

LANGUAGE TESTING: PURPOSES, EFFECTS, OPTIONS, AND CONSTRAINTS¹

James Dean Brown

University of Hawai'i at Manoa

Abstract

The purpose of this paper is to examine the importance of the language testing and explore the choices that teachers and administrators must make in doing language testing well. The *importance of language testing* is examined in terms of the purposes and effects of testing as follows: (a) the purposes of language testing include two types of testing decisions (norm-referenced decisions for aptitude, proficiency, and placement, as well as criterion-referenced decisions for diagnosis, progress, and achievement); (b) the effects of testing are the positive and negative washback effects of tests on language curriculum particularly in analyzing students' needs, setting goals and objectives, developing program level tests, producing materials, delivering instruction, and evaluating program effectiveness. The *choices that teachers and administrators must make* in doing language testing are explored in terms of the options and constraints with which they must deal: (c) the options in language testing include selected-response (true-false, multiple-choice, and matching), constructed-response (fill-in, short-answer, and performance), and personal-response (conference, portfolio, and self/peer) types of tests; (d) the constraints in language testing are functional, political, and economic in nature. Ultimately, the purposes, effects, options, and constraints are different for administrators (who must make program-level decisions) and teachers (who are responsible for giving classroom level feedback and making pedagogical decisions). The paper ends with a brief discussion of the conflicts that can arise between administrators and teachers on testing issues and how such collisions can be resolved.

Introduction

As you may have read in the abstract for this talk, my overall purpose in this speech is to explore the importance of the language testing and consider the choices that teachers and administrators must make in doing a good job in their language testing. To those ends, I will organize the talk into two primary sections: one on the importance of language testing and the other on the choices that teachers and administrators must make in doing language testing.

¹ This paper was delivered as a plenary speech at the Sixth National Conference on Community Languages and English for Speakers of Other Languages in Palmerston North, New Zealand on 28 September 1998.

Turning first to the importance of language testing, I will frame the discussion within the larger contexts of the purposes and effects of language testing in general.

Purposes of language testing

In my view, the *purposes* of all language tests have to do with making decisions of one kind or another about students' lives. Such decisions can be subdivided into two overall categories: norm-referenced and criterion-referenced decisions (see for instance, Brown, 1988, 1989, 1990, 1995a, 1995b, or 1996).

NORM-REFERENCED TESTING PURPOSES

- Aptitude testing
- Proficiency testing
- Placement testing

CRITERION-REFERENCED TESTING PURPOSES

- Diagnostic testing
- Progress testing
- Achievement testing

Table 1: Purposes of norm-referenced and criterion-referenced testing

Norm-referenced decisions

As shown in Table 1, norm-referenced decisions are typically focused on the administrative decisions that we make about our students' language aptitude, proficiency, and placement.

Aptitude testing helps us to make decisions about who will most benefit from language teaching, or put another way, who will be the best investment, given limited resources, for language training. An example of an aptitude test is the *Modern Language Aptitude Test* developed by Carrol and Sapon way back in 1958 (which is of course long before any of us were born).

Proficiency testing most often helps us to make decisions about who has sufficient language ability to be admitted to our institutions. An example of a proficiency test is the TOEFL test battery, which is used to decide who can be admitted to many universities, especially in North America.

Placement testing most often helps us to make decisions about who should study in which level of language studies once they are admitted to our institutions. An example of a placement test is the ELIPT at the University of Hawaii, which we use to decide if students should be in our intermediate or advanced listening, reading, or writing courses, or should be exempted altogether.

For the most part, norm-referenced decisions about students' language learning aptitude, proficiency, and placement are the responsibility of administrators. While teachers often help in the processes of administering and scoring such tests, the decisions themselves are primarily made by administrators in order to manage the logistics of getting students properly situated in a language learning institution or passing students from institution to institution, both of which are administrative concerns rather than pedagogical decisions.

Criterion-referenced decisions

In contrast, criterion-referenced decisions are centered on the pedagogical issues of diagnosis, progress, and achievement (also shown in Table 1). *Diagnostic testing* helps us make decisions about what students already know and what they still need to learn. For instance, when I was teaching in China during the 1980s, our courses had clearly defined instructional objectives that we tested at the beginning of the course to determine each student's strengths and weakness so as to better tailor our courses to their needs.

Progress testing is similar to diagnostic testing, except that it helps us to make decisions about what students have learned and what they still need to learn so we can adjust the curriculum to their needs as the course progresses. For example, when I was teaching in China, we would also test the students at the five-week point, the midpoint in our 10 week courses. We did so to see how well the students were mastering the objectives of our courses. Our purpose was to make any adjustments in focus that would help them to better learn the material so as to meet the course objectives.

