
The Alternatives in Language Assessment

JAMES D. BROWN and THOM HUDSON
University of Hawai'i

Language testing differs from testing in other content areas because language teachers have more choices to make. The purpose of this article is to help language teachers decide what types of language tests to use in their particular institutions and classrooms for their specific purposes. The various kinds of language assessments are classified into three broad categories: (a) selected-response assessments (including true-false, matching, and multiple-choice assessments); (b) constructed-response assessments (including fill-in, short-answer, and performance assessments); and (c) personal-response assessments (including conference, portfolio, and self- or peer assessments). For each assessment type, we provide a clear definition and explore its advantages and disadvantages. We end the article with a discussion of how teachers can make rational choices among the various assessment options by thinking about (a) the consequences of the washback effect of assessment procedures on language teaching and learning, (b) the significance of feedback based on the assessment results, and (c) the importance of using multiple sources of information in making decisions based on assessment information.

A variety of alternative assessments have become popular in recent years. Alternative assessment procedures listed by Huerta-Macías (1995) include checklists, journals, logs, videotapes and audiotapes, self-evaluation, and teacher observations. We would add at least portfolios, conferences, diaries, self-assessments, and peer assessments. But what makes these *alternative assessments*, whereas other types of assessments are called *traditional assessments*? In other words, what common characteristics make these types of assessments special and different? Various authors have different answers to this question. Aschbacher (1991) lists several common characteristics of alternative assessments, stating that they

1. require problem solving and higher level thinking,
2. involve tasks that are worthwhile as instructional activities,
3. use real-world contexts or simulations,

4. focus on processes as well as products, and
5. encourage public disclosure of standards and criteria.

Herman, Aschbacher, and Winters (1992, p. 6) offer a somewhat different set of characteristics. In their view, alternative assessments

1. require students to perform, create, produce, or do something;
2. tap into higher level thinking and problem-solving skills;
3. use tasks that represent meaningful instructional activities;
4. approximate real-world applications;
5. ensure that people, not machines, do the scoring, using human judgment; and
6. call upon teachers to perform new instructional and assessment roles.

Huerta-Macías (1995) says that alternative assessments

1. are nonintrusive in that they extend the day-to-day classroom activities already in place in a curriculum,
2. allow students to be assessed on what they normally do in class every day,
3. provide information about both the strengths and the weaknesses of students, and
4. are multiculturally sensitive when properly administered.

Combined, the characteristics listed in the three papers cited above provide an impressive list of positive characteristics for alternative assessments that should appeal to most language teachers and testers alike:

1. require students to perform, create, produce, or do something;
2. use real-world contexts or simulations;
3. are nonintrusive in that they extend the day-to-day classroom activities;
4. allow students to be assessed on what they normally do in class every day;
5. use tasks that represent meaningful instructional activities;
6. focus on processes as well as products;
7. tap into higher level thinking and problem-solving skills;
8. provide information about both the strengths and weaknesses of students;
9. are multiculturally sensitive when properly administered;
10. ensure that people, not machines, do the scoring, using human judgment;

11. encourage open disclosure of standards and rating criteria; and
12. call upon teachers to perform new instructional and assessment roles.

RELIABILITY AND VALIDITY ISSUES

However, other claims made by advocates of alternative assessments may not be quite so universally acceptable. For instance, Huerta-Macías (1995) argues that the

trustworthiness of a measure consists of its credibility and auditability. Alternative assessments are in and of themselves valid, due to the direct nature of the assessment. Consistency is ensured by the auditability of the procedure (leaving evidence of decision making processes), by using multiple tasks, by training judges to use clear criteria, and by triangulating any decision making process with varied sources of data (for example, students, families, and teachers). Alternative assessment consists of valid and reliable procedures that avoid many of the problems inherent in traditional testing including norming, linguistic, and cultural biases. (p. 10)

Although we are excited about the possibilities of developing new assessment procedures that provide opportunities for students to demonstrate their abilities to use language for meaningful communication (in ways that are consonant with the particular curriculum in which they are studying), we must take issue with the statements made about reliability and validity.

We agree that, in part, the “trustworthiness of a measure consists of its credibility and auditability.” However, we believe that trustworthiness so defined is not enough. We also agree that consistency is aided by “the auditability of the procedure (leaving evidence of decision making processes), by using multiple tasks, by training judges to use clear criteria, and by triangulating any decision making process with varied sources of data (for example, students, families, and teachers),” but that is not enough either. We are very concerned about the attitudes expressed above that somehow the consistency of alternative assessments is “ensured” by the various strategies listed and that somehow such procedures are “in and of themselves valid, due to the nature of assessment.” These statements are too general and shortsighted to fit with our experiences as decision makers who, from lifelong habit, rely on the guidelines set forth in the *Standards for Educational and Psychological Testing* (American Psychological Association, 1985, 1986) for designing measures that will be used to make responsible decisions about students’ lives.

Certainly, we would agree that credibility, auditability, multiple tasks, rater training, clear criteria, and triangulation of any decision-making procedures along with varied sources of data are important ways to improve the reliability and validity of any assessment procedures used in any educational institution. In fact, these ideas are not new at all. What is new is the notion that doing these things is enough, that doing these things obviates the necessity of demonstrating the reliability and validity of the assessment procedures involved.

