JALT Journal, Vol. 12, No.1 (May 1990)

Where Do Tests Fit Into Language Programs?

James Dean Brown
University of Hawaii at Manoa

The central thesis of this paper is that testing should be used to guide the development of a sound language program. The paper begins by examining the four types of decision making processes used in any language teaching institution: (a) who should be admitted into the program; (b) which level is appropriate for each student; (c) what should be taught so that resources are maximally utilized; (d) which students should be promoted to the next level? In order to help staff make these decisions, four categories of tests (one for each type of decision) are discussed: proficiency, placement, diagnostic, and achievement. Each is examined in terms of the purpose of the decision (norm-referenced or criterion-referenced), and the type of information that it provides. The paper concludes with a discussion of a hypothetical case study. The main point is that testing is not the enemy. On the contrary, tests can provide guidance in making informed and responsible decisions.

言語教育におけるテストの位置付け

本研究は、主に、言語プログラムの発展・充実に際して、テストが活用されるべきである、ということについて論ずるものである。まず、いかなる教育機関においても決定を下す過程で採用し得る、以下に掲げる4つの事項について検討をする。

1．そのプログラムに入れるのは誰であるべきなのか。
2．個々の学生に適したレベルはどれなのか。
3．学生の力を最大限に伸ばすためには何を教えるべきなのか。
4．上のレベルに進めるのはどの学生であるべきなのか。

これらの決定を下す際の参考に供するために、実力(proficiency)、クラス分け（placement)、診断（diagnostic)、成就（achievement）の4種類のテスト（1つの事項に1つのテスト）を取り上げる。さらに、そのテストの提供し得る情報の種類及び、(全般的な能力もしくは、学習範囲内での能力といったことも考慮して）決定を下す際の妥当性といった観点からそれぞれのテストに検討を加える。締めくくりとして、ある学習者を想定し、そのケースに沿って、いかに本論が適用し得るかを論ずるが、肝心なのはテストは「敵」なのではなく、情報に富み、妥当な判

断を下す上での有効なガイドラインになり得るということである。

1. Introduction

Many language teachers dislike the whole idea of testing. This distaste may be due to the fact that testing is so often viewed as a "necessary evil" rather than something that we do because we want our students to learn. This paper tries to turn that idea around. The central theme is that testing can and should be used to guide the development of a sound language program and help the students, as well. Thus the paper begins by examining some of the very real-life decision making processes that go on in any language program, such as deciding:

1. who should be admitted into the program;
2. which ability grouping, or level, is appropriate for each student;
3. what should be taught so that resources are most effectively allocated;
4. and, which students have accomplished enough to be advanced to the next level of study.

Such decisions are of particular importance in students' lives and should therefore be based on the best available information. Language tests are an important source of that information.

Hence, another main topic of discussion will be the types of tests that are best suited for helping teachers and administrators make decisions. Four categories of tests will be defined and discussed here. They are (a) proficiency tests, (b) placement tests, (c) diagnostic tests, and (d) achievement tests. These decision making procedures will be examined in terms of the purpose of each type of decision and the type of information that is needed. In addition, examples will be drawn from my personal experience to suggest a number of ways that the different types of tests can be effectively implemented, sequenced, and administered. As always, it will be necessary to consider some practical and political realities as well—realities that exist in any language program, particularly with regard to its tests.

The main point here is that testing is not just a "necessary evil."

On the contrary, tests, if effectively used, can assist language teachers and administrators in making informed and responsible decisions about students, curriculum, and future policy.

2. Two Families of Decision Making

Let us start by thinking about the processes involved in decision making. Administrators may feel that there are infinite decisions that must be made on a daily basis. These vary from minor decisions like what kind of paper clips to buy, to rather major ones like what theoretical directions a program should take. On the other hand, teachers often feel that the really important decisions are those directly related to teaching and to students' learning processes. When I am teaching, decisions about paper clips or even about theoretical perspectives suddenly seem unimportant to me. However, there are administrative decisions that bear upon teaching and vice versa. The focus, here, will be on those types of truly important and practical decisions that affect the lives of all participants in a program: the administrators, the teachers, and the students alike.