Achievement testing is similar to progress testing, except that it is typically done at the end of a course of study and it helps us to make decisions about what students have learned, which students should pass the course, and which grades should be assigned to each student's work or achievement in the course. For example, when I was teaching in China, we would test the students at the end of our 10 week courses to see how well they had learned the content and skills taught in our courses. Since we did no grading in our program, the purpose of our achievement tests was to make pass-fail decisions about our students and to learn what we could about the appropriateness of the overall syllabuses of our courses as well as the effectiveness of each and every objective. We could then make decisions about any

adjustments in focus that would help future students to better learn the material so as to meet the course objectives.

Effects of language testing

Now, I would like to turn to the *effects* of testing. Testing can affect all elements of a language curriculum. Such effects of testing on curriculum are often referred to as washback effects. I will begin here with some general discussion of the nature of washback effects and then turn to some of the more salient effects of testing on curriculum.

Washback effects in general

What exactly are washback effects? As I noted earlier, tests are used to make a variety of different types of decisions. In making such decisions, one of our many responsibilities as language teaching professionals is to recognize the effects that our tests and related decisions are having on everybody involved, including students, teachers, administrators, parents, politicians, and any other stakeholders in the language education process. These are the effects that are referred to as *washback effects*. Shohamy, Donitsa-Schmidt, and Ferman (1996) define washback as “the connections between testing and learning” (p.298), while Gates (1995) defined it as the “influence of testing on teaching and learning” (p.101). Most teachers will recognize that the washback effects of tests can be both negative and positive (as shown in Table 2).

NEGATIVE WASHBACK CAN AFFECT:

- Teaching
- Course content
- Course characteristics
- Class time

POSITIVE WASHBACK CAN BE FOSTERED BY MODIFYING:

- Test design factors
- Test content factors
- Test logistics factors
- Test interpretation factors

Table 2: General effects of washback on curriculum

Negative washback

As I explained in Brown (1997, 1998a), *negative washback* can influence teaching, course content, course character, and class time (synthesized from Alderson & Hamp-Lyons, 1996; Bailey, 1996; and Shohamy et al, 1996). Let's consider each of those negative effects in slightly more detail:

1. Teaching is affected when tests cause teachers to do any of the following: (a) narrow the curriculum, (b) stop teaching new material and instead review test related material, (c) replace course textbooks with worksheets based on previous tests, or (d) teach unnaturally.
2. Course content is affected when tests cause students to: (a) learn "examination-ese", (b) practice items similar to those on the test, (c) learn test-taking strategies in class, or (d) study grammar and vocabulary (while excluding other important aspects of language).
3. Course characteristics are affected when tests cause the inclusion in courses of: (a) any inappropriate language learning and language use strategies, (b) reduced emphasis on skills that require complex-thinking or problem-solving skills, (c) emphasis on raising exam scores without providing the language needed to interact in future overseas situations, or (d) a tense classroom atmosphere.
4. Class time is affected when: (a) test-preparation classes replace language learning classes, (b) test review sessions are added to regular class hours, (c) classes are skipped by students so they can study for the test, or (d) instructional time is lost.

Positive washback

Naturally, testing can also have *positive washback* effects. As I explained in Brown (1997), positive washback can be fostered by modifying factors like test design, test content, logistics, and test interpretation (synthesized from Bailey, 1996; Heyneman & Ransom, 1990; Hughes, 1989; Kellaghan & Greaney, 1992; Shohamy, 1992; and Wall, 1996). Let's consider each of those positive effects in slightly more detail:

1. Test design factors might involve any or all of the following: (a) making a test criterion-referenced, (b) building a test to measure specific teaching points, (c) constructing a test according to sound theoretical principles, (d) basing criterion-referenced tests on course objectives, (e) using direct tests, or (f) using self-assessment and learner autonomy.
2. Test content factors involve: (a) testing those abilities you want to encourage, (b) emphasizing open-ended items rather than selected-response items (like true-false, multiple-choice, etc.), (c) making tests reflect the full curriculum, not just a small portion, (d) assessing higher-order cognitive skills so they will be taught, (e) using a variety of testing formats, (f) expanding the skills that are tested to include non-academic out-of-school tasks, and (g) using authentic texts and tasks.
3. Test logistics factors include: (a) seeing that all interested parties understand the purpose of the test, (b) insuring that learning goals are clear, (c) helping teachers themselves to

understand the test results, (d) providing feedback to teachers so that meaningful change can be take place, (e) providing detailed and timely feedback to schools on students' scores, (f) making sure teachers and administrators are involved in all phases of the test development and administration, and (g) providing detailed score reports.