Those strategies are not enough. Like all other forms of assessment, the so-called alternative assessments are used to make decisions, sometimes very important decisions, about peoples' lives. As in all other forms of assessment, the designers and users of alternative assessments must make every effort to structure the ways they design, pilot, analyze, and revise the procedures so the reliability and validity of the procedures can be studied, demonstrated, and improved. The resulting decision-making process should also take into account what testers know about the standard error of measurement and standards setting.¹ Precedents exist for clearly demonstrating the reliability and validity of such procedures in the long-extant performance assessment branch of the educational testing literature, and the field of language testing should adapt those procedures to the purposes of developing sound alternative assessments. These existing procedures for showing the reliability and validity of performance assessments are not new, nor are they difficult from logical or technical perspectives. Hence, we find the views that the consistency of alternative assessments is "ensured" and that they are "in and of themselves valid" to be incredible. Such a stance could easily lead to irresponsible decision making. As we point out elsewhere, "the issues of reliability and validity must be dealt with for alternative assessments just as they are for any other type of assessment—in an open, honest, clear, demonstrable, and convincing way" (Norris, Brown, Hudson, & Yoshioka, 1998, p. 5).²

ALTERNATIVES IN ASSESSMENT

The literature on language testing is full of examples of new and innovative types of tests being introduced. Examples are, to cite just a

¹For more information on standards setting, especially for performance assessments, see Bond (1995), Brennan and Johnson (1995), Burger and Burger (1994), Green (1995), Guion (1995), Jaeger (1995), Linn (1994), Linn and Burton (1994), Messick (1995), and Yen (1997).

²Indeed, numerous other criteria exist for evaluating the quality and effectiveness of assessments, such as consequences, fairness, transferability and generalizability, cognitive complexity, content quality, and content coverage (see, for instance, Linn, Baker, & Dunbar, 1991).

few, various types of composition tests, cloze tests, c-tests, cloze elides, dictations, reduced-forms dictations, oral proficiency interviews, simulated oral proficiency interviews, role-play tests, group tests, task-based tests, and performance assessments. Portfolios, conferences, diaries, self-assessments, and others are becoming increasingly prominent in the literature, even appearing in some language testing books (e.g., Cohen, 1994; Genesee & Upshur, 1996; McNamara, 1996). New assessment alternatives are always exciting and interesting, but one should not view them as somehow magically different.

In our view, the phrase *alternative assessments* may itself be somewhat destructive because it implies three things: (a) that these assessment procedures (like alternative music and the alternative press) are somehow a completely new way of doing things, (b) that they are somehow completely separate and different, and (c) that they are somehow exempt from the requirements of responsible test construction and decision making. We view procedures like portfolios, conferences, diaries, self-assessments, and peer assessments not as *alternative assessments* but rather as *alternatives in assessment*.³ Language teachers have always done assessment in one form or another, and these new procedures are just new developments in that long tradition.

LANGUAGE TEACHERS' ALTERNATIVES IN ASSESSMENT

Language testing practices are fundamentally different from assessment practices in most other disciplines, not only because of the complexity of the domain being tested but also because of the different types of tests that language teachers and administrators can and do use. From discrete-point tests like the multiple-choice and true-false tests used predominantly in the 1950s and 1960s, to the integrative tests like cloze and dictation used in the 1970s and early 1980s, to the more communicative tests like task-based and other new assessments used in the 1980s and 1990s, language testers have tried out, researched, and argued about a wide variety of different types of tests. Which tests are most valid? Which tests are most reliable? Which tests are easiest to score? Which tests measure what skills? These are all legitimate questions. But the one idea that seems to get lost in the shuffle is that virtually all of the various test types are useful for some purpose, somewhere, sometime. In other words, all of the different types of tests are important to keep because all of them have distinct strengths and weaknesses.

³This distinction between the labels *alternative assessments* and *alternatives in assessment* was first suggested by our coauthor, John Norris, in Norris et al. (1998).

In search of a way to explain the relationships among these various types of tests to students in our language testing and advanced language testing courses, we have organized the discussion around the three basic assessment types: (a) *selected-response* (including true-false, matching, and multiple-choice assessments); (b) *constructed-response* (including fill-in, short-answer, and performance assessments); and (c) *personal-response* (including at least conference, portfolio, and self- and peer assessments). In this article we define each type of assessment, highlight the relative strengths and weaknesses of each, and discuss how teachers can choose among the many options with two primary considerations: the importance of the washback effect and the crucial need to use multiple sources of information in making decisions. The article ends with suggestions to help teachers choose among the various options.

Selected-Response Assessments

Selected-response assessments present students with language material and require them to choose the correct answer from among a limited set of options. In selected-response assessments, students typically do not create any language. Thus these assessments are most appropriate for measuring receptive skills like listening and reading.

In general, selected-response assessments are relatively quick to administer. In addition, scoring them is relatively fast, easy, and relatively objective. However, these assessments have two disadvantages: (a) They are relatively difficult for the test writer to construct, and (b) they do not require students to use any productive language. Three types of selected-response assessments are commonly used: true-false, matching, and multiple-choice.

True-False

True-false assessments present a sample of language and require the students to respond to that language by selecting one of two choices, true or false. The primary strength of true-false assessments is that they focus on the students' abilities to select the correct answer from two alternatives. Thus true-false assessments provide simple and direct indications of whether a particular point has been understood. One problem with true-false assessments is that, in order to produce items that discriminate well, test writers may be tempted to write items that are tricky, that is, that turn on the meaning of a single word or phrase or that depend on some ambiguity. Most teachers prefer to create straightforward assessments wherein students who know the answer get it correct and students who do not know the answer get it wrong.

The relatively large guessing factor is another problem for true-false assessments. In fact, the examinees have a 50% chance of answering correctly even if they do not know the answer. However, if a large number of carefully designed true-false items are used, the overall score should overcome much of the guessing factor's influence.

