

SECTION V

Repairing and Improving Externally Developed Assessments

Kiyas oche kapetotimak kisteumowin.

[The way we do things is our knowledge.]

Old Metis saying quoted
by Elmer Ghostkeeper

W

hen the starting point is an existing assessment, Section V helps educators evaluate its quality, identify pitfalls for diverse students, and provide specific ideas for getting around the barriers. In the example below, a group of teachers concerned about the district's test banded together to develop an alternative assessment that would meet each of the characteristics of good assessment, to be a more accurate measure of student achievement (Trumbull, Koelsch, & Wolff, 1999).



VIGNETTE: THE REDWOOD CITY ASSESSMENT

In Redwood City, districtwide assessments are given in the fifth grade. By this grade most of the district's English language learners (nearly all of whom speak Spanish as a first language) have transitioned to English-only instruction.

The fifth-grade reading assessment (Performance-Based Assessment — PBA) requires students to read a narrative text and an expository text and respond in writing to a series of questions about them. The prototype PBA, containing 15 questions, addresses standards related primarily to vocabulary, comprehension, summarizing, interpreting literary elements, connecting text meaning to one's own life, and using structural and design elements of books (maps, charts, captions, etc.) to enhance comprehension.

The group of concerned teachers judges that the texts used on the PBAs are at too high a reading level, vocabulary is arcane, and the content and themes are

not culturally appropriate for these students. In addition, the directions are too long and visually dense with text, and responses call for considerable writing.



THINGS TO CONSIDER

- The Redwood City teachers are challenged to ensure that all of their students have an equitable opportunity to demonstrate their reading proficiency. However, is there more at stake here than demonstrating proficiency?
 - What are the issues they must address as they move through the process of meeting student needs?
 - What choices do they have?
-

Like others facing a challenge to be accurate and fair in the assessment of their students, Redwood City educators had a number of options to consider as they developed the Transitional Performance-Based Assessment for their students. The previous section demonstrated why culture and language should be considered in all aspects of the testing of cultural and linguistic minority students. This section offers practical recommendations for promoting equitable testing when teachers select, use, administer, and develop assessments.

Assessments are usually developed based on assumptions about the values, experiences, beliefs, and learning styles of the students. Some of these assumptions may not hold for the entire population. As a consequence, the way in which exercises are designed, prompts are phrased, and student responses are scored may produce an inaccurate picture of the knowledge and skills of cultural minority students.

One effective way in which teachers can contribute to more equitable assessment practices consists of being critical consumers of assessments. From our perspective, being a critical assessment consumer comprises three activities: knowing the challenges that we need to overcome to improve our assessment practices, being aware that often assessments for minority students are defective because of flawed adaptation and translation procedures, and using certain assessment development and scoring strategies intended to address student diversity.

There are some actions that we can incorporate into our daily practice to promote more equitable testing practices. These actions concern our activities when we review the assessments we use, when we develop assessments of our own, and when we score student responses.

How to Make Assessment Work for Everyone: Reviewing Assessments

Accomplished teachers know very well that developing an assessment is a never-ending process. Assessments for English language learners should not be an exception.

To attain more equitable testing, we need to become critical reviewers of all assessments we use, whether we are the authors of these assessments, we borrow them from colleagues, or we purchase them from testing companies.

In addition to using our own judgment, we can take three actions to review assessments: using external reviewers, interpreting students' responses, and obtaining student verbalizations. These actions should be taken continuously with any assessment, even if it has been used many times in the past and is regarded as “non-culturally biased.”

Using External Reviewers

People with different skills and backgrounds can be sensitive to flaws that others may overlook. In addition to colleagues, external reviewers can include translators, content specialists, native speakers of the language targeted, and parents. Use more than one of each. We have observed that many errors made in tests designed for English language learners occur because organizations or development teams rely on the opinion of only one parent or only one native speaker.

Use as many reviewers as possible. Each person may give a different opinion. We will need to use our judgment to decide how the exercises can be improved.