4. Test interpretation factors include: (a) making sure the test results are credible and fair to test takers and score users, (b) considering factors other than teaching effort in judging examination results, (c) conducting predictive validity studies of tests to make sure they are fulfilling their purposes, (d) improving the professional competence of test developers, especially in test design, (e) insuring that each testing group has the research capacity to investigate the impact of tests on teaching (among other things), (f) having test developers work closely with curriculum developers and administrators, and (g) developing professional networks to share common concerns and interests.

Specific Effects of Testing on Curriculum

Here, I will turn to the more specific effects of testing on how teachers and administrators analyze students' needs, set goals and objectives, develop program-level tests, produce materials, deliver instruction, and evaluate program effectiveness. These six elements of curriculum development (shown in Figure 1) are covered in considerably more depth in Brown (1995a).

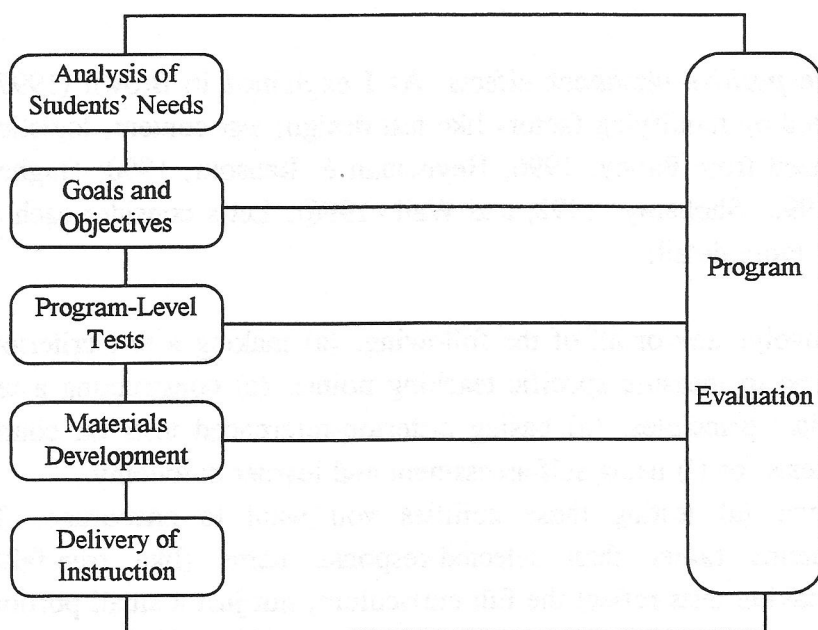


Figure 1: Elements of Curriculum Design (adapted from Brown, 1995a, p.20)

Effects on analysis of students needs

Whether analyzing linguistic or situational needs (for an explanation of the difference, see Brown, 1995a), the entire needs analysis process can and should be informed by test scores. Aptitude tests can be used to help us understand who would be the best investment for language training, or at least who would benefit most from such training. Proficiency tests can help us understand the general outlines of the ability levels of the students in a program particularly in terms of the ranges of overall language proficiency involved. For example, at the University of Hawai'i at Manoa (UHM), we know the range of students' abilities will fall between 500 and 600 on the TOEFL because the students must have 500 to be admitted to the university and they are automatically exempted from any ESL training once they attain 600 on that test. In addition, our placement tests help us create homogeneous groups of students with relatively specific ability levels within skill or content areas thereby helping us to define and meet the students' needs.

Effects on goals and objectives

Criterion-referenced diagnostic, progress, and achievement tests can help us understand our course objectives by forcing us to operationalize those objectives in an observable, measurable way. Criterion-referenced diagnostic tests in particular can help at the beginning of a course in determining the degree to which objectives we have set for a course on the basis of needs analysis are actually needed by the students. For instance, if 95% of the students score very high on a particular objective-based subtest at the beginning of a course, they clearly do not need to study that objective further, and so it can be eliminated from the curriculum. Given such information, we can then select those objectives that students do need to learn and eliminate those objectives that students do not need to learn thereby making the structure of the course much more efficient.

Effects on program-level tests

Systematic development, administration, and analysis of norm-referenced placement tests and criterion-referenced diagnostic, progress, and achievement tests can lead to revision and improvement of the effectiveness of the various types of tests through item analysis techniques (for instance, see Chapter 3 of Brown, 1996). I would like to emphasize that I am not necessarily talking about multiple-choice tests here, or even about pencil-and-paper tests; I am talking about all types of tests including: selected-response (true-false, multiple-choice, & matching), productive-response (fill-in, short-answer, & performance), and personal-response (portfolios, conferences, and self/peer-assessments). [See the **Options in Language Testing** section below for further explanation, or Brown (1998b) for many examples of the wide variety of different types of language testing procedures available to language teachers].

