NW, Washington, DC 20016 202-362-0700) with funding from the U.S. Department of Education (ED), Office of Vocational and Adult Education (OVAE), under Contract No. ED-99-CO-0008. The opinions expressed in this report do not necessarily reflect the positions or policies of ED. This document is in the public domain and may be reproduced without permission.