Leni Puppin
BRAZIL

A Paradigm Shift: From Paper-and-Pencil Tests to Performance-Based Assessment

Even though many English as a Foreign Language (EFL) programs use Communicative Language Teaching, all too often their assessment methods do not correspond to this approach. This was the situation in early 2000 at the Language Center (CLC) of the Espirito Santo Federal University in Brazil. At the time, teachers were encouraged to utilize communicative methods and make language instruction interactive and relevant to the students' real-world interests, but at the same time they were asked to assess their students with traditional paper-and-pencil test items such as multiple choice questions and fill-in-the-blanks. Teachers, administrators, and students became dissatisfied because they recognized that there was a mismatch between paper-and-pencil tests that assessed what McNamara (1996, 6) calls the "abstract demonstration of knowledge," and the "actual performances of relevant tasks," which are commonly known as *performance tests*. (See Table 1 for examples of paper-and-pencil and performance test items.)

In addition to involving students in actual communication, performance tests are carefully designed to pose tasks "that are based directly on the learners' intended (or hypothesized) use of the target language" (Bailey 1998, 215). Since the communicative classroom focuses on exercises that are reflective of students' needs and real-world interests, performance tests that are designed around authentic tasks are a more valid way to assess the students' language learning progress.

Teachers and administrators agreed on the need for change in their assessment methods, and together they created the Assessment Project, which entailed contracting with an external specialist in program evaluation and language assessment for over two years to hold regular meetings with the CLC's teachers and coordinators. In 2000, this specialist stated that the goals of the Assessment Project were "to provide the EFL program at CLC

with a system that will permit the collection of valid, reliable, relevant and useful information about the performance of the students. This information should be analysed in a manner that will bring improvement to the current program" (M. Raupp, unpublished document).

This article will discuss the procedures the CLC undertook to make a positive change to the assessment system. Other teachers and administrators may find that reviewing these procedures will be a useful guide to create more meaningful ways to assess their students.

Table 1: Past and Present Assessment Procedures at the Language Center

1990–2000 **Traditional Paper-and-Pencil Test Items**	**2000–Present** **New Performance-Based Test Tasks**
1. Fill in the Blanks. Complete the five sentences below with the correct personal pronoun or possessive adjective (choose from the following: he/she/we/ they/it/ his/ her/our/their). a. **John** is American. _____ last name is Stevens. b. **Lisa and I** are in Room 10. _____ room is very nice. c. **Robert and David** are brothers. _____ are from California. d. The keys are in **Linda's** bookbag. _____ bookbag is brown. e. **Joseph and Kate's** phone number is 555-6608. _____ are brother and sister.	**1. ORAL PERFORMANCE** **SKILL:** Speaking **LEVEL:** Beginner **TASK:** You have five minutes to prepare a brief presentation about yourself. In one to two minutes, state in complete sentences: a) Your full name and the way you spell your last name b) Your age and phone number c) Where you and your family are from Students can also be given an interesting topic and time to prepare and then be interviewed or asked to make an oral presentation (one minute or less for beginners and longer for intermediate and advanced students).
1. Multiple Choice Listening Exercise. Listen to people talking and check (✓) the correct information. a) The woman is short and in her thirties. The woman is medium height and in her twenties. The woman is fairly short and about twenty-five. b) The man had a great vacation in Paris last year in July. The man hasn't been to Paris, France yet. The man can't wait to go to Paris in August. c) You shouldn't go to Las Ramblas because that's a very long street. You shouldn't miss some of the wonderful museums in Barcelona. You should visit Spain in January.	**2. LISTENING TASK** **SKILL:** Listening **LEVEL:** Intermediate **TASK:** Interesting and relevant audio- and video-taped material followed by open-ended questions and/or multiple-choice items. The teacher selects a familiar topic to conduct a *partial dictation* (certain omitted words are filled in by the students as the teacher reads) and a *graduated dictation* (students write down the dictation as the teacher reads progressively longer sentences). (Adapted from Bailey 1998, 15-18)
3. Complete the Dialogue. Write the questions for the following answers: a) _____? Yes I do. I play volleyball. b) _____? I play volleyball very well. c) _____? I usually spend about two hours a day. d) _____? Yes, Leila and Virna are pretty good at volleyball. e) _____? Well, I have two sisters and one brother. f) _____? No, we didn't. We stayed home and relaxed.	**3. WRITING TASK** **SKILL:** Writing **LEVEL:** Intermediate **TASK:** A Hypothetical Interview. What famous person would you like to interview? Why? In two paragraphs prepare an interview plan. In the first paragraph mention *who* you would like to interview and *why*. In the second paragraph, prepare five questions you would like to ask this person (things you think other people would like to know). **4. INTEGRATED SKILLS TASK** **SKILL:** Reading and Writing **LEVEL:** Advanced **TASK:** After reading a job announcement, write a business letter requesting an application and then fill out the application using the attached form. (Adapted from Bailey 1998, 209-212)