Figure 1
Decision Making Using the Four Testing Types

Level of Decision	Program Level		Classroom Level	
Type of Decision	Admissions or Institutional Comparisons	Grouping Ss into Class Levels	Grading, Promotion & Graduation	Diagnosis of Strengths & Weaknesses
Family ofTest	Norm-Referenced		Criterion-Referenced	
Type of Testing	Proficiency Testing	Placement Testing	Achievement Testing	Diagnostic Testing

As shown in Figure 1, there are at least four major types of decisions that affect an entire program. Two of these are the primary responsibility of the administrators. These are essentially

questions about who should be admitted into the program and about which ability grouping, or level, is most appropriate for each student. These will be called *program level decisions* because they have to do with getting the students into the program and into appropriate courses. Such decisions must fit an unknown student into a known program. For example, if students have had previous training at other institutions, it is first necessary to decide if their level of general language ability is too low, too high, or just right for them to fit into the known program. This is the nature of admissions decisions.

Once it has been decided that certain students belong in the program, it is necessary to determine what level within the program is most appropriate for each of them, given their previous training and the training that is available in the program. Should the students be put into the elementary, intermediate, or advanced level? Should they be put into the reading course or listening, writing, and speaking courses? Such decisions are most often entrusted to the administrators who are responsible for the logistics of registration and enrollment in the program. Nevertheless, the teachers will ultimately work most closely with the students, and therefore should be involved in these types of decisions, even if the primary obligation rests with the administrators.

Other decisions (also shown in Figure 1) are the primary responsibility of the teachers. These are usually questions about what should be taught in a course, about students' progress, or about which students have accomplished enough at the end of the course to be advanced to the next level of study. These will be labeled *classroom level decisions* because they have to do with the teaching-learning processes. Such decisions are based on behavior that takes place in the classroom, where the teacher knows best what is going on. Therefore, these decisions are most often made during or at the end of instruction, and are almost entirely internal to the program. In short, they are decisions that help the students progress smoothly through the program.

For example, once the students are in a particular course, it is necessary to make judgments about who needs to work hard on which objectives. Such decisions are usually made early in the course in order to help students use their energies most efficiently. At the midterm, decisions about student progress may require

assigning a grade or otherwise reporting to the administration. At the end of a course, evaluations of each student's success in the course may be necessary, either in terms of a grade or in terms of promotion to the next level of study, or both. They are classroom level decisions and should be primarily the responsibility of the teachers. Nevertheless, the administrators should provide as much assistance to the teachers (in terms of test design and logistics) as possible—even though the primary responsibility for classroom decisions should always remain with the teachers.

In my work, I have repeatedly found that well-designed tests provide one source of valuable information for decision making. I have also found that the two families of decisions, program level and classroom level, are aided by two more-or-less parallel families of tests, called norm-referenced and criterion-referenced tests. In all cases, the tests in our English Language Institute (ELI) have been developed cooperatively with input from administrators, teachers, and students. The goal of our testing program is to help with the types of program level decisions that the director and assistant director must make, as well as the classroom decisions that the teachers must make. Let us consider each of these test types separately.

3. Two Families of Language Tests

Perhaps the most important thing to remember about language tests is the fact that the results can be viewed from a variety of different perspectives and, therefore, can serve a variety of decision making functions within any language program. The problem is matching the correct type of test with the type of decision. To this end, let us consider the two families of tests, *norm-referenced* and *criterion-referenced,* in much more depth.

These two terms are relatively new in our field (see Bachman, 1989; Brown, 1984a, 1988, & 1989 a & b; Cziko, 1983; Davidson, Hudson, & Lynch, 1985; and Hudson & Lynch, 1984), though advocated for years in some educational testing circles (see Berk, 1980; and Popham, 1978 & 1981, for much more on this topic). In fact, the contrast between norm-referenced and criterion-referenced tests dates back to Glaser (1963) and is the focus of many articles and studies in educational and psychological testing jour-

nals (see any recent issue of *Journal of Educational Measurement* or *Applied Psychological Measurement*). The distinction is of increasing importance in those circles, and hopefully will become more significant in language teaching. Thinking about tests in these two distinct ways can help us understand the differences and similarities between the various types of tests that are administered, and the decisions that must be made.

