

Dorry Kenyon, Center for Applied Linguistics
Carol Van Duzer, Center for Adult English Language Acquisition
Sarah Young, Center for Applied Linguistics & Center for Adult English
Language Acquisition

PERFORMANCE-BASED ASSESSMENTS IN THE ADULT ESOL WORLD: MEETING STANDARDS AND ACCOUNTABILITY REQUIREMENTS

The primary objective of adult ESOL (English to Speakers of Other Language) programs is to enable adult learners who are not fully fluent and literate in English to become proficient, so that they can meet personal, community, academic, and employment goals. The adult ESOL field is turning to content standards to provide instructional guidelines and learning outcomes to match these goals. The U.S. Department of Education put forth strong recommendations for states to develop and implement content standards that align with appropriate assessments. Selecting an appropriate assessment means considering such factors as reliability, construct validity, authenticity, interactiveness, impact, and practicality (Bachman & Palmer, 1996). These criteria are also reflected in the National Reporting System (NRS) guidelines for evaluating assessments (NRS, 2005). This article looks at the characteristics of high quality adult ESL assessments used for NRS purposes and examines two such assessments, the BEST Plus Oral English Proficiency Test and the REEP Writing Assessment.

Introduction

Selecting a valid, reliable, and appropriate assessment for adult English language learners (ELLs) requires careful consideration of a variety of criteria. What language construct does the assessment measure, and how is the construct defined? Does the assessment measure language proficiency, defined

as “the ability to use the language effectively and appropriately in real-life situations” (Buck, Byrnes, & Thompson, 1989, p. 11), or achievement, which is more closely tied to classroom instruction and does not necessarily reflect the examinee’s potential? What is the format of the test items and types of tasks, and do they support the test construct? Does the test assess authentic skills that are applicable to other contexts?

What research studies support the test's claims for reliability and validity? Can the test be used for accountability, placement, or diagnostics? How can the test inform classroom instruction? How much does the assessment cost, in terms of training, materials, personnel, time, scoring, and technology needs? An assortment of assessments are available to adult English language programs nationwide, measuring everything from speaking, listening, reading, and writing, to grammar, vocabulary, life skills knowledge, and functional literacy – or some combination thereof. How is the best assessment selected for a given adult ESL (English as a Second Language) or ENL (English as a New Language) program's needs?

According to Florida's Department of Education, approximately 126,000 adult students were enrolled in ESL courses in Florida in 2003-2004, or about 35% of the adult education population (http://www.firn.edu/doe/workforce/pdf/booklet_at_a_glance.pdf). ESL and other adult education programs that receive federal funding, channeled through states, are held to the accountability requirements of the Workforce Investment Act (WIA) of 1998 (Mislevy & Knowles, 2002) and represented by the National Reporting System (NRS). The accountability systems in place require the use of approved standardized assessment instruments. Since 1998, federal guidelines have stated that assessment procedures that fulfill the accountability requirements of the WIA must be valid, reliable, and appropriate (U.S. Department of Education, 2001). Content standard, defined as clear statements of what learners should know and be able to do, must be developed and implemented. Federal guidelines recommend that assessments be aligned 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), are still in place, although proposed changes to the Beginning ESL and Advanced ESL functioning levels may affect NRS reporting and assessment scores in the future (see www.nrsweb.org for more information).

As the field of adult English as a second language (ESL) or English for Speakers of Other Languages (ESOL) instruction move 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 article 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:

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

The article concludes with a look at two adult ESL assessments used for NRS reporting purposes, the *BEST Plus Oral Proficiency Test* and the *REEP Writing Assessment* (RWA).

Proficiency vs. Achievement?

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 for communicative competence. 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 skills have their place, but knowing a language involves being able to put all of these pieces together 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 in their daily lives. Therefore, for adult language learners, the goal of learning is to develop proficiency. Proficiency distinguishes itself from achievement; when measuring language skills, proficiency is not

necessarily confined to what is taught in the classroom. Language acquisition, or 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. 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 when proficiency is being 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.