If the language knowledge or skills a teacher wants to test lend themselves to two-way choices and enough items can be written, true-false items may turn out to be useful. However, because true-false assessments tend to place emphasis on details and unimportant facts, it may be difficult to find 25 nontrivial points to assess (in, for example, a listening or reading passage).

Matching

Matching assessments present students with two lists of words or phrases from which they must select the words or phrases in one list that match the ones in the other list. The main advantages of matching assessments are that they are relatively compact in terms of space and have a low guessing factor (for instance, only 10% for 10 items if extra options are supplied). Matching assessment is generally restricted to measuring students' abilities to associate one set of facts with another, which in language testing usually means measuring passive vocabulary knowledge (i.e., the students' ability to match definitions to vocabulary items).

Multiple Choice

Multiple-choice assessments require students to examine a sample of language material and select the answer that best completes a sentence or best fills in a blank in the sentence from among a set of three, four, or five options. Multiple-choice assessment, like matching assessment, has the advantage of a relatively small guessing factor. Whereas true-false assessment has a 50% guessing factor, multiple-choice assessment typically has a 33%, 25%, or 20% guessing factor depending on whether there are three, four, or five options. Multiple-choice assessment also has the advantage of being useful for measuring a fairly wide variety of different kinds of precise learning points.

Multiple-choice assessments are frequently criticized by language teachers because real-life language is not multiple choice. In truth, authentic productive language use rarely offers options from which speakers can select, so avoiding the use of multiple-choice assessment (or true-false or matching assessments, for that matter) for measuring productive skills like writing and speaking is just common sense.

Nonetheless, many aspects of language, particularly the receptive skills, can be tested using multiple-choice assessment. Testing reading, listening, grammar knowledge, and phoneme discrimination with multiple-choice items can provide useful information about students' abilities or knowledge in those areas with relative efficiency. Unfortunately, because reading, listening, and grammar skills are often the only assessments measured on the commonly used proficiency and placement tests, multiple-choice assessments have often been overused. Looked at in reverse, the pervasive use of multiple-choice items (usually because of ease of administration and scoring as well as objectivity) may often have limited the types of language skills that were tested in reading, listening, and grammar. In addition, multiple-choice items have sometimes been twisted to uses that seem quite inappropriate (for instance, multiple-choice assessment of writing ability).

Constructed-Response Assessments

Constructed-response assessments require students to produce language by writing, speaking, or doing something else. Hence, these assessments are probably most appropriate for measuring the productive skills of speaking and writing. Constructed-response assessments can also be useful for observing interactions of receptive and productive skills, for instance, the interaction of listening and speaking in an oral interview procedure or the interaction of reading and writing in a performance assessment in which students read two academic articles and write an essay comparing and contrasting them.

There are certain trade-offs in deciding whether to use selected-response or constructed-response assessments. For example, selected-response items allow for some guessing, but they are relatively objective; on the other hand, constructed-response items eliminate some of the guessing factor but create problems of subjectivity, especially when human judgments get involved in deciding what is a correct answer for a blank or short answer or when raters score the language samples.

The guessing factor is less of a problem on constructed-response assessments. However, they are not completely immune from guessing, though guessing on constructed-response assessments might better be called bluffing. For example, on a composition examination, some students might try to use key words in the prompt to write around the topic or take a shotgun approach to answering in the hope of hitting something that will be counted as correct. Although this is a type of guessing, it is guessing that scorers or raters can see if they are alert to its possibility.

In general, constructed-response assessments have virtually no guessing factor, and they measure productive language use as well as the interaction of receptive and productive skills. However, bluffing is a possible problem, and scoring may be relatively difficult and time-consuming. Constructed-response assessments may also be fairly subjective depending on the type. Three types of constructed-response assessments are commonly used in language testing: fill-in, short-answer, and performance assessments.

Fill-In

Fill-in assessments give a language context with part of the context removed and replaced with a blank. To answer, students are required to fill in the blanks. Fill-in assessment comes in many shapes and forms, from single-word fill-in items in single sentences to cloze passages with many blanks embedded in a longer stretch of text.

Fill-in assessments have the advantages that they are fairly easy to construct, are flexible in what they can assess, and are quick to administer. Moreover, like the other constructed-response types, fill-in assessments measure the students' abilities to actually produce language, albeit small amounts of language, and open up the possibility of assessing interactions between receptive and productive skills (for example, in a listening cloze, students must listen to a passage while reading it and filling in the blanks).

One limitation to fill-in assessment is that it is generally very narrowly focused on testing a single word or short phrase at most. Another problem is that a blank to fill in may have a number of possible answers. For instance, in the process of conducting one study (Brown, 1980), as many as 28 possible answers were found for a particular cloze test blank.

Short Answer

Short-answer assessments require the students to scrutinize a question or statement and respond with a one or more phrases or sentences. The advantages of short-answer assessments are that they are easy to produce and are relatively quick to administer. One disadvantage of short-answer assessments is that they focus on assessing a few phrases or sentences. A second disadvantage is that multiple answers are possible, which means that if the prompts are not carefully crafted, each student may produce a completely different answer.

Performance

Performance assessments require students to accomplish approximations of real-life, authentic tasks, usually using the productive skills of speaking or writing but also using reading or writing or combining skills. Performance assessments can take many forms, including fairly traditional tasks like essay writing or interviews or more recent developments like problem-solving tasks, communicative pair-work tasks, role playing, and group discussions.

In short, by definition, the performance assessment has three requirements: (a) Examinees are required to perform some sort of task, (b) the tasks must be as authentic as possible, and (c) the performances are typically scored by qualified raters. (For more on performance assessment in language testing, see Shohamy, 1995, and Wiggins, 1989.)