Dissatisfaction with traditional assessment

The CLC's desire to transform its assessment procedures grew out of dissatisfaction with the following aspects of traditional assessment:

- Teachers were encouraged to utilize communicative approaches but assessed their students with traditional paper-and-pencil tests.
- Teachers observed that traditional assessment was not reflecting students' actual potentials.
- Many traditional test items had poor *content validity*, which means the tests did not adequately measure the language skills that were being taught in the classroom (Popham 1981; Davies 1990; Heaton 1990).
- Different teachers who rated the same student compositions did not have an agreed upon judging process and therefore gave significantly different scores, resulting in highly subjective assessment and low *interrater reliability* (McNamara 1996; Gamaroff 2000).
- Teachers were either not provided with or had not developed *level descriptors*, which are concise statements describing the character of a minimally acceptable performance of an oral or written presentation (McNamara 2000). (See the Appendix for examples of level descriptors for an oral presentation.)
- Many traditional test items did not have *construct validity*, which means that they were not grounded on the theory of language acquisition that informs the communicative approach or the communicative methods being applied in the classroom (Popham 1981; McNamara 2000).
- The traditional testing produced negative instead of positive *washback* (also known as *backwash*), which is "the impact of tests on the teaching programme" (McNamara 1996, 23). In other words, traditional tests became the main focus of language instruction and did not contribute to student learning in a positive way.

Initiation of performance-based assessment

At the beginning of the transition to performance-based testing, the specialist provided the Assessment Project participants with literature pertinent to testing and measurement (see Davies 1990; Bailey 1998; and McNamara 2000). This research served two main purposes: (1) it provided an initial frame of reference on testing and measurement, and (2) it became a guide throughout the process of elaboration, administration, and refinement of the new performance-based testing program.

During the initiation of the new performance-based program, the specialist presented an essential overview of certain concepts fundamental to language assessment, including practicality, validity, and reliability.

Practicality

A quality assessment program typically requires the allocation of many resources, including materials, funds to hire outside experts, and the time of administrators and teachers. Therefore, an educational institution must be practical as it determines how to best dedicate the available resources while developing valid and reliable tests that promote positive washback (Bailey 1998).

Validity and reliability

All assessment programs must consider the validity and reliability of the testing instruments under creation. According to Popham (1981), *validity* is obtained if it can be demonstrated that the testing instrument is appropriate for the skills that one wants to measure, and *reliability* occurs when the testing instrument yields consistent results over repeated administrations.

Valid tests have a clear and demonstrable relationship with the actual skills being assessed, and the developers must follow precise guidelines to ensure this relationship. Data is collected at every stage of test development to document validity, which is also measured by the statistical results obtained from questionnaires and pre- and post-testing scores.

Reliable tests are administered in a consistent manner to all test-takers, and the developers must eliminate any conditions that might make the testing experience different from one student to the next. This includes

making sure that the testing environment is identical for all students, that all raters administer and score the tests in a standardized manner, and that all students have a clear picture of what is expected of them. These conditions become possible by publishing test development and administration guidelines for teachers and test content information for the students.