When we review or have someone review a translation, we should always have the original version in English with us. As basic as this idea may sound, many organizations give their reviewers only the translated version. Without the original version in English, the reviewer can review, at the most, the correctness of the language used, not the accuracy of the translation. It is not unusual to run into translations in which the

original, intended meaning has been altered, but the scoring rubrics or criteria are the same across languages.

In addition to reviewing language, have reviewers check for cases in which exercises:

- Exclude or offend underrepresented groups (e.g., the characters used are only male)
- Contribute to perpetuating stereotypes about those groups (e.g., the characters are a male and a female with divergent points of view; the male is right)
- Promote the notion that the content area is not open to all groups (e.g., the characters' names in an exercise sound like White, European American names)
- Include descriptions of situations, words, sentence structures, or pieces of equipment that are familiar to only a specific group of students

Interpreting Student Responses

Interpreting English language learners' responses accurately requires a recognition that their native language may strongly influence how they interpret an item and how they respond to it. Being able to identify those influences allows teachers to obtain valuable information of their students' strengths and weaknesses.



VIGNETTE: AN EXAMPLE FROM THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (NAEP)

Using vocabulary that is age or linguistically inappropriate may result in exercises that penalize certain groups of English language learners. Here is an example provided by Solano-Flores and Nelson-Barber (2000):

One of the 1996, fourth-grade science questions NAEP released reads as follows:

“A nail becomes warm when it is hammered into a piece of wood. Tell why the nail becomes warm.” ()*

One student responded as follows: “because in side of the tree is warm.”

Most likely, this student's native language is Spanish. In addition to having a Spanish syntactical structure, the response reflects two typical mistakes made by English learners whose native language is Spanish: confusing "nail" and "snail" and confusing "warm" and "worm," whose sounds they confuse. Because of this and the fact that "tiered into" is unfamiliar to beginning English learners, quite probably the student interpreted the item as if it read:

"A snail becomes a worm when it is (?) into a piece of wood. Tell why the snail becomes a worm."

The student clearly missed the fact that the item was about physics and tried to make up a biology response.

-- In the same document (NAEP, 1996 Assessment: Science, Public Release, Grade 4), the prompt for one of the student response samples reads: "A nail becomes warm when it is tiered into a piece of wood. Tell why the nail



THINGS TO CONSIDER

Although it is impossible to know whether the student would have answered correctly had she or he understood what the item was asking, the example illustrates how an unfamiliar word may cause a student to perform poorly on an essay. It also illustrates how the interpretation of student responses can be used to improve an exercise.

Obtaining Student Verbalizations

The reasoning used by students in their responses reveals whether an exercise is eliciting the intended kind of thinking. Asking students to talk about their responses can help clear up problems in the assessment as well as reveal possible student misconceptions or flawed problem-solving strategies and identify the linguistic and cultural influences that shape student responses.

We can ask the students to tell us in their own words what they think the question or task is about. This strategy allows us to detect possible misinterpretations of the question or task due to inappropriate translation. A good, effective way to improve a question or task consists

of asking the students to rephrase it (e.g., How would you ask this question to a friend?).

Also, we should ask students to “think aloud” as they engage in responding to the items. By asking them to describe what they are thinking, we will be able to identify whether they respond to the item as they are expected to.

An alternative to this strategy is the approach of interviewing students individually, right after they have completed the item, and asking them to explain their thinking retrospectively.

How to Make Assessment Work for Everyone: Developing Assessments

Concurrent Assessment Development

Developing two language versions of the same assessment concurrently is the best approach to promote equitable testing when developing an assessment from scratch for English language learners. We should not wait until we have developed the English version to start developing a version in another language. Educators must develop both versions at the same time and with the same team of assessment developers. By allowing both language versions to go through the same process of development (which includes revising the assessments iteratively based on the results obtained from piloting them with the students), we give them equal opportunities for refinement and for capturing sociocultural aspects relevant to language (Solano-Flores, Trumbull, & Nelson-Barber, 2000).