The principal advantage of performance assessments is that they can come close to eliciting authentic communication (at least insofar as authentic communication can be elicited in any testing situation). Advocates of performance assessments maintain that performance assessments provide more valid (a) measures of students' abilities to respond to real-life language tasks, (b) estimates of students' true language abilities than traditional standardized multiple-choice assessments, and (c) predictions of students' future performances in real-life language situations. Performance assessments can also be used to counteract the negative washback effects of standardized testing, like bias and irrelevant content. In fact, well-designed performance assessments can provide strong positive washback effects (see discussion below), especially if they are directly linked to a particular curriculum. (For details on the positive aspects of performance assessment, see the discussion in Norris et al., 1998.)

Performance assessments are relatively difficult to produce and relatively time-consuming to administer. Considerable costs may also be incurred in developing performance assessments, administering them, training raters, conducting rating sessions, reporting scores, and so on. Still another disadvantage is that logistics involve a number of complex issues like collecting and storing audio- or videotapes of the performances, providing special equipment and security, and planning and conducting rating sessions. Reliability may be problematic because of rater inconsistencies, limited numbers of observations, subjectivity in the scoring process, and so on. Validity may also be problematic because of (a) inadequate content coverage; (b) lack of construct generalizability; (c) the sensitivity of performance assessments to test method, task type, and scoring criteria; (d) construct underrepresentation (i.e., the problem of generalizing from a few observations to the whole spectrum of

real-life performances); and (e) construct-irrelevant variance (i.e., performance characteristics that have nothing to do with the students' real abilities). Test security may also be problematic because of a small number of prompts (each prompt may be very easy for examinees to remember and pass on to others), the difficulty of creating and equating new prompts for each administration, and the potential effects of teaching to the test. (For details on the negative aspects of using the performance assessment, see Educational Testing Service, 1995, and Norris et al., 1998.)

Personal-Response Assessments

Like constructed-response assessments, personal-response assessments require students to actually produce language, but personal-response assessments also allow each student's responses to be quite different. In a real sense, these assessments allow students to communicate what they want to communicate.

In general, personal-response assessments are beneficial in that they provide personal or individualized assessment, can be directly related to and integrated into the curriculum, and can assess learning processes in an ongoing manner throughout the term of instruction. However, personal-response assessments also have the general drawbacks of being relatively difficult to produce and organize and of involving subjective scoring. The most common types of personal-response assessments are conferences, portfolios, and self- and peer assessments.

Conferences

Conference assessments typically involve the student visiting the teacher's office, usually by appointment, to discuss a particular piece of work or learning process, or both. More importantly, conferences are different from other forms of assessment in that they focus directly on learning processes and strategies (Genesee & Upshur, 1996). For example, consider a series of conferences conducted to discuss multiple drafts of students' compositions. During the conferences, the focus could be on students' views and worries about the learning processes they are experiencing while producing and revising their compositions.

In total, the advantages of conferences are that teachers can use them to (a) foster student reflection on their own learning processes; (b) help students develop better self-images; (c) elicit language performances on particular tasks, skills, or other language points; or (d) inform, observe, mold, and gather information about students. Naturally, such advantages

are offset by certain disadvantages. In the case of conferences, the disadvantages are that they are relatively time-consuming, difficult and subjective to grade, and typically not scored or rated at all.

Portfolios

For decades, photographers, models, graphic artists, and practitioners of similar vocations have collected portfolios of their work in order to show their work and skills in a compact and convenient form. Recently, language teachers have begun using portfolios in order to encourage their students to select, compile, and display their work. We define portfolio assessments here as purposeful collections of any aspects of students' work that tell the story of their achievements, skills, efforts, abilities, and contributions to a particular class. However, several other definitions exist for this fairly new type of assessment, which might more aptly be called a family of assessments. (For other definitions, see Arter & Spandel, 1992; Brown & Wolfe-Quintero, 1997; Camp, 1993; Shaklee & Viechnicki, 1995; and Wolf, 1989.)

The literature reports at least three advantages for portfolio assessments. We see these advantages as falling into three categories: strengthening students' learning, enhancing the teacher's role, and improving testing processes.

Portfolio assessments may strengthen student learning in that they (a) capitalize on work that would normally be done in the classroom anyway; (b) focus learners' attention on learning processes; (c) facilitate practice and revision processes; (d) help motivate students, if well-planned, because they present a series of meaningful and interesting activities; (e) increase students' involvement in the learning processes; (f) foster student-teacher and student-student collaboration; (g) provide means for establishing minimum standards for classroom work and progress; and (h) encourage students to learn the metalanguage necessary for students and teachers to talk about language growth.

Portfolio assessments may enhance the teacher's role to the degree that they (a) provide teachers with a clearer picture of students' language growth, (b) change the role of the teacher (in the eyes of students) from that of an adversary to that of a coach, and (c) provide insights into the progress of each individual student.

Portfolio assessments may improve testing processes to the extent that they (a) enhance student and teacher involvement in assessment; (b) provide opportunities for teachers to observe students using meaningful language to accomplish various authentic tasks in a variety of contexts and situations; (c) permit the assessment of the multiple dimensions of language learning (including processes, responses, and activities); (d) provide opportunities for both students and teachers to work together

and reflect on what it means to assess students' language growth; (e) increase the variety of information collected on students; and (f) make teachers' ways of assessing student work more systematic. (For more on the advantages of the portfolio assessments, see Chittenden, 1991; Genesee & Upshur, 1996; LeMahieu, Eresh, & Wallace, 1992; Valencia, 1990; and Wolf, 1989.)