Appropriate Assessment According to the NRS

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, or validate, because many variables are 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, authentic task, or a poor task? Are conditions provided so that it can be successfully completed? Will the student be given enough time to complete the task? The next concern 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 that it be done?

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?

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, and 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?

Validity

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 regard to the NRS, the main questions that need to be answered seem to be the following: How well does the performance elicited by the test align with the NRS descriptors? How well can the test assess yearly progress? Are the performances on the assessment indicative of program quality?

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, that is, what the learners can do, given their level of 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 evidence 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-à-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 through raw data.

Reliability

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 later, the learner should get about the same score each time, provided nothing else has changed.

As the diagram in figure 1 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 might be 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, 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 proper 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.

Measuring Adult Performance Assessments

BEST Plus Oral English Proficiency Test and REEP Writing Assessment (RWA)

To illustrate the above discussions of reliability, validity, and appropriateness in assessments in practical terms,, descriptions of the *BEST Plus Oral*

English Proficiency Test and REEP Writing Assessment (RWA) follow. Both of these assessments can be used for placement, assessment of student progress, diagnosis, and program evaluation. In addition, the scores from both assessments are correlated to Student Performance Levels and NRS ESL educational functioning level descriptors. As such, they are two of only a few oral proficiency and writing assessments accepted for reporting level gains in the NRS. *BEST Plus* and *RWA* feature high reliability, validity, and applicability to real-world contexts, as required by NRS and language testing criteria for performance assessments.

Real world use

BEST Plus consists of an individually administered face-to-face scripted oral interview. The trained test administrator scores examinees' responses based on a standardized rubric that evaluates listening comprehension, (how well the examinee understood the setup and question), language complexity, (how the examinee organized and elaborated the response), and communication, (how clearly the examinee communicated meaning). The computer-adaptive software selects items of varying degrees of difficulty for examinees, based on their performance on the previous items.

As an oral proficiency test, *BEST Plus* is not linked to any particular curriculum or textbook. Rather, it assesses the ability to understand and use unrehearsed and conversational language within topic areas generally covered in adult ESL courses. Examinees are administered questions drawn from several "folders" of thematically-related questions, dealing with topics like health, family / parenting, consumerism, getting a job, community services, weather, and education. Each thematic folder contains different question types that allow examinees to demonstrate their full range of proficiency.

The *RWA* is a performance-based writing test that uses the real-world skill of writing a letter to a close friend or family member on a familiar topic, such as living in the United States, visiting the home country, or moving to a new city. Examinees begin with standardized warm-up tasks conducted involving authentic brainstorming activities and pair discussion

focused on the writing topic. Students are then scored using the REEP Writing Rubric, which analyzes the examinee's performance on content and vocabulary (comprehensive and comprehensible information), organization and development (paragraph writing and linked ideas), structure (grammar, syntax, verb tense), mechanics (punctuation, capitalization), and voice (personal style, engaging). Test administrators undergo rigorous training on standardizing the test administration, scoring, and annual recertification. Programs that use the *RWA* can share the scoring rubric with students. This metacognitive strategy helps engage students in the process of evaluating writing and improves their understanding of their own writing (Grant, 2005).

Reliability of BEST and RWA

Both *BEST Plus* and *RWA* were developed, standardized, and validated to ensure the highest technical and content quality. Reliability studies show high interrater reliability, high test/retest reliability, and high reliability for equivalent forms for both assessments (Center for Applied Linguistics, 2005; Arlington Education and Employment Program, 2005). Research studies on *RWA* and *BEST Plus* show that level gains in writing and oral proficiency, respectively can reliably be shown within 120-180 hours of instruction for *RWA* and 80-100 for *BEST Plus*.