The rationale for concurrent assessment development is that adapting tests to address cultural and linguistic diversity has a serious methodological weakness that prevents testing from being equitable: the process used to develop the original version of the assessment is different from the process used to develop the adapted version. Developing an assessment is a cyclical process. During each iteration, developers observe the students or ask them to talk aloud as they perform in order to gain access to the reasonings they use to solve the problems. Developers also interview students to investigate how well they understand the problems posed and the kind of knowledge and thinking skills they use in solving the problems. In each iteration, developers discuss their findings and create a more refined version of the assessment. Among other things, they may rephrase questions; add, replace, or eliminate words; or include examples in the directions for equipment use (Solano-Flores & Shavelson, 1997). When an assessment



DID YOU KNOW?

Researchers from the Language and Cultural Diversity program at WestEd are currently investigating the potential and limitations of certain alternatives to translating assessments (Solano-Flores, Trumbull & Nelson-Barber, 2000). They contend that a more equitable testing of linguistic minority students in their first language can be attained only if both the source and secondary language versions are given the same status and are developed concurrently. Their model for concurrent assessment development is currently being tested with bilingual teachers in the Wenatchee School District, in the state of Washington.

is translated from English into another language, this delicate process that allows assessment developers to refine language does not take place for the other language.

Designing Appropriate Ways for Students to Respond

Nothing is trivial when it comes to deciding the characteristics of an assessment's response format — the methods intended to capture the student responses. This is especially the case in situations in which students have different cultural backgrounds. Different cultures tend to produce different ways to interpret and respond to questions. Consider the following cases. They have been identified from observing students with different cultural backgrounds taking science assessments designed for either research purposes or large-scale testing (Solano-Flores & Shavelson, 1997).

- English language learners can benefit from constructed-response exercises that allow them to provide their answers with drawings. They may provide more information if the spaces for drawing are not bounded by boxes.
- Students are more likely to make drawings, computations, and notes on observations and thoughts when the question provides a blank half a page than when it provides half a page with lines to write on. Lines tend to elicit written responses only. In addition, the number of lines suggests that a specific response length is expected. Also, too many lines may have the undesired effect of intimidating some students, thus discouraging them from attempting to write an elaborate response.

- Opposite pages in notebooks with double-sided pages allow students to keep information close at hand. If that information is provided on the left page and the questions are arranged on the right page, students can use the information to answer the questions without having to turn to back pages.

Finding the response format characteristics that best work for a specific cultural group may take a great deal of experience. In this process, we always need to be aware that any decision on the way a question is worded and the physical appearance of the response format is relevant to the testing of English language learners.

Striking a Balance between Reading Demands and Contextual Information

At the beginning of the 1990s, when multiple-choice testing was under scrutiny, high hopes were placed on constructed-response assessments as potential tools for promoting more equitable testing. It was thought that, by their concrete, well-contextualized nature, verbal-linguistic skills might play less of a role than with abstract, decontextualized, multiple-choice, and other traditional tests.

In some cultures, knowledge occurs in a contextualized manner; isolated pieces of information — typical of multiple-choice items — are meaningless unless they are related to a specific situation. Therefore, it makes sense to assume that, for students from these cultural groups, constructed-response assessments can provide a better opportunity for cultural minority students to demonstrate their knowledge.

However, providing too much contextual information to make a task meaningful increases the reading demands and may place those students at a disadvantage if they are, in addition, poor readers.

As with any step taken in the process of assessment development, striking a balance between the reading demands posed by an assessment and the contextual information needed to make its task meaningful takes a good number of tryouts and revisions.



VIGNETTE: BAGELS

In the following vignette, a teacher suspects that a district test may prove challenging for his students.

Mr. Powell, a third-grade teacher in New Orleans, is preparing his students for a new districtwide reading assessment. He just earned his license to teach English language learners, and for the first time, most of his students are from immigrant families. He has a combination of students whose families are from Vietnam and Laos. The district has released a list of books that may be used on the exam. His class has read most of them already, but there are still a few that he wants to make sure they have a chance to experience. One of the books is Mrs. Katz and Tush by Patricia Polacco. It is one of his favorite books, and he is surprised that he hasn't used it yet with these students. As he flips through the book at home, he notices many references to traditionally Jewish foods. He believes his students will understand the story, given the pictures and context clues, but he is worried that on an assessment with multiple-choice questions, his students may get stuck on the vocabulary and have trouble comprehending the reading passages. He decides to bring in bagels for his class as a treat and a basis for discussion before he reads the book. He wants to demystify the language that will be strange to them as much as possible. Predictably, as he covers the bagels with cream cheese and passes them out, his students stare at him with wonder and ask, "Teacher, what is that?"