The literature also addresses at least five disadvantages of using portfolio assessments: the issues of design decisions, logistics, interpretation, reliability, and validity. Design decision issues include deciding (a) who will determine grading criteria, (b) how grading criteria will be established, (c) who will determine what the portfolios will contain, and (d) how much of daily authentic classroom activities will be included in the portfolios. Logistical issues involve finding (a) the increased time and resources needed to support portfolio assessments, (b) ways to rely on the training and abilities of teachers to implement portfolio assessments, and (c) the time for teachers to read and rate portfolios on a regular basis throughout the school year while simultaneously helping students develop those portfolios. Interpretation issues include (a) grading students' achievements as represented in their portfolios; (b) setting standards and interpreting the portfolios in a way that is equally fair to all students; (c) training teachers to make fair interpretations; and (d) reporting portfolio assessment results so that all interested audiences (e.g., students, parents, administrators, politicians) can understand them. Reliability issues involve (a) ensuring sufficient reliability across raters and occasions when ratings occur; (b) encouraging objectivity; (c) preventing mechanical errors, especially those that could affect decisions; (d) standardizing the rating and grading processes; and (e) ensuring equal access to resources for all students. Validity issues include (a) demonstrating the validity of the portfolios for the purposes of making decisions about students; (b) determining how adequately the portfolios exemplify students' work, development, and abilities; (c) identifying and controlling any potential intervening variables that might affect students' achievements; and (d) separating out which student abilities lead to which performance characteristics in what amounts. (For more details on the disadvantages of portfolio assessments, see Arter & Spandel, 1992; Camp, 1993; and Valencia & Calfee, 1991.)

Self- and Peer Assessments

Self-assessments require students to rate their own language, whether through performance self-assessments, comprehension self-assessments, or observation self-assessments. Performance self-assessments require students to read a situation and decide how well they would respond in

it. Recent examples of self-assessments of performance can be found in Hudson, Detmer, and Brown (1992, 1995) and Yamashita (1996). Similarly, comprehension self-assessments require students to read a situation and decide how well they would comprehend it. (For examples, see Bergman & Kasper, 1993, and Shimamura, 1993.) In contrast, observation self-assessments require students to listen to audio- or videotape recordings of their own language performance (perhaps taped in natural situations or in role-play activities) and decide how well they think they performed. Recent examples of observation self-assessments can be found in Hudson et al. (1995) and Yamashita (1996). A variant of the self-assessment is the peer assessment, which is similar to the self-assessment except that, as implied by the label, students rate the language of their peers.

Self-assessments have a number of advantages. First, self-assessments can be designed to be administered relatively quickly. Second, they inevitably involve students directly in the assessment process. Third, in turn, such involvement may help students understand what it means to learn a language autonomously. Finally, both the students' involvement and their greater autonomy can substantially increase their motivation to learn the language in question. (For more information about designing self-assessments, see Blanche, 1988; Blanche & Merino, 1989; Gardner, 1996; Hudson et al., 1992, 1995; McNamara & Deane, 1995; Oscarson, 1989; and Oskarsson, 1978.)

Self-assessments also have a number of disadvantages. For instance, Blanche (1988), in a comprehensive literature review, concluded that "the accuracy of most students' self-estimates often varies depending on the linguistic skills and materials involved in the evaluations" (p. 81). Both Blanche (1988) and Yamashita (1996) noticed that those students who were more proficient tended to underestimate their language abilities.

In addition, Blanche (1988) warned that "self-assessed scores may often be affected by subjective errors due to past academic records, career aspirations, peer-group or parental expectations, lack of training in self study, etc." (p. 81). Such subjective errors can probably be overcome to some degree if the scoring grids the students are using to rate themselves describe clear and concrete linguistic situations in which they are to consider their performance in terms of precisely described behaviors. However, such subjective errors may be difficult to surmount in some situations, that is, when the consequences of the self-assessment become an integral part of the assessment itself. For instance, a self-assessment might turn out to be quite successful when used for research purposes, but the same self-assessment might not function well at all in a higher stakes setting where students are asked to assess their placement into levels of study in a language program. Any students with

a vested interest in being exempted from study might rate themselves considerably higher in such a placement situation than they would in a research setting. (For examples of self-assessments used in real testing and research, see Bachman & Palmer, 1981; Bergman & Kasper, 1993; Davidson & Henning, 1985; Heilenman, 1990; and LeBlanc & Painchaud, 1985.)

FITTING ASSESSMENT TYPES TO CURRICULUM

Testing and curriculum very often do not match very well in language programs. To correct such a situation, three sets of issues are worth considering: the negative and positive consequences of washback, the significance of feedback, and the importance of using multiple sources of information.

The Negative and Positive Consequences of Washback

Washback is the effect of testing and assessment on the language teaching curriculum that is related to it. Washback is also called *backwash*, *test impact*, *measurement-driven instruction*, *curriculum alignment*, and *test feedback* in various places in the literature.⁴

Recently, Alderson and Wall (1993a) called into question the existence of washback, and rightly so, given that little if any actual research had ever demonstrated its existence. Alderson and Wall themselves discussed four studies that had empirically addressed the issue of washback in the past (Hughes, 1988; Khaniya, 1990; Wall & Alderson, 1996; Westdorp, 1982). More recently, a number of studies have further confirmed the existence and complex nature of the washback effect (e.g., Alderson & Hamp-Lyons, 1996; Shohamy, Donitsa-Schmidt, & Ferman, 1996; Wall, 1996; Watanabe, 1992, 1996a, 1996b). All in all, the empirical studies to date seem to confirm the existence of the washback effect in various places with a variety of different effects, but these studies also indicate that washback is not a simple or straightforward phenomenon that conforms neatly to popular notions about the effects of tests on language learning.