Valid and reliable tests are likely to produce positive washback. For example, when a test is linked to what students are learning in class, they will experience testing as an extension of classroom work. In contrast, if the test is not specifically related to classroom instruction, the experience will be stressful and will cause students to attempt to memorize a large number of language items, which leads to short-term learning. In addition, when students know what to expect and the grading criteria are clear, testing will be more informative and result in positive washback. For example, if a student taking a paper-and-pencil test graded on a 0 to 10 scale receives a score of 7.7, he or she might ask: "What does this tell me?" The student does not know specifically what a good performance means because the criteria for grading has not been made clear and there is a disconnect between teaching and assessment. On the other hand, if a student writes a composition about summer vacation and notes: "I think I deserve a good grade on this essay because it has a good title, an introduction and a conclusion, and it's not boring," then it is clear the student knows the required elements of good writing and how to achieve a good score. The second student assessment demonstrates the inseparable nature of teaching, learning, and assessment (Raupp 2003).

Stages to implement performance-based assessment

As described below, the implementation of the new performance-based assessment plan at the CLC took place in three stages over a period of one year. (See Table 2 for a summary of these stages.)

Stage 1 activities

1. Select teachers who represent English course beginner, intermediate, and advanced levels (based on individual competencies and length of experience with different levels) to participate in test development.
2. Define the hours of instruction required for each level: 200 for Beginner, 200 for Intermediate, and 100 for Advanced.
3. Develop a performance objectives continuum to guide teachers and describe what is expected from students at the end of each cycle in (a) Oral Production; (b) Reading Comprehension; (c) Listening Comprehension; and (d) Written Production.
4. Discuss, revise, and finalize the performance objectives continuum.
5. Construct a new four-point grading system (A, B, C, and D) to measure student performance.
6. Develop rating grids with level descriptors based on the four-point grading system. To reduce subjective interpretations, this requires carefully worded descriptors as well as the repeated training and monitoring of raters to make sure they assign consistent scores and achieve adequate levels of interrater reliability (McNamara 1996). (See the Appendix for a sample rating grid.)
7. Establish two periods for the assessment of four skills twice during the course: one at the end of the first 8 weeks and the other at the end of the course, or at 16 weeks.
8. Establish a pass/fail cutoff score for the Beginner, Intermediate, and Advanced level students.

Stage 2 activities

1. Identify and select appropriate testing instruments to measure the desired performance.
2. Create an *instrument bank*, or a collection of testing items and tasks that meet the pre-established criteria of validity and reliability.
3. Begin development of an *Assessment Guide* with thorough information for teachers about the new test development and administration procedures.
4. Plan to identify signs of failing as early in the course as possible and provide remedial class interventions for students who need extra assistance. (Because the CLC had teacher train-

ees available, it was feasible to offer remedial classes throughout the year to students whose performance needed improvement.)

Stage 3 activities

1. Revise and produce final drafts of support materials, including the Assessment Guide and a letter explaining the testing changes to students and parents.
2. Conduct training sessions with the English teaching staff.
3. Pilot test instruments on pre-selected groups of Beginner, Intermediate, and Advanced students.
4. Analyze qualitative and quantitative data, identify problems, and recommend solutions.
5. Publish results and implement changes to improve the quality of the instructional program.
6. Extend training sessions to the teachers of the other languages.