Table 1
Differences Between Norm-referenced and Criterion-referenced Tests

Characteristic	Norm-referenced	Criterion-referenced
1. Type of interpretation	Relative: A student's performance is compared to that of all other students by that student	Absolute: A student's performance is compared only to the amount, or percent of material learned
2. Type of information	General Language abilities or proficiencies	Specific Objectives-based language points
3. Purpose of Testing	To spread students out along a continuum of general abilities or proficiencies	To assess the amount of material known, or learned by each
4. Score Distribution	Normal distribution of scores around a mean score	Ideally, if students know all of the material, they should 100%
5. Knowledge of questions	Students have little or no idea of what content to expect in questions	Students know exactly what content to expect in test questions

(adapted from Brown 1984a, 1989a)

There are a number of important contrasts between norm-referenced tests and criterion-referenced tests in terms of both the theory and practice of language testing. Some of the practical

concerns are summarized in Table 1, which shows that the two sorts of tests differ in the (a) ways scores are interpreted; (b) types of information gathered; (c) purposes for testing; (d) score distributions; and (e) knowledge that students have beforehand about the items.

In brief, norm-referenced tests (NRTs) are designed to measure global language skills or abilities (e.g., overall English language proficiency, reading comprehension, and so forth). A student's score on a norm-referenced test is interpreted in relation to the scores of the other students who took the test. This interpretation is typically done using the statistical concept of normal distribution (the "bell curve") of scores dispersed around the mean, or average, sometimes in terms of percentiles or other standardized scores. The purpose of a norm-referenced test is to spread students out along a continuum of scores so that those with "low" abilities end up at one end of the normal distribution, while those with "high" abilities are found at the other (with most of the students falling somewhere between the extremes near the mean). On a norm-referenced test it is also common that, even though the students may know the general form that the examination questions will take (i.e., the types of items), they typically do not know what specific content will be tested by those questions. For example, they may know that the questions will be multiple-choice, true-false, and so forth, but they usually have no idea what the questions will actually test, except in the broadest terms.

Criterion-referenced tests (CRTs) are produced to measure clearly defined instructional objectives. Often those objectives are unique to a specific program and serve as the basis or outline of the curriculum. Thus, it is usually important for the teachers and students to know exactly what the objectives are so that suitable amounts of time can be spent on them. The interpretation of criterion-referenced test scores is absolute in the sense that each student's score is meaningful by itself, without reference to any other students' scores. In other words, a student's score on a particular objective indicates the percentage of the skill or knowledge in that objective which has been learned or acquired. Furthermore, the distribution of scores on a criterion-referenced test will not necessarily be normal. In theory, if every student has learned 100 percent of each objective, it follows that all the students will

attain the same score. Finally, the purpose of criterion-referenced tests is to measure the degree to which students have developed skills or knowledge in relation to a specific set of objectives. Hence, the students should know what to expect on the test in terms of question types, as well as in terms of the tasks and content for each objective. This information would probably be implied, if not explicitly stated, in the objectives of the course.

The discussion here of norm-referenced and criterion-referenced tests has centered on practical and important differences in the type of measurement involved, the way scores are interpreted and distributed, the purpose for giving each type of test, and the students' knowledge of question content. There are also numerous contrasts between norm-referenced tests and criterion-referenced tests in the ways that they are viewed empirically and treated statistically (see Brown, 1989c; Hudson & Lynch 1984), but for the purposes of this paper, this basic description of the practical differences between the two families of tests will suffice.

4. Matching Tests to Decision Purposes

So far, the discussion has centered on two sorts of decisions that must be made in almost any language program and has shown that there are two families of tests that can be used to help us make those decisions. Before deciding which specific type of test to use for a particular type of decision, it is essential to carefully consider the true purposes of making that decision. Only then is it possible to match the correct type of test to that purpose. In this section, suggestions will be made for ways of making such matches within the context of a language program. It will begin by reviewing the four most commonly used varieties of language tests, while summarizing the main points that should be kept in mind when matching the correct variety of test to the type of decision to be made (see Table 2).