Effects on materials development

I argue that materials development can be made much more efficient through prudent use of tests. If course objectives are tested in a diagnostic test before the materials are actually written, the materials developer(s) may save a great deal of energy by not having to develop materials for those objectives that students have already learned. Also important is the idea that developing the tests and other observation strategies can help focus the materials developers on the types of exercises that students will need to practice and master in order to demonstrate that they have mastered the course objectives by passing the final achievement test. Yes, such thinking will probably lead to "teaching to the test", but if the teachers are judiciously involved in the test development processes and the tests are good ones, why is teaching to the test a problem?

Effects on delivery of instruction

I believe that teaching can also be improved through judicious use of testing procedures. Criterion-referenced tests can help us understand which objectives students already know and which they need to focus on in three ways: (a) diagnostic tests can help students and teachers allocate resources where they are needed at the beginning of the course, (b) progress tests can help in monitoring the gain that students have made up to a certain point in a course, and (c) achievement tests help in understanding how much students have learned overall in a course. In addition, criterion-referenced progress and achievement tests (especially when combined with the diagnostic results) provide us with curriculum development information about how well students are learning the course objectives. This information can show us which parts of the curriculum are and are not being effectively imparted to the students, and in turn, can lead to changes in teaching strategies that will increase the effectiveness of materials, teaching methods, tests, materials development, and so forth.

Effects on program evaluation

Program evaluation comes in two forms: summative evaluation and formative evaluation. Typically, summative program evaluation compares diagnostic tests with achievement tests in a pretest/posttest manner, which can help us to understand how much students have learned in a particular course. Such information can then be used to defend the program from outside political influences or funding changes. In contrast, formative program evaluation typically draws on a variety of information sources to revise/reform curriculum. Formative program evaluation will benefit from using tests to regularly study the adequacy of needs analyses, the usefulness of goals and instructional objectives, the effectiveness of the tests themselves, the strengths and weaknesses of the materials, and the quality of instruction. I argue that good testing is at the heart of any effective program evaluation effort, whether summative or formative.

Clearly then, language testing has many purposes and effects within any well-organized language program. But, given all that, what are the choices that teachers and administrators must make with regard to testing?

Choices that teachers and administrators must make

The choices that teachers and administrators must make in doing language testing will be explored here in terms of the options that language professionals have in language testing and the constraints that they must face.

Options in language testing

In my experience, language teaching professionals often have negative attitudes toward tests in general. My guess is that they feel this way because they associate testing mostly with multiple-choice testing or true-false testing. Since most language teachers recognize that real language, as it is used in real life, is not multiple-choice, their distrust of tests would seem to be justified. However, when I talk about testing, I am not restricting myself to multiple-choice or any other single type of test. In fact, the options are so numerous that, in order to make them clear, I will subdivide them into three categories (after Brown and Hudson, in press & 1998): selected-response, constructed-response, and personal-response.

RESPONSE TYPE examples	ADVANTAGES	DISADVANTAGES
<i>Selected-response types</i> true-false multiple-choice matching	Quick to administer; Scoring is fast and easy; scoring is objective	Relatively difficult to construct; No productive language
<i>Constructed-response types</i> fill-in short-answer performance	Guessing not a major factor; Measures productive language use; Measures the interaction of receptive and productive skills	Bluffing is possible; Scoring is difficult, time-consuming, and subjective
<i>Personal-response types</i> conferences portfolios self/peer assessments	Personal aspect to assessment; Integrated into and part of curriculum; Can assess learning processes	Difficult to produce and organise; Scoring is subjective

Table 3: Options in classroom assessment

Selected-Response

Selected-response tests are those that require students to circle the correct answer by filling in an oval, or otherwise mark the correct answer. Selected-response tests, as they are defined here, can all be classified into three categories: true-false, multiple-choice, or matching tests. As shown in Table 3, the advantages of selected-response tests are that they are relatively fast to administer, scoring them is relatively quick, accurate, and easy, and scoring them is objective (even a machine can usually do it). Unfortunately, selected-response tests also have disadvantages. For instance, writing selected-response test questions is relatively difficult, and they require absolutely no productive language abilities from the students. Nonetheless, true-false, multiple-choice, and matching can be useful for testing things like knowledge of grammar and vocabulary, or the receptive skills of reading and listening (or combinations of listening and reading).