Washback effects can be either negative or positive. If the assessment procedures in a curriculum do not correspond to a curriculum's goals and objectives, the tests are likely to create a negative washback effect on those objectives and on the curriculum as a whole. For example, if a

⁴Dan Douglas once considerably lightened the mood of a very serious meeting at Educational Testing Service by referring to the washback effect as the *bogwash effect*.

program sets a series of communicative performance objectives but assesses the students at the end of the courses with multiple-choice structure tests, a negative washback effect will probably begin to work against the students' being willing to cooperate in the curriculum and its objectives. Students soon spread the word about such mismatches, and they will generally insist on studying whatever is on the tests and will ignore any curriculum that is not directly related to it. We have each seen this occur in numerous settings.

A positive washback effect occurs when the assessment procedures correspond to the course goals and objectives. For instance, if a program sets a series of communicative performance objectives and tests the students using performance assessments (e.g., role plays, interviews) and personal-response assessments (e.g., self-assessments, conferences), a powerful and positive washback effect can be created in favor of the communicative performance objectives. Positive washback occurs when the tests measure the same types of materials and skills that are described in the objectives and taught in the courses.

Language teachers can use the information given above about the advantages and disadvantages of selected-response assessments, constructed-response assessments, and personal-response assessments when designing course objectives. If teachers consider how those objectives will be assessed or observed at the end of the course and follow through by using the assessment format that best matches each objective, they will be helping to create a strong relationship between the assessment procedures and the objectives and therefore helping to produce a positive washback effect. (For more information on the washback effect, see Alderson & Hamp-Lyons, 1996; Alderson & Wall, 1993a, 1993b; Gates, 1995; Messick, 1996; Shohamy et al., 1996; Wall, 1996; Wall & Alderson, 1996; and Watanabe, 1996a. For summary articles, see Bailey, 1996, and Brown, 1997.)

The Significance of Feedback

The purpose of feedback will differ in different situations, but feedback is nonetheless important (see, e.g., Shohamy, 1992). For example, the purpose of feedback in the form of scores from a diagnostic pretest administered at the beginning of a course will be to inform students of their strengths and weaknesses vis-à-vis the knowledge or skills covered in the course. In other words, the scores will be interpreted diagnostically: A low score on a particular objective indicates that a student needs to work hard on that objective, and a high score on another objective shows that the student already has mastered the knowledge or skill involved in that objective (and that the student should

focus energy on other, weaker objectives). Thus in a diagnostic pretest, the feedback is given in terms of what the students need to do about each course objective.

On the other hand, with scores derived from an achievement test given as a posttest, the purpose of the feedback will be quite different. If the scores are referenced to the objectives of a particular course, they will be interpreted in terms of what the students have learned or learned how to do in the course. Thus a low score on a particular objective will indicate that the student did not get the knowledge or skills necessary to master that objective. Such a student may be advised to work hard on the perceived weakness or may be required to do remedial training on it. Alternatively, if some students have low scores on a number of objectives, the teacher may decide that those students should not be promoted to the next level or that they should fail the course and be required to take it again.

The decisions made with such test scores are often a matter of policy within a given institution, and the making of those decisions should relate directly to the curriculum: The feedback from the tests will not just be a number but will also provide an warning that the student did not achieve, say, Objectives 2, 3, 8, 11, and 13. Hence achievement tests provide feedback to the students in terms of what they have learned in the course and provide feedback that the teachers can use for grading.

Clearly, feedback is important in diagnostic and achievement testing, particularly in objectives-based testing (Brown, 1990, 1996). Students want to know how they did on a particular test. To the extent that feedback can be couched in terms more meaningful than a single score (e.g., by reporting subscores related to particular course objectives), that feedback can become an integral part of the learning process. Such integration is one of the particular strengths of the personal-response types of assessments described above. Conferences, portfolios, and self-assessments all provide rich forms of feedback to the students that can be integrated into their learning. But some mixture of different types of tests and feedback may prove best in a particular curriculum.

The assessment procedures used within a particular language program must be directly related to the curriculum if that feedback is to be maximally useful. In some programs that we have observed, scores on the Test of English as a Foreign Language (TOEFL) or Test of English for International Communication (TOEIC) serve as pretests and posttests for language courses as well as a means to assess student improvement (gain), teacher effectiveness, and so on. In the vast majority of cases, such tests will not be appropriate for such purposes. They are norm-referenced tests, which are by definition very general tests (Brown, 1996, pp. 2–8). Therefore, much of what is being tested on the TOEFL or TOEIC will not be directly related to the knowledge or skills that the

students are learning in a particular course. Moreover, such norm-referenced tests are very global in nature and are not designed to make the fine distinctions that would reflect the amounts and types of learning that take place during a single term in a single language course. Furthermore, such norm-referenced tests are not level specific, in the sense that the material tested is typically not at exactly the correct level of difficulty for the students involved in a particular course. Because the TOEFL and TOEIC must spread students out along a continuum of proficiency levels, these tests must have items with a wide variety of difficulty levels. As a result, many of the items on such a test will be too easy or too difficult for the students in a particular course, which means that those items are not appropriate for assessing performance in that course or for assessing the learning gains that students make in that course.