Over the years, the four categories which have been most prominent in language testing are proficiency, placement, achievement, and diagnostic tests (e.g., Alderson, Krahnke, & Stansfield, 1987, pp. iii-iv). As shown in Table 2 these four types of language tests can be categorized very neatly into the two families of tests just discussed: norm-referenced tests tend to be more useful in

Table 2
Matching Tests to Decision Purposes

Test Qualities	Type of Decision			
	Norm-referenced		Criterion-referenced	
	Proficiency	Placement	Achievement	Diagnostic
Detail of Information	Very general	General	Specific	Very specific
Focus	General skills pre-requisite to entry	Abilities, levels and skills in particular program	Objectives of the course or program	Strengths & weaknesses in course objectives
Purpose of Decision	To compare individuals with other groups/ individuals	To find each student's appropriate level	To determine the degree of learning vis-à-vis program objectives	To inform student and teachers of objectives needing more work
Relationship to Program	Comparisons with other institutions	Comparisons within program	Directly related to achievement of program objectives	Directly related to progress on program objectives
When Administered	Before entry and sometimes at exit	Beginning of program	End of courses	Beginning and/or middle of courses
Interpretation of Scores	Percentile position of scores	Percentile position of scores	Amount or percent of objectives learned	Amount or percent of objectives known

making program level decisions (i.e., proficiency and placement); and criterion-referenced tests, are most effective in helping to make classroom level decisions (i.e., diagnostic and achievement). Let us consider each of these categories separately.

4.1 Program Level Decisions: Proficiency and Placement

Proficiency decisions: Who should be admitted into the program?

Sometimes in making decisions, we need to know the students' general levels of language proficiency. The focus of such decisions

is usually on the general knowledge or skills prerequisite to entry into, or exit from, some type of institution, for example, an American university. Making such proficiency decisions may be necessary in setting up entrance and exit standards for a program, in adjusting the level of program objectives to the students' abilities, or in making comparisons across programs. In other words, a variety of curricular and administrative questions may be usefully answered on the basis of overall proficiency information.

One way to gather information for these types of decisions is to compare the overall language performances of individuals to those of other individuals or groups. For this reason, proficiency decisions are often based on proficiency tests specifically designed for such decisions. These tests are constructed to assess general skills commonly required or prerequisite for entry into (or exemption from) a group of similar institutions. One example is the Test of English as a Foreign Language (TOEFL), which is used by many American universities that have English language proficiency prerequisites in common (see ETS, 1987). Such a test will necessarily be very general in nature and cannot be specific to any particular program. Another example of just how general such a test can and must be is found in the ACTFL Proficiency Guidelines (ACTFL, 1986). Though this type of test may contain subtests for each skill, the approach to these skills is still very general and the resulting scores can only be used as overall indicators of proficiency.

Since proficiency decisions are based on knowing the general level of language students in comparison to other students, a test is needed which will provide scores that form a wide distribution so that our interpretations of the differences between students can be fair. Thus proficiency decisions are best based on norm-referenced tests because norm-referenced tests have all of the qualities desirable for such decisions. Proficiency decisions may sometimes seem unfair, but they are often necessary to protect the integrity of the institution, as well as to protect the students from entering a program beyond their abilities or one which they really do not need.

Proficiency decisions include determining how well arriving students will fit into the program, how well the goals of the program meet the actual language needs of the students, and whether students are learning enough, as they go along, to meet overall program goals. When students leave, such tests can be used to

decide whether their level of language proficiency is high enough for their purposes, or alternatively, to discover if the program has had little impact on the language skills that students will be using in the real world. Either way, the information may prove useful to overall curriculum revision.