Constructed-Response

Constructed-response tests are those that require students to supply the correct answer by filling in a word, phrase, or short answer. At more advanced levels of study, constructed-response tests can become even more elaborate and require students to perform some oral or written task. Constructed-response tests as they are defined here can all be classified into three categories: fill-in, short-answer, or performance tests. As shown in Table 3, the advantages of constructed-response tests are that guessing is not as big a problem as for other types of tests, they can be used to measure productive language use, and they can be used to assess the interaction of receptive and productive skills. The disadvantages of constructed-response tests are that bluffing is a very real possibility (that is, the students can produce an answer without actually knowing the correct response through clever use of avoidance or other strategies), scoring them is relatively difficult (especially if there is more than one possible answer), administering and scoring them is time-consuming, and scoring them is sometimes quite subjective. Nevertheless, fill-in, short-answer, performance tests can be useful for testing the productive skills of speaking and writing (or combinations of reading, writing, listening, and speaking).

Personal-Response

Like constructed-response tests, personal-response tests may require students to actually produce language, but personal-response tests also allow for the responses to vary from student to student. In a real sense, personal-response tests allow students to communicate what they want or need to communicate. The most commonly used types of personal-response tests to date are conferences, portfolios, and self/peer assessments. As shown in Table 3, the advantages of personal-response tests are that they add a personal, or authentic, aspect to the testing process, they can be integrated directly into the curriculum, and they can be used to assess learning processes. The disadvantages of personal-response tests are that they are difficult for teachers to plan, they are relatively difficult for students to produce, they must be

carefully organized, and scoring them is relatively subjective. Nonetheless, conferences, portfolios, and self/peer assessments can be useful for motivating students to produce language, and are useful for assessing all four skills as well as higher order organizational and thinking skills.

Constraints in Language Testing

Recently, I was reading an article by Cronbach (1988) in which he discusses five relatively new perspectives that he said we should consider in our arguments for the validity of any test. While reading it, I began to associate his perspectives with constraints that interfere with or at least minimize the effectiveness of language tests. As I thought about it, I began to realize that these *constraints* fell into basically three categories: functional, political, and economical constraints.

Functional Constraints

According to Cronbach (1988, pp.5-6), testers have traditionally worried about the truthfulness of their score interpretations. In other words, they have sought the answer to the following question: To what degree are the test scores representative of the construct(s) being tested? More recently, testers have begun to also consider the worth of their score interpretations. In other words, we are seeking the answer to the following question: To what degree are the decisions being made with the scores worthy and whose values are they based on? Clearly test developers today are duty bound to consider both the truthfulness and worth of their score interpretations. However, we must all recognize that the two are not necessarily related in a direct and clear manner. For example, the degree of truthfulness of a criterion-referenced classroom test might be examined by studying the match between the test items and the associated course objectives. However, the worth of that criterion-referenced test might lie "in its contribution to the learning of students working up to the test, or to next year's quality of instruction." (Cronbach, 1988, p.5). Truthfulness and worth may be associated, but then, they may not.

Also of functional concern, test developers must recognize that examinations have a sort of built-in conservatism in that the construct(s) that are being tested are defined and operationalized at a specific point in time. Theories of language teaching and learning are changing constantly and rapidly. Consider for instance the grammar/translation and task-based communicative teaching movements that are separated by only a few decades historically and certainly coexist in the world today. Given the state of the art in language teaching, the ideas of truthfulness and worth are likely to change over time (or over geography), depending on differences in construct definitions or in the social norms for construct definitions. Thus, test developers of all kinds (including teachers, administrators, and professional testers) must always be mindful of the consequences of their tests for the people and institutions that are

affected by the results as well as any potential conflicts in value systems that may arise among the various stakeholders in a particular situation (after Messick, 1980).

Political Constraints

Other constraints on testing that we often overlook are the political constraints. These constraints are important in my view because all language testing decisions are essentially political. They are political within the language teaching institution for two reasons: (a) because we are making decisions (aptitude, proficiency, placement, diagnosis, progress, and achievement) that are important to our students, their parents, and the students' futures; and (b) the teachers and administrators may vary in their views of how those decisions should be made and what they should be based on, both between those two groups and within. In addition, such decisions may become political in the more general sense of the word if they are brought to the attention of politicians, the media, the general public, etc.