Table 2: Stages in Implementing Performance-Based Assessment at the Language Center

STAGES	ACTIVITIES	AGENTS OF CHANGE
1. JULY 2000	■ Determination of the instruction cycles ■ Development of a performance objectives continuum ■ Construction of a new grading system (A, B, C, and D) ■ Development of rating grids with level descriptors for each of the four language skills ■ Establishment of a pass/fail cutoff score	■ A specialist in program evaluation and language assessment ■ 5 permanent staff who are teachers of English ■ 4 teacher trainees ■ 2 pedagogic coordinators
2. JANUARY 2001	■ Creation of an instrument bank of valid and reliable items and tasks ■ Identification and selection of appropriate instruments ■ Production of a draft Assessment Guide ■ Creation of a plan for remedial classes	
3. JULY 2001 onwards	■ Revision and final draft of Assessment Guide and other support materials ■ Training sessions for teaching staff ■ Piloting of instruments on pre-selected groups ■ Analysis of data, identification of problems, and recommendations ■ Publication of results and implementation of changes ■ Introduction of training session to teacher of other languages	

Overview of results

Six years have passed since the CLC first began the transition from paper-and-pencil tests to performance-based assessment; during that time, pilot testing helped identify where to revise and improve the program. Pilot testing "instruments before actually employing them in final data collection is paramount" (Weir and Roberts 1994, 138). After the piloting in Stage 3 on pre-selected groups, the Assessment Project participants collected and analyzed qualitative data (test-taker feedback, teachers' reports on student progress, administration reports) and quantitative data (statistical analysis of scores and interrater reliability, reliability of pre- and post-tests). As a result, we made the following adjustments:

1. We identified and revised some tests or test components that were too easy or difficult.
2. We revised the four-letter grading scale by replacing the letter grades with numbers to allow for averaging of scores for pass/fail decision-making and to achieve more reliable overall results.
3. Instead of assessing students twice (at the midpoint and at the end), we now assess student performance at three intervals during the 16-week course.

Benefits of the new assessment process

As one of the participants and informal evaluators of the new assessment process, I observed the following beneficial results:

- The assessment procedures are clearer for everyone since the desired level of student performance and scoring criteria are clearly established.
- The mismatch between testing and teaching is greatly reduced because teaching activities are geared to the performance objectives and assessment.
- Teachers utilize fewer grammar driven activities and more real-world communicative tasks.
- The assessment instruments strongly correspond with the subject matter being taught and how it is being taught, increasing the content validity of the tests.
- The testing changes allow the teachers to document student progress systematically through formative assessment (daily in the classroom) and summative assessment (at the end of each level).
- The standardized administration, rating, and grading of the tests have increased the reliability of the assessment process.
- Teachers who participated in the process have a sense of "ownership" of the project.

Table 3 summarizes some of the benefits that resulted from the transition from traditional paper-and-pencil assessment to performance-based assessment.

Table 3: Transition from Traditional to Performance-Based Assessment
(adapted from Bailey 1998, 207)

One-shot tests	→	Continuous assessment
Textbook based tests	→	Classroom performance test
Inauthentic tests	→	More real-world assessment
Decontextualized test task	→	Contextualized test tasks
No feedback provided to learners	→	Feedback provided to learners in four skills
Subjective correction and grading	→	Standardized scoring criteria
No test follow-up	→	Remedial classes available
Negative washback	→	Positive feedback

Conclusion

Teachers and students have reacted positively to the new assessment procedures at the CLC, where testing has become a lever for instructional improvement. The EFL program now has a valid and reliable testing system to diagnose student strengths and weaknesses and identify staff development needs. Most importantly, the changes have not yielded a finished product because they are related to performance objectives and not to a specific textbook, which leaves room for an adaptation and further change if necessary.

References

Bailey, K. M. 1998. *Learning about language assessment: Dilemmas, decisions, and directions.* Boston: Heinle and Heinle.

Davies, A. 1990. *Principles of language testing.* Oxford: Blackwell.

Gamaroff, R. 2000. Rater reliability in language assessment: The bug of all bears. *System* 28 (1): 31–53.

Heaton, J .B. 1990. *Writing English language tests.* 2nd ed. London: Longman.

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

——. 2000. *Language testing.* Oxford: Oxford University Press.

Popham, W. J. 1981. *Modern educational measurement.* Englewood Cliffs, NJ: Prentice Hall.