THINGS TO CONSIDER

In this case, Mr. Powell was not able to redesign an assessment that he felt would ensure student success, but he was able to prepare his students for the exam. What further steps would you have taken to prepare

your class for this reading test? If Mr. Powell had been on the assessment design committee, what suggestions should he have made?

Using Multiple Forms of Assessment

An increasing body of evidence (Baxter & Shavelson, 1994; Dalton, Morocco, Tivnan, & Rawson, 1994; Ruiz-Primo & Shavelson, 1996) shows that different types of tasks (e.g., multiple choice, short answer, essay, hands on, computer simulations, concept maps) tap into different types of knowledge and skills. For example, a student may get high scores in a hands-on assessment and low scores in a computer version of that assessment, even if both assessments pose exactly the same problem and vary only on whether the student manipulates real or virtual objects. Seemingly parallel ways of designing assessment questions can result in different responses because they require different, unintended skills.

The reasons that account for differences in score due to type of task are still being investigated. A possible reason is that, because of their different experiences and skills, some students perform better on tasks that, say, involve reading long paragraphs; others perform better on tasks that involve manipulating objects; and still others perform better on tasks that involve interpreting visual representations of objects.

Gender and racial differences in performance have been investigated for only a few of the new alternative assessments (e.g., Jovanovic, Solano-Flores, & Shavelson, 1994; Klein, Jovanovic, Stecher, McCaffrey, Shavelson, Haertel, Solano-Flores, & Comfort, 1997). There is scant information on how different tasks elicit from students different styles of thinking and problem-solving strategies depending on their cultural and linguistic backgrounds.

Since each culture promotes different sets of skills and values among its members, it is possible that individuals from some cultural groups may tend to perform better on some tasks than on others. Using a limited variety of assessments in a multicultural classroom may privilege some students and penalize others.

The solution to this potential dilemma consists of using a wide variety of tasks in classroom assessment. This not only renders more dependable measures of student academic achievement, but also ensures that all students are given the opportunity to demonstrate their learning.

How to Make Assessment Work for Everyone: Scoring Assessments

Assessing to Understand (Not Just Grade) Our Students

Alternative, constructed-response forms of assessment allow for focusing not only on the product but also on the process (reasoning) used by students to respond to an item or solve a problem.

Since constructed-response assessments permit a variety of responses varying in degree of correctness, the universe of possible correct responses may be vast and, in many cases, undetermined. This means that some students can come up with unanticipated responses or solutions that are, nonetheless, good.

Unfortunately, probably because of the long tradition of multiple-choice testing in the United States, many American teachers still limit their grading of constructed-response exercises to judgments about the product, not the process involved in the students' responses.

Assessing students based solely on product is not consistent with the rationale for using constructed-response assessments. This may adversely affect students who are different from the mainstream population. This is especially the case for students who, because of their linguistic and cultural backgrounds, may provide their responses in novel ways.

Understanding what students are trying to say before assigning a grade to that response is a good approach to promote equity in the scoring of constructed-response exercises. A careful reading of student responses (even those that at first glance look inadequate) may reveal in some cases that a student has a reasonably good understanding of a given topic but cannot articulate a response in the same way as native speakers of English. In addition, a careful reading of student responses provides teachers with valuable information that they can use to help students improve their written communication skills.

Separating Content Knowledge From Writing Skills

One of the major challenges for objective scoring in large-scale assessment is the difficulty of training scorers to focus on the skill or knowledge that a given assessment intends to measure. This is probably due to the fact that, whereas assessments are supposed to assess knowledge on a specific skill or knowledge domain, educators are used to assessing individuals as a whole. As a result, they may tend to judge an

individual's performance based on things that are irrelevant to the targeted learning.