Why is the political nature of testing important for us to consider? Kleiman and Faley (1985) pointed out that, if professional test developers (whether teachers, administrators, or researchers) do not explain their testing practices and results adequately, the nonprofessionals (i.e., the politicians, the public, the students, their parents, etc.) may take over the decision making processes without professional help. Typically, fairness is the central issue, but the degree to which students are being treated fairly and the very definition of the word *fairness* are both necessarily political decisions in their own right. We owe it to ourselves, as language teaching professionals, to shape those decisions about fairness by supplying the best quality tests available, understanding the testing information ourselves, and explaining test results clearly to anybody who may ultimately be involved in decision making.

Economic Constraints

Cronbach (1988, pp.9-12) couches his discussion of economic perspectives on validity in terms of employment testing, focusing especially on classifications and the making of qualitative judgments in real-life employment decisions. In my view, we must broaden our conceptualization to encompass economic constraints of all kinds including at least: the institutional costs of testing, the costs that we pass on to the students and their parents, the hidden costs of test preparation, as well as the costs of unsuitable or unreliable testing in terms of bad decisions that are made or inefficient learning processes.

The costs of testing burden students and parents in a variety of ways, but most often such costs take the form of test fees, the costs of test preparation courses, remedial courses to improve test scores, and so forth. Other emotional and psychological costs may exist as well, but they are beyond the scope of this paper. Here, I am referring only to the monetary costs of inappropriate testing in terms of the costs of bad decisions and inefficient learning processes.

For instance, students may incur costs in the form of extra tuition costs (and class time spent) because of a badly constructed achievement test that they flunked (and thereby had to repeat the course). In short, the economic constraints on language testing seem to me to be important, yet they appear to have been completely overlooked in the language testing literature.

Discussion

In my experience, the purposes, effects, options, and constraints that I have discussed briefly here are viewed quite differently by people in different educational roles, especially administrators and teachers. Since I spent seven years as an ESL/EFL teacher before I became a professor at UHM and since I have also spent seven years as an administrator in the EFL and ESL programs, I think I have come to understand both points of view. In my opinion, the purposes, effects, options, and constraints of language testing are viewed quite differently by these two groups because administrators necessarily concern themselves with making program-level decisions, while teachers are more often responsible for giving classroom level feedback and making pedagogical decisions. This fundamental difference in responsibilities seems to lead to different, and sometimes conflicting points of view on what testing is and how it should be used. I will end this paper by exploring those differences briefly. Table 4 compares what I think are the typical administrators' and teachers' points of views with regard to purposes, options, effects, and constraints of language testing.

Purposes

As I mentioned above, the first way in which administrators and teachers vary is in the types of decisions they must make with test scores. The work of administrators typically leads them to be more interested and concerned about program-level grouping decisions or about inter-institutional comparisons based on tests. Hence, administrators will naturally take most interest in aptitude, proficiency, and placement decisions, as well as the types of testing procedures that such decisions are commonly based on. You may have noticed that those three types of tests were described at the beginning of this talk as norm-referenced tests. In my view, administrators are most often interested in norm-referenced testing because of the types of decisions they must make and also because of their training.

The work of teachers typically leads them to be more interested and concerned about classroom-level learning and pedagogical decisions based on tests. As a consequence, they focus most often on diagnostic, progress, and achievement decisions, as well as the types of testing procedures that such decisions are usually based on. As indicated at the beginning of this speech, such classroom decisions are typically based on criterion-referenced tests.

CATEGORY Subcategory	ADMINISTRATORS	TEACHERS
PURPOSES		
Decisions	Program-level grouping decisions	Classroom level learning pedagogical decisions
Types of tests used	Aptitude, proficiency, & placement Focus on norm-referenced testing	Diagnostic, progress, & achievement Focus on criterion-referenced testing
EFFECTS		
Washback	Most often worry about the negative washback effects of test design, test content, logistics, & test interpretation, but should be concerned with using those factors for positive washback	Rightly worry about the negative washback effects of testing on teaching, course content, course characteristics, & class time
Effects on curriculum	Norm-referenced perspective tends to keep administrators focused on need analysis, goals setting, program-level testing, & program evaluation	Criterion-referenced perspective tends to keep teachers interested in objectives setting, course-level testing, materials development, & delivery of instruction
OPTIONS		
Scale	Scale of testing large in terms of numbers tested	Scale of testing relatively small in terms of numbers tested
Emphasis	Institutional orientation and perspective	Pedagogical orientation and perspective
Options of most interest	Most interested in selected-response options (true-false, multiple-choice, & matching)	More willing to consider options like constructed-response (fill-in, short-answer, and performance) and personal-response (conference, portfolio, and self/peer)
CONSTRAINTS		
Functional constraints	Content and truthfulness of tests are still main concerns Conservative with regard to changing testing domains	Have always been more interested in the worth of a test (especially, in the contribution of a test to student learning) and have always been painfully more aware of the consequences of testing; More progressive in terms of changing testing domains
Political constraints	Politics of testing is basically about fairness, to administrators fairness means making accurate decisions about student groupings (aptitude, admissions, placement)	Fairness means giving all students the same opportunities to learn and achieve (diagnosis, progress, achievement)
Economic constraints	Costs are considered very high for erroneous "high stakes" decisions like aptitude, proficiency, & placement in terms of time & money wasted by students doing unneeded courses, or students failing because they find themselves in over their heads Resources for norm-referenced test development are usually found to accomplish these types of decisions from institution, students, parents	Costs are considered low for erroneous "low stakes" decisions like diagnosis, progress, & achievement, but are just as important cumulatively in terms of time & money wasted studying material students already know or failing because of automatic promotion The costs of criterion-referenced test development are considered minimal because they are born by the teachers

Table 4: Administrators vs. teachers views on language testing

Effects

Administrators most often worry about the negative effects of washback due to test design, test content, logistics, and test interpretation factors. In my view, however, they should be more concerned about using those factors to create positive washback effects. Typically, it is only administrators who are in a position to affect the sorts of norm-referenced tests that create such negative washback effects. In the curriculum development area, the administrators' norm-referenced perspective also tends to keep them focused on aspects like needs analysis, goals setting, program-level testing, and program evaluation.

Teachers are more likely to worry about the negative washback effects of testing on teaching, course content, course characteristics, and class time. In terms of curriculum development, teachers' natural predilection for criterion-referenced testing procedures tends to keep them interested in objectives setting, course-level testing, materials development, and delivery of instruction.

Options

Typically, administrators are concerned with large scale testing, both in terms of the numbers of people being tested and the numbers of test items being used. In addition, administrators usually have an institutional orientation. As a result, they are often most interested in selected-response options (true-false, multiple-choice, & matching) because they are relatively easy to administer and score.

Teachers on the other hand, are usually dealing with testing on a relatively small scale in terms of both the numbers of students to be tested and the number of test items to be used. Furthermore, they tend to have a more pedagogical orientation. Hence, teachers are typically more willing to consider options like constructed-response testing (fill-in, short-answer, and performance) and personal-response testing (conference, portfolio, and self/peer).

Constraints

As I pointed out earlier in this speech, constraints can be functional, political, or economic. With regard to functional constraints, administrators tend to be more interested in the content and truthfulness of tests in terms of test validity. I believe that administrators also tend to be conservative with regard to changing their definitions of testing domains, which naturally leaves tests like the TOEFL in the dark ages with regard to the domains being tested. In contrast, I believe that teachers have always been more interested in the worth of a test, especially in terms of the contribution of a test and the experience of taking a test to student learning. In my view, teachers have for years also been painfully aware of the consequences of testing in terms of what actually happens to their students as people. Perhaps as a consequence

of dealing with such pedagogical concerns, teachers seem to me to be more progressive and flexible in their view of which domains should be tested and how they should be tested.

With regard to political constraints, as I pointed out above, the politics of testing is basically about fairness. To most administrators, fairness means making accurate decisions about student groupings, which is to say fairness is making aptitude, proficiency, and placement decisions in the same way for all students. To teachers, I think fairness is more often about giving all students the same opportunities to learn and achieve, which means providing all students with diagnostic feedback, progress reports, and achievement testing that is based directly on what they have been learning.

With regard to economic constraints, administrators seem to me to be most worried about the very high costs of erroneous "high stakes" decisions like aptitude, proficiency, and placement decisions in terms of time and money wasted by students doing unneeded course work, or students failing because they find themselves in "over their heads" in their studies. Because of this perceived importance of administrative decisions, resources for norm-referenced test development are almost always found to accomplish these types of decisions (either from the institution, students, or their parents). In contrast, teachers are typically thought of as being more worried about the relatively low costs of erroneous "low stakes" decisions like diagnosis, progress, and achievement decisions. However, the notion that these are "low stakes" decisions may be erroneous because such decisions may be cumulatively very important in terms of time and money wasted by students who end up studying material they already know or students who fail because they have been "automatically" promoted to a level above their abilities. Because of the perceived lack of importance of such classroom decisions, the costs of criterion-referenced test development are seldom supported by anybody but the teachers. Perhaps this last economic constraint is or should be a major issue in language teaching because better economic support for classroom tests might well lead to better tests, which in turn would no doubt lead to much better teaching and learning.

Conclusions

Clearly then, major differences exist between the views of administrators and teachers on the purposes, effects, options, and constraints of language testing. One important issue that needs to be addressed is what administrators and teachers can do to reconcile those differences in ways that will most benefit the students that all parties claim to care about. The answers that immediately spring to mind sound so cliché as to be laughable. But alas, I think it is true that the only way for administrators and teachers to reconcile their different views on testing and resolve any collisions that result from those differences is for them to: (a) recognize their

differences, (b) open clear channels of communication, (c) foster mutual understanding, and (d) (somehow or other) maintain a spirit of cooperation.

To those ends, I feel administrators should take an interest in what is happening in the classrooms and even try to organize teachers so they can work together on curriculum projects in general and testing projects in particular. When teachers work in isolation, they are not only less productive, but also more prone to *teacher burnout* and leaving the field.

If administrators will not take the lead, then the teachers themselves need to do so. If they can manage to work together on tests for courses that they teach in common, or at least agree to proofread each others tests, they will at least be avoiding the trap of working in isolation and will be taking a step in the right direction. To paraphrase a comment Charles Alderson once made in a conference at RELC, testing is far too important to be left to the testers. I might add that testing is far too important to be left to the administrators as well. Indeed, testing is probably too important to be left to any single group of people; their points of views on the purposes, effects, options, and constraints of language testing are far too diverse.

I would like to end today by thanking the organizers of the CLESOL Conference for inviting me to speak here in Palmerston North. New Zealand has been every bit as friendly and hospitable as I had heard it would be. Thank you again. It has been a pleasure and privilege to get to know you all.

References

- Alderson, J.C. & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing*, 13, 280-297.
- Bailey, K.M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, 13, 257-279.
- Brown, J.D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. London: Cambridge University Press.
- Brown, J.D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly*, 23(1) 65-83. [An earlier version appeared in 1988 *University of Hawaii Working Papers in ESL*, 7(1), 239-260.]
- Brown, J.D. (1990). Where do tests fit into language programs? *JALT Journal*, 12(1), 121-140.
- Brown, J.D. (1995a). *The elements of language curriculum: A systematic approach to program development*. New York: Heinle & Heinle Publishers.
- Brown, J.D. (1995b). Differences between norm-referenced and criterion-referenced tests? In J.D. Brown & S.O. Yamashita (Eds.), *Language Testing in Japan* (pp. 12-19). Tokyo: Japan Association for Language Teaching.

- Brown, J.D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.
- Brown, J.D. (1997). Do tests washback on the language classroom? *TESOLANZ Journal*, 5, 63-80. [An earlier version appeared in 1997 under the title "The washback effect of language tests" in *University of Hawaii Working Papers in ESL*, 16(1), 27-45.]
- Brown, J.D. (1998a). University entrance examinations and their effect on English language teaching in Japan. In J. Kahny & M. James (Eds.), *Perspectives on Secondary School EFL: A publication in commemoration of the 30th anniversary of the Language Institute of Japan* (pp. 20-27). Odawara, Japan: Language Institute of Japan.
- Brown, J.D. (Ed.). (1998b). *New ways of classroom assessment*. Washington, DC: Teachers of English to Speakers of Other Languages.
- Brown, J.D., & Hudson, T. (1998). The alternatives in language testing: Advantages and disadvantages. *University of Hawai'i Working Papers in ESL*, 16(2), 79-103.
- Brown, J.D., & Hudson. (In press). Alternatives in language assessment. *TESOL Quarterly*.
- Carroll, J.B., & Sapon, S.M. (1958). *Modern language aptitude test*. New York: The Psychological Corporation.
- Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Gates, S. (1995). Exploiting washback from standardized tests. In J.D. Brown & S.O. Yamashita (Eds.), *Language testing in Japan* (pp. 101-106). Tokyo: Japanese Association for Language Teaching.
- Heyneman, S.P., & Ransom, A.W. (1990). Using examinations and testing to improve educational quality. *Educational Policy*, 177-192.
- Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.
- Kellaghan, T., & Greaney, V. (1992). *Using examinations to improve education: A study of fourteen African countries*. Washington, DC: The World Bank.
- Kleiman, L.S. & Faley, R.H. (1985). The implications of professional and legal guidelines for court decisions involving criterion related validity: A review and analysis. *Personal Psychology*, 38, 803-833.
- Messick, S. (1980). Test validation and the ethics of assessment. *American Psychologist*, 35, 1012-1027.
- Shohamy, E. (1992). Beyond performance testing: A diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal*, 76(4), 513-521.
- Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, 13, 298-317.
- Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing*, 13, 234-354.