The Importance of Multiple Sources of Information

Basing any decision on a single source of information is dangerous and maybe even foolish. For instance, hiring a new teacher on the basis of a single recommendation letter would be foolish because that letter might be motivated by friendship with the teacher, by a desire to get rid of the teacher (due to incompetence), by a desire to make a particular MA program look good, or by any number of other factors. Generally, most teachers realize that multiple sources of information are more reliable than any single piece of information. Hence, administrators typically gather many different types of information about teachers when making hiring decisions. For example, in recent hires at the University of Hawai'i, the department has required three letters of recommendation, a résumé, graduate school transcripts, a personally written statement of teaching philosophy, a sample lesson plan, an interview with the director, a teacher portfolio (see Brown & Wolfe-Quintero, 1997), and even a live demonstration lesson for those teachers on the short list. The faculty feels that those multiple sources of information help make decisions about hiring much more dependable. Multiple sources of information are important to think about in selecting assessment strategies and in interpreting their results.

In Selecting Assessment Strategies

The general educational testing literature shows repeatedly that tests should be made up of a sufficient number of observations, or bits of information, to increase the chances that they will collectively be reliable. A one-item multiple-choice test would never seem fair or

reliable to any teacher or student. Intuitively, they would feel that a single-item test could never do a really good job of testing. That is why tests are usually made up of 40–50 items instead of just one. When thinking about the advantages and disadvantages of the various assessment types discussed above, especially which ones to select and how to use them in a particular curriculum, language teachers should remember that assessments based on multiple observations are generally more reliable than assessments based on a few observations. Hence, a single interview done once and providing a single score is likely to be less reliable than, say, the multiple scores of a video portfolio of oral work created and rated on multiple occasions over an entire semester. Similarly, an interview rated by one rater is less likely to be less reliable than a score on a composition rated by three raters.

In Interpreting Assessment Results

The use of multiple sources of information in designing and selecting assessments is also a key factor in interpreting assessment. One important type of decision that we make at the University of Hawai'i concerns the admission of thousands of international students. Although TOEFL scores are used in deciding whether an international student should be admitted, basing that decision solely on a single TOEFL score would be highly irresponsible. To get around this problem, we use other types of information, like the students' high school grade point average; statement-of-purpose essays; recommendation letters; transcripts of high school performance; and information about sports, clubs, and other extracurricular activities. These pieces of information along with the TOEFL scores help us make much more reliable admissions decisions. No responsible educator, least of all the testing professionals at Educational Testing Service, would advocate using a single test score in making important decisions because using multiple sources of information of varying types increases the collective reliability of that information and of any decisions that may result from interpreting it. As McNamara and Deane (1995) put it, "Using these complementary assessment tools—traditional measures and student self-assessment information—we have a more complete picture of our students' ability, effort, and progress" (p. 21).

CONCLUSION

In this article, we have organized and tried to expand the options that teachers have in classroom assessment. Teachers and testers might better be served by thinking of all types of language tests as alternatives in

assessment rather than viewing some types as being special. These alternatives, including selected-response, constructed-response, and personal-response assessments, each exhibit advantages and disadvantages. The consequences of the washback effect of assessment on curriculum, the significance of feedback in assessment, and the importance of using multiple sources of information in making important decisions are important considerations for choosing a means of assessment.

Tests are neither good nor evil in and of themselves. They are simple tools. Teachers must look with clear eyes at all of these tools as alternatives in assessments. They are by no means magical, but they are alternatives that teachers should consider within an overall framework of responsible assessment and decision making.

THE AUTHORS

James Dean (JD) Brown is Professor in the Department of ESL at the University of Hawai'i at Manoa. His research interests are language assessment, research design, and curriculum development. He is the author, coauthor, or editor of eight books on those topics. He has taught in the U.S., France, China, Saudi Arabia, and Japan.

Thom Hudson is Associate Professor in the Department of English as a Second Language at the University of Hawai'i at Manoa. His research interests are language assessment and testing, second language reading and literacy, and program development. He has taught in the U.S., Egypt, Mexico, and Japan.

REFERENCES

- Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13*, 280–297.
- Alderson, J. C., & Wall, D. (1993a). Does washback exist? *Applied Linguistics, 14*, 115–129.
- Alderson, J. C., & Wall, D. (1993b). Examining washback: The Sri Lankan impact study. *Language Testing, 10*, 41–69.
- American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: Author.
- American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: Author.
- Arter, J. A., & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice, 11*, 36–44.
- Aschbacher, P. A. (1991). Performance assessment: State activity, interest, and concerns. *Applied Measurement in Education, 4*, 275–288.
- Bachman, L. F., & Palmer, A. S. (1981). A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 149–165). Washington, DC: TESOL.
- Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*, 257–279.
- Bergman, M. L., & Kasper, G. (1993). Perception and performance in native and