The influence of grammar and spelling mistakes in scoring is a case in point. Teachers may find it difficult to resist the temptation to assign a low score to, say, performance on a science item if it is plagued with grammar and spelling mistakes. Obviously, this tendency may affect English language learners more than other students in the classroom.

Being Sensitive to Cultural Values

Knowing the way cultures view and make sense of the world may lead to more accurate inferences about an individual's capabilities. Cultures' norms may influence the way in which students interpret and solve an item (Kopriva & Sexton, 1999). For example, students may be asked to create a fair race. Students are expected to create a racecourse in which each contestant runs the same distance. Those from cultures that do not emphasize competition may interpret the word "fair" in a different way and create shorter distances for slower runners, as if the item asked them to create a racecourse in which all contestants have equal chances of winning.

How to Make Assessment Work for Everyone: Giving Students Opportunities to Communicate Learning in Their Most Fluent Language

Finally, students who are learning English frequently have trouble expressing their knowledge and skills in English. Their progress in English must be assessed, of course, but their learning in other academic areas must also be assessed. Therefore, in order to truly discover what they know and are able to do, English language learners often need accommodations in assessment, including assessment in native language. The decisions made by Redwood City teachers and bilingual staff continue to provide a good example of how to meet the needs of English language learners.



VIGNETTE:
REDWOOD CITY: GETTING DOWN TO SPECIFICS

The Redwood City Department of Bilingual Education became convinced that merely introducing special accommodations for transitional students, such as allowing them to read texts in Spanish as well as English or take more time to complete the assessment if they so desired, would not address students' actual needs. Mediation, such as rephrasing prompts or instruction or responding to student questions, also seemed inadequate to solving what teachers saw as the real problems with the assessments. Staff believed that an alternative assessment specifically designed for students transitioning into English-only instruction was the best answer.

Teachers reasoned that an alternative reading assessment should have the following characteristics:

- *Appropriate texts (in terms of themes, level, and length)*
- *Clear instructions and accessible vocabulary*
- *Response strategies similar to what students had been exposed to in instruction (appropriate for English language learners)*
- *A similar sampling of the district's reading standards to that of the PBA*
- *A format as similar as possible (in terms of numbers and types of items) to that of the regular PBA*

The intent was to create an English reading assessment that would perform the same functions as the PBA, sample many of the same standards, and replicate the PBA in form to the degree possible. It needed to be appropriate for English language learner students in their first year of English reading. The Transitional Performance-Based Assessment would not be comparable to the PBA in any sense of equating of the assessments, but neither would it operate on an entirely different set of premises. In fact, it would be parallel in many ways.

The Redwood City assessment design group decided that to read aloud one paragraph of the reading selection prior to distributing the full passage would improve performance. The group felt strongly that hearing the word in the context of the story would elicit greater involvement in the pre-assessment activities than putting the word on the board and asking students if they knew its meaning. Moreover, reading aloud the paragraph in question provided students with a definition of the word “elaboration.” The technique of defining a word through elaboration and description, rather than saying the word and asking for definitions, is commonly used in the teaching of second languages. Thus, the preassessment activity paralleled sound instructional practice. Teachers believed that using this technique in the assessment would put students at ease and lower their anxiety about taking an important test.



THINGS TO CONSIDER

English language learners can be at a wide array of levels of proficiency in English and may need very different accommodations. When is native language assessment needed? When are modified instruction, simplified vocabulary/language, and/or longer time periods appropriate? When should students be allowed dictionaries or interpreters?

Summary

Unfortunately, teachers do not have complete control over all of our students' assessment experiences. We all have to use some external tests and materials as well as submit to large-scale state or district assessment, even though it may sometimes be against our better judgment. Regardless, teachers can still be critical consumers of external assessments and can strive to make student experience as positive as possible. Some things to consider include strategies for reviewing assessments, developing assessments, scoring assessments, and assessing students in their own language.

Finally, when we discover errors in assessments that may potentially affect students from cultural minorities, we should make their creators aware of them. Many errors may go unnoticed by testing companies because they do not get enough feedback from teachers. By providing this feedback on a continuous basis, teachers will make companies realize that they need to improve their translation and cultural adaptation procedures and will contribute to a more equitable assessment.