There are also times when comparisons are made across programs. Since proficiency tests, by definition, are general in nature, they can be used to compare regional branches of a particular language program or to compare different language centers nationwide. It is essential that such decisions be made with extreme care because serious problems can arise. These problems are often due to the very quality that makes such comparisons possible at all, that is, the fact that such tests are not geared to any particular language program. Depending on the test involved, this may result in one program which has students who score very low on the test because the teaching and learning that is going on (though perfectly effective and useful) is simply not assessed by the test. Hence such comparisons must be made with conscientious attention to the validity and appropriateness of the tests for the decisions being made. As with all tests, the information should be assessed carefully and used as part of a larger overall system of information gathering and decision making (see Brown, 1989b).

The general nature of proficiency decisions makes it essential that such tests be designed so that the general abilities or skills of students are reflected on a continuum, or spread of scores. This is desirable so that comparisons among students, or groups of students, can be rationally made. This need for a spread of scores most often leads to the construction of a test that produces a normal distribution (bell curve). All of this is to say that proficiency tests are, and most often should be, norm-referenced in purpose and character.

At the University of Hawaii, proficiency scores, in the form of TOEFL results, are used for two purposes: admission to the university and exemption from ELI training. Students must score a minimum of 500 on the TOEFL to be admitted to studies at our university (some individual departments require more). Once admitted, all international students are subject to clearance from the ELI. In many cases, that means that they must take the ELI Placement Test. However, students who scored 600 or higher on

the TOEFL are automatically exempt from ELI study.

Proficiency decisions should not be dismissed lightly. On the contrary, they must be based on the best obtainable proficiency tests as well as on other types of information. These are decisions that can dramatically affect the students' lives, so it would be grossly unprofessional to do a slipshod job with the proficiency aspect of any language program.

Placement decisions: Which ability grouping, or level, is appropriate for each student?

Still relatively general in purpose, placement decisions are those made for grouping students of similar ability levels together. Often, this is so that teachers can focus in each class on the problems and learning points appropriate to students at a more or less homogeneous level. To help in making such decisions, there is a category of tests designed to aid in deciding what each student's appropriate level will be within a specific program, skill area, or course. The purpose of such tests, then, is to discover which students have more of, or less of, a particular ability, knowledge, or language skill than other students in a particular program.

In thinking about placement tests, it is important to consider the similarities and differences between proficiency and placement testing. As they are defined here, proficiency and placement tests may look very similar and may both be general in nature. However, proficiency tests are designed to assess extremely wide bands of abilities. Hence they are usually applicable across a wide array of institutions. Placement tests, on the other hand, must be more specifically related to a given program, particularly in terms of the range of abilities assessed, so that students can be efficiently separated into approximately homogeneous groups which can effectively be taught similar levels of material.

In other words, a general proficiency test might be useful for determining which language program is most appropriate for a student. Once in the program, the student would take a placement test so that his or her level of study could be determined. Different levels might be appropriate for different skills.

Both proficiency and placement tests should be designed as norm-referenced instruments so that decisions can be made on the students' relative abilities or skill levels. However, as demon-

strated in Brown (1984b), the degree to which a test is effective in spreading students out along a continuum is directly related to the degree to which that test fits the ability levels of the students. Hence it is important to remember that a proficiency test will typically be norm-referenced to a population of students with a very wide band of language abilities and a variety of purposes for learning the language. This is the case with TOEFL, and with the ACTFL guidelines. However, a placement test must be norm-referenced to a narrower band of abilities and purposes—usually those found in students at the beginning of studies in a particular language program.

This distinction becomes particularly important in programs which have tracks and levels. For instance, in the ELI at Hawaii, students who arrive have already been fully admitted. In order to be admitted, they have taken the TOEFL and scored at least 500. In other words, it has been decided in language proficiency terms that these students are eligible to study in the ELI and simultaneously take a few courses at the University. Those students who score 600 or above on the TOEFL are informed that they are completely exempt from ELI training. Thus, with only rare exceptions, most ELI students have scored between 500 and 600 on the TOEFL.