- nonnative apology. In G. Kasper & S. Blum-Kulka (Eds.), *Interlanguage pragmatics* (pp. 82–107). Oxford: Oxford University Press.
- Blanche, P. (1988). Self-assessment of foreign language skills: Implications for teachers and researchers. *RELC Journal*, 19, 75–96.
- Blanche, P., & Merino, B. J. (1989). Self assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning*, 39, 313–340.
- Bond, L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. *Educational Measurement: Issues and Practice*, 14(4), 21–24.
- Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, 14(4), 9–12, 27.
- Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *Modern Language Journal*, 64, 311–317.
- Brown, J. D. (1990). Short-cut estimates of criterion-referenced test consistency. *Language Testing*, 7, 77–97.
- Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.
- Brown, J. D. (1997). Do tests washback on the language classroom? *TESOLANZ Journal*, 5, 63–80. Earlier version published as Brown, J. D. (1997). The washback effect of language tests. *University of Hawaii Working Papers in ESL*, 16(1), 27–45.
- Brown, J. D., & Wolfe-Quintero, K. (1997). Teacher portfolios for evaluation: A great idea or a waste of time? *The Language Teacher*, 21(1), 28–30.
- Burger, S. E., & Burger, D. L. (1994). Determining the validity of performance-based assessment. *Educational Measurement: Issues and Practice*, 13(1), 9–15.
- Camp, R. (1993). The place of portfolios in our changing views of writing assessment. In R. S. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing and portfolio assessment* (pp. 183–212). Hillsdale, NJ: Erlbaum.
- Chittenden, E. (1991). Authentic assessment, evaluation, and documentation of student performance. In V. Perrone (Ed.), *Expanding student assessment* (pp. 22–31). Alexandria, VA: Association for Supervision and Curriculum Development.
- Cohen, A. D. (1994). *Assessing language ability in the classroom*. Boston: Heinle & Heinle.
- Davidson, F., & Henning, G. (1985). A self-rating scale of English difficulty. *Language Testing*, 2, 164–179.
- Educational Testing Service. (1995). *Performance assessment: Different needs, difficult answers*. Princeton, NJ: Educational Testing Service.
- Gardner, D. (1996). Self-assessment for self-access learners. *TESOL Journal*, 5(1), 18–23.
- Gates, S. (1995). Exploiting washback from standardized tests. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 101–106). Tokyo: Japan Association for Language Teaching.
- Genessee, F., & Upshur, J. (1996). *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.
- Green, B. F. (1995). Comparability of scores from performance assessments. *Educational Measurement: Issues and Practice*, 14(4), 13–15, 24.
- Guion, R. M. (1995). Commentary on values and standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 25–27.
- Heilenman, L. K. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing*, 7, 174–201.
- Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

- Hudson, T., Detmer, E., & Brown, J. D. (1992). *A framework for testing cross-cultural pragmatics*. Honolulu: University of Hawai'i Press.
- Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics*. Honolulu: University of Hawai'i Press.
- Huerta-Macias, A. (1995). Alternative assessment: Responses to commonly asked questions. *TESOL Journal*, 5(1), 8–11.
- Hughes, A. (1988). Introducing a needs-based test of English language proficiency into an English-medium university in Turkey. In A. Hughes (Ed.), *Testing English for university study* (ELT Document No. 127, pp. 134–153). London: Modern English Publications.
- Jaeger, R. M. (1995). Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. *Educational Measurement: Issues and Practice*, 14(4), 16–20.
- Khaniya, T. R. (1990). The washback effect of a textbook-based test. *Edinburgh Working Papers in Applied Linguistics*, 1, 48–58.
- LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly*, 19, 673–687.
- LeMahieu, F. G., Eresh, J. T., & Wallace, R. C. Jr. (1992). Using student portfolios for a public accounting. *The School Administrator*, 49(11), 8–15.
- Linn, R. L. (1994). Performance assessment: Policy, promises and technical measurement standards. *Educational Researcher*, 23(9), 4–14.
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15–21.
- Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, 13(1), 5–8, 15.
- McNamara, M. J., & Deane, D. (1995). Self-assessment activities: Toward autonomy in language learning. *TESOL Journal*, 5(1), 17–21.
- McNamara, T. (1996). *Measuring second language performance*. London: Longman.
- Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5–8.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13, 241–256.
- Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Honolulu: University of Hawai'i Press.
- Oscarson [Oskarsson], M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing*, 6, 1–13.
- Oskarsson, M. (1978). *Approaches to self-assessment in foreign language learning*. Oxford: Pergamon Press.
- Shaklee, B. D., & Viechnicki, K. J. (1995). A qualitative approach to portfolios: The early assessment for exceptional potential model. *Journal for the Education of the Gifted*, 18, 156–170.
- Shimamura, K. (1993). *Judgment of request strategies and contextual factors by Americans and Japanese EFL learners*. Unpublished master's thesis, University of Hawai'i at Manoa, Honolulu.
- Shohamy, E. (1992). Beyond performance testing: A diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal*, 76, 513–521.
- Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, 15, 188–211.
- Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, 13, 298–317.

- Valencia, S. (1990). A portfolio approach to classroom assessment: The whys, whats, and hows. *The Reading Teacher*, 1, 338–340.
- Valencia, S. W., & Calfee, R. (1991). The development and use of literacy portfolios for students, classes, and teachers. *Applied Measurement in Education*, 4, 333–345.
- Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing*, 13, 234–354.
- Wall, D., & Alderson, J. C. (1996). Examining washback: The Sri Lankan impact study. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 194–221). Clevedon, England: Multilingual Matters.
- Watanabe, Y. (1992). Washback effects of college entrance examinations on language learning strategies. *JACET Bulletin*, 23, 175–194.
- Watanabe, Y. (1996a). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing*, 13, 318–333.
- Watanabe, Y. (1996b). Investigating washback in Japanese EFL classrooms: Problems and methodology. *Australian Review of Applied Linguistics*, 13, 208–239.
- Westdorp, H. (1982). *Backwash effects of language testing in primary and secondary education*. Unpublished manuscript, University of Amsterdam, Stichting Center for Educational Research.
- Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, 70, 703–713.
- Wolf, D. P. (1989). Portfolio assessment: Sampling student work. *Educational Leadership*, 46, 35–39.
- Yamashita, S. O. (1996). *Six measures of JSL pragmatics*. Honolulu: University of Hawai'i Press.
- Yen, W. M. (1997). The technical quality of performance assessments: Standard errors of percents of pupils reaching standards. *Educational Measurement: Issues and Practice*, 16(3), 5–15.