Validity of BEST Plus and RWA

Both *BEST Plus* and *RWA* have a history of adult ESOL teacher input and work in the development and validation of the assessments, and as such, have strong face validity. The content and topics of these assessments, as well as the difficulty level of each item and writing prompt, reflect those concepts and domains that are covered in adult ESOL courses around the country. These assessments have been correlated to SPLs and NRS functioning level descriptors through standard setting activities and input from the field. In addition, both assessments have been correlated, and consequently validated, to other adult ESL assessments used in the field.

Moderate to high correlations were found between *BEST Plus* and program placement levels of 24 adult ESOL programs nationwide.

Conclusion

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

Meanwhile, as program staffers 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.
- Arlington Education and Employment Program. (2005). *REEP Writing Assessment Technical Manual*. Arlington, VA: 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.
- Center for Applied Linguistics. (2005). *BEST Plus Technical Report*. Washington, DC: Author.
- Gass, S. M. (1997). *Input, interaction, and the second language learner*. Mahwah, NJ: Erlbaum.
- Grant, S. (2005). The REEP writing story. *Vertex: The Online Journal of Adult and Workforce Education*, 1(1). Retrieved on October 31, 2005 from <http://vawin.jmu.edu/vertex/article.php?v=1&i=1&a=2>
- Messick, S. (1989). Validity. In R. Linn, (Ed.), *Educational measurement* (3rd ed., pp. 11-103). New York: Macmillan.
- Mislevy, R.J. & Knowles, K.T. (eds.) (2002). *Performance assessments for adult education: Exploring the measurement issues*. Washington, DC: National Academy Press.
- National Reporting System for Adult Education. (2005, February). *Evaluating assessments for use in the National Reporting System* (Sect. 11-B). Retrieved October 31, 2005 from <http://www.nrsweb.org/thirdwave.asp>
- 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/aebprint2.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.

The Authors

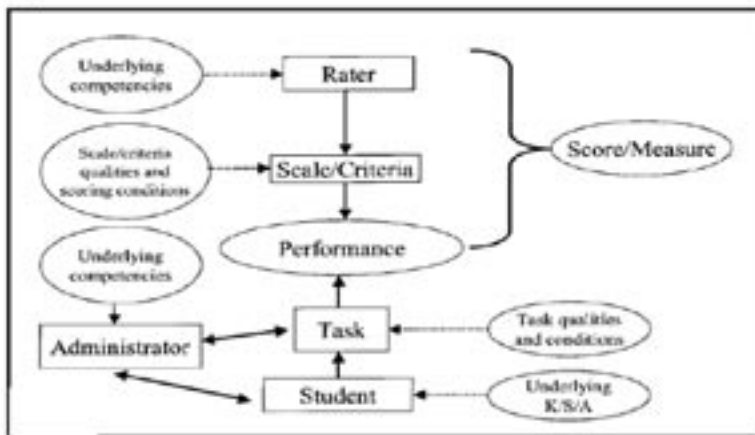
Dorry Kenyon (PhD, Measurement, Applied Statistics and Evaluation, University of Maryland; MA, Teaching English as a Foreign Language, American University in Cairo) is the Director of the

Language Testing Division at the Center for Applied Linguistics in Washington, DC.

Carol Van Duzer (MATESOL, Georgetown University) is an adult ESL assessment specialist with the Center for Adult English Language Acquisition and has extensive experience in teacher training and curriculum and materials development.

Sarah Young (MATESOL, Monterey Institute of International Studies) is an adult ESL specialist at the Center for Applied Linguistics in adult ESL assessment, research, and professional development projects, as well as a part-time adult ESL teacher.

Figure 1. Performance-Based Assessment Model



From introductory Remarks at the 14th Language Testing Research Colloquium on Development and Use of Rating Scales in Language Testing, by D. M. Kenyon, February–March 1992, Vancouver, Canada; from Measuring Second Language Performance, by T. McNamara, 1996, London: Longman; and from A Cognitive Approach to Language Learning, by P. Skehan, 1998, Oxford, England: Oxford University