Raupp, M., and A. Reichle. 2003. *Avaliação: Ferramenta para melhores projetos* [Evaluation: A tool for project improvement]. Santa Cruz do Sul, Brasil: Editora Edunisc.

Weir, C., and J. Roberts. 1994. *Evaluation in ELT*. Oxford: Blackwell.

LENI PUPPIN, who has a Master's in Education from the University of Chichester, UK, is a teacher trainer at the CLC and a Professor of English at the TEFL Specialization Program at the Espírito Santo Federal University in Brazil.

See Appendix on next page

Continued from page 9

Appendix Sample Dialogue

Using Replacement Performance Role-Plays… • **Maria Snarski**

Scene: Two students walking toward class and talking about the upcoming exam.

Student A: Good Morning!

Student B: Morning, are you ready for the exam?

Student A: No, I didn't really have a chance to study, but I have a little help in case I need it. (flashes a cheat sheet)

Student B: You're going to *cheat*?

Student A: Only if I have to. I didn't have time to study last night.

(They walk into the classroom, and Student A takes a seat next to Student B.)

Teacher: Good morning, class. As you know, there is an exam today. Please remove your books from your desks and just have your pencils ready. You will have 30 minutes for the exam. When you are finished, you may leave.

Scene: Student A visibly needs to cheat and tries looking at Student B's paper and looking at the cheat sheet, avoiding being caught by the teacher.
Student A finishes first and accidentally drops the cheat sheet. It lands near Student A. Student A leaves. Later, the teacher sees the cheat sheet and believes it belongs to Student B. The teacher questions Student B about the paper.

Appendix Oral Performance Rating Grid with Level Descriptors

Purpose: To identify student skill level in important language components and to record language progress.

Task: Given a familiar topic and five minutes to prepare, the student will make a coherent one- or two-minute oral presentation.

LANGUAGE COMPONENT	Level D	Level C	Level B	Level A
Fluency	Hesitant, makes repeated long pauses searching for ways to express him/herself. Often forced into silence by language limitations. Discourse is disconnected.	Speech is frequently disrupted by the student's search for the correct manner of expression. Frequently has problems linking ideas together in a logical sequence.	Speech is generally fluent with occasional lapses while student searches for the correct manner of expression. Can on occasion link ideas together in a logical sequence.	Speech is fluent and speech is rarely hesitant. Ideas are linked in a logical sequence.
Vocabulary	Misuse of words and very limited vocabulary make comprehension quite difficult. Resorts to L1 to fill in vocabulary gaps.	Frequently uses the wrong words. Conversation is somewhat limited because of insufficient vocabulary. Words are often repeated.	In general, uses appropriate terms and words. Occasionally must rephrase ideas because of vocabulary limitation.	Choice of words indicates a broad knowledge of vocabulary. Uses appropriate terms and words to express ideas.
Pronunciation	Very hard to understand because of pronunciation problems. Consistently needs to repeat words or sentences to be understood. Rarely uses appropriate intonation.	Makes him/herself understood, though pronunciation problems necessitate concentration on the part of the listener and occasionally lead to misunderstandings. Frequently uses inappropriate intonation.	Intelligible most of the time, though a definite foreign accent is noticed in his/her speech. Occasionally uses inappropriate intonation.	Always intelligible, although a foreign accent that does not impede communication is noticeable in his/her speech. Errors in pronunciation are rare. Almost always uses appropriate intonation.
Grammar	Grammar, word order, and verb tense errors make comprehension difficult. Restricts him/herself to the simplest grammatical structures or leaves sentences unfinished. Uses isolated words to express ideas.	Makes grammar, word order, and verb tense errors, which frequently obscure meaning and impede communication. Restricts him/herself to simple grammatical structures.	Makes occasional grammar, word order, and verb tense errors, which do not always obscure meaning.	Rarely makes grammar, word order, and verb tense errors that obscure meaning. Shows some degree of sophistication in the sequencing of tenses.