Within the ELI, there are three tracks, each of which is focused on one skill (reading, writing, or listening). Within those skill areas, there are also levels: two levels each for reading and listening, and three levels of writing instruction. As a result, the placement decisions and the tests upon which they are based must be much more focused than the information provided by TOEFL scores. The placement tests must provide information on each of the three skills involved, as well as on the language needed by students in the relatively narrow proficiency range reflected in their TOEFL scores (see Brown, 1987a & b for more details on the entire placement process in the ELI).

There is a dramatic difference between our general proficiency decisions and our placement decisions. While the contrasts between proficiency and placement decisions may not be quite so clear in all programs, the ELI's definitions, and the way we distinguish between proficiency and placement decisions, may prove useful in thinking about testing in other institutions.

In the case of placement, it is particularly important to examine each test in terms of how well it fits the abilities of the students and how well it matches what is actually taught in the classrooms. If there is a mismatch between the tests and the curriculum (see Brown, 1981), the placement of students into levels may be based on something entirely different from what is taught in the levels of the program. So, typically, it is a good idea to make sure that placement decisions are based on placement tests which are either designed or adapted for the specific program involved, or, at least, seriously examined for their appropriateness for the particular program (see Brown, 1989a).

4.2 Classroom Level Decisions: Achievement and Diagnosis

Achievement decisions: Which students have accomplished enough to be promoted?

In a sense, we are all in the business of fostering achievement in the form of language learning. In fact, the purpose of most language programs is to maximize the possibilities for students to achieve a high degree of language learning. As a result, a time always arrives when teachers become interested in making achievement decisions. These are decisions related to the achievement (that is, the amount of learning) of our students. Teachers may also find themselves wanting to make rational decisions that will help improve achievement in the form of deciding or justifying changes in curriculum, staffing, facilities, materials, equipment, and so forth.

In order to make such decisions about student achievement and how to improve it, tests are essential. For instance, achievement tests may help in discovering how much of the language material in the program has been absorbed by each person. Thus achievement tests must be designed with specific reference to a particular program. This link with a specific program usually means that the achievement tests will be directly based on program objectives. Such tests will typically be given at the end of a course or program to determine how effectively students have mastered the objectives.

The tests used to monitor such achievement must be flexible in the sense that they can readily be made to change in response to what is learned from them about the other elements of the curricu-

lum. Within the ELI at the University of Hawaii, criterion-referenced achievement tests have been developed for each of our courses. In each case, the test items were developed directly from the objectives of the course involved. However, in the process of constructing and using these tests, we have discovered many things, not just about the students' learning, but also about the appropriateness of the objectives for those students and about the effectiveness of instruction in each of the objectives.

In other words, well thought out achievement decisions are ones based on tests from which a great deal can also be learned about the curriculum and the program as a whole. These tests should, in turn, be very responsive in the sense that they must be used to affect changes and continually test those changes against the program realities.

Diagnostic decisions: What should be taught so that resources are most effectively allocated?

From time to time, we may also be interested in assessing the strengths and weaknesses of each individual student vis-à-vis the instructional objectives for purposes of correcting the individual's weaknesses "before it is too late." Diagnostic decisions are aimed at fostering achievement by promoting the strengths and eliminating the weaknesses of individual students. Naturally, the curriculum is designed for the entire group of students, but in practice, attention is given to each individual. Thus, with this type of test, it is sound practice to report the performance level on each objective (as a percent) to individual students so that they can decide how and where to invest their time and energy most profitably.

Remember that this last category of decisions is concerned with diagnosing problems that students may be having in the learning process. While this type of decision is definitely related to achievement, diagnostic testing generally requires more detailed information about the specific areas in which students have strengths and weaknesses. The purpose is to help students and their teachers to focus their efforts effectively.

As with achievement tests, diagnostic tests are designed to determine the degree to which the specific instructional objectives have been accomplished. While achievement decisions are usually focused on the degree to which these objectives have been accomplished at the end of the program or course, diagnostic decisions are

normally made as the students are learning the language. As a result, diagnostic tests are typically administered at the beginning or in the middle of a course. In fact, if well constructed to reflect the course objectives, one criterion-referenced test in three equivalent forms could serve as a diagnostic tool at the beginning and midpoints in a course, and as an achievement test at the end.

Within the ELI at the University of Hawaii, the criterion-referenced achievement tests discussed above have been developed in two forms (A and B) for each course so that they can be administered at the beginning and end of each course. At the beginning, randomly selected students take Form A and the remaining students take Form B. At the end of the course, all students take the form they did not take earlier. This type of "counterbalanced" design insures that each student takes a relevant pre-test and post-test without taking the same one twice. It also allows us to diagnose potential areas of student difficulty by analyzing the results of the criterion-referenced tests, particularly the pre-test results.

Again, this system of a diagnostic/achievement testing is important, not only to provide information for the students and teachers, but also to improve the tests themselves, and to implement effective language curriculum planning and development (see Brown 1989b). Such tests are particularly important in examining and revising the program goals and objectives. The information gained can be useful in reexamining the language learning needs of the students, in selecting or creating materials and teaching strategies, as well as in evaluating program effectiveness. Thus it can be argued that the development of systematic diagnostic and achievement tests is crucial to the development of a systematic curriculum. A needs analysis is just a needs analysis and objectives are just so many notions unless they are implemented and tested against the realities of the language learning situation.

5. A Case Study

To review the testing processes described above, let us take a slightly different point of view for a moment. The ultimate user of the services delivered by any language program is the student. Let us consider for a moment a hypothetical student, Toshi, who wants

to study for a Master of Science degree in Oceanography at the University of Hawaii.

In order to be admitted to the program, Toshi must fill out an application, get letters of reference, have transcripts of previous studies sent to the university, and demonstrate adequate English language proficiency by taking the TOEFL. Once his application is complete and TOEFL results are received, the Graduate Division and Oceanography Department are responsible for determining whether or not Toshi is academically admissible—including whether he has a high enough TOEFL score. As mentioned above, the lowest TOEFL score allowed is 500, and Toshi scored 561.

Once Toshi is admitted to his degree program, the TOEFL proficiency test results are sent to the ELI. If he had scored over 600, he would automatically have been exempt from ELI training. However, since he did not, we must consider him for further training in ESL. In his acceptance letter, Toshi is told that he must report to the ELI as soon as he arrives in Honolulu. When Toshi reports to the ELI office, he discovers that he must take the English Language Institute Placement Test (ELIPT) three days later. He signs up for the test and receives a pamphlet describing the various parts of the ELIPT. Toshi reads this information so that he will know what to expect.

He arrives at the placement examination at 7:30 in the morning and takes the following six subjects: Academic Listening Test, Dictation, Reading Comprehension Test, Cloze, Writing Sample, and Academic Writing Test (see Brown, 1987b). The ELIPT is administered and scored primarily by the ELI teachers, so this is Toshi's first contact with that staff. He is finished with the test battery about 11:30 and goes for lunch. At 1:30, he returns for a placement interview with one of the teachers. At that time, Toshi is told that his scores indicate that he is exempt from the listening and reading courses, but that he must take the advanced writing course for graduate students (ELI 83). He agrees, and signs up for the ELI course at the same time that he registers for his oceanography courses.

During the second week of his ELI class, he takes another test. Toshi notices that it is Form B. On the basis of this diagnostic pretest, Toshi's particular strengths and weaknesses can be determined. Perhaps he is told that his grammar is very good, but that

he should concentrate on improving his organization and mechanics. Toshi studies advanced writing for 15 weeks. In the process, he is required to learn word processing and gets a great deal of practice in writing, proofreading, and revising in English.

At the end of the course, he must take another test, this time for achievement. Toshi notices that the test is very similar to the one he took at the beginning of the course, but different too. He also notices that it is labeled Form A. He does very well on this achievement test; that is good because it counts for 20 percent of his grade. Toshi's overall performance in the course, including achievement, is recorded by the teacher and reported to his Oceanography Department advisor. Toshi has completed all ELI requirements.

However, two years later, when he is finishing his master's degree, he decides to apply for a Ph.D. program at Scripts Institute near San Diego, California. It also requires information about his overall English language proficiency in the form of a TOEFL score. Because Toshi has taken the ELI course and spent two years in the United States, his score on the TOEFL is much better. He gets a score of 617, which tells the ESL professionals associated with Scripts that Toshi should be exempt from any further ESL training.

The central message in this hypothetical case study is that the student's viewpoint is important. We must always remember that it is the student who will be most affected by our tests and decisions. It is also important to note that, though there was a great deal of testing involved, it all seemed quite natural. That will be true if the testing is set up as an integral part of the overall curriculum.

6. Conclusion

Clearly, test results can be viewed from a variety of perspectives and serve a variety of different decision making functions within a language program. All language tests are actually based on one of two completely different families of tests: norm-referenced tests designed to help make program level determinations (like proficiency and placement decisions), and criterion-references tests constructed to help make classroom level judgments (like diagnostic and achievement decisions). There are also four primary decision making functions that tests can serve (proficiency, placement, diagnostic, and achievement). We have explored how we might best go about matching our tests to the purposes and decision making

requirements of our language programs, the desires of our fellow teachers, and, above all, the needs of our students.

Decision making processes differ widely from program to program, and only the teachers and administrators involved in a program can determine what types of tests are needed. However, I have tried to outline some of the ways that testing and decision making are handled in the ELI at the University of Hawaii where testing has become more than just a "necessary evil." Tests have been incorporated into the ELI curriculum to perform a variety of necessary decision making functions. Since we want to make these decisions in the most responsible manner possible—precisely because we do care about our students—testing has become an integral and essential part of the language teaching that we provide.

James Dean Brown is director of the English Language Institute, Department of English as a second Language, University of Hawaii at Manoa. He is the author of *Understanding Research in Second Language Learning: A Teacher's Guide to Statistics and Research design* and numerous articles on the role of testing in language instruction.

References

ACTFL. (1986). *ACTFL proficiency guidelines.* Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

Alderson, J.C., Krahnke, K.J., & Stansfield, C.W. (1987). *Reviews of English language proficiency tests.* Washington, DC: TESOL.

Bachman, L.F. (1989). The development and use of criterion-referenced tests of language proficiency in language program evaluation. In K. Johnson (Ed.), *The second language curriculum.* London: Cambridge University Press.

Berk, R.A. (1980). *Criterion-referenced measurement: The state of the art.* Baltimore: Johns Hopkins University Press.

Brown, J.D. (1981). Newly placed versus continuing students: comparing proficiency. In J.C. Fisher, M.A. Clarke & J. Schachter (Eds.), *On TESOL '80* (pp. 111-119. Washington, DC: TESOL.

Brown, J.D. (1984a). Criterion-referenced language tests: What, how and why? *Gulf Area TESOL Bi-annual, I,* 32-34.

Brown, J.D. (1984b). A cloze is a cloze is a cloze? In J. Handscombe, R.A. Orem, & B.P. Taylor (Eds.), *On TESOL '83* (pp. 109-119). Washington, DC: TESOL.

Brown, J.D. with Christensen, T. (1987a). Interview: James D. Brown. *The Language Teacher, 11* (7), 6-10.

Brown, J.D. (1987b). False beginners and false starters: How can we identify them? *The Language Teacher, 11* (14), 9-11.

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. (1989a). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23,* 65-83.

Brown, J.D. (1989b). Language program evaluation: A synthesis of existing possibilities. In K. Johnson (Ed.), *The second language curriculum* (pp. 222-241). Cambridge: Cambridge University Press.

Brown, J.D. (1989c, March). *Short-cut estimates of criterion-referenced reliability.* Paper presented at the Language Testing Colloquium in San Antonio, Texas.

Cziko, G.A. (1983). Psychometric and edumetric approaches to language testing. In J.W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 289-307). Rowley, MA: Newbury House.

Davidson, F., Hudson, T., & Lynch, B. (1985). Language testing: Operationalization in classroom measurement and L2 research. In M. Celce-Murcia (Ed.), *Beyond basics: Issues and research in TESOL* (pp. 137-152). Rowley, MA: Newbury House.

ETS. (1987). *Test of English as a foreign language.* Princeton, NJ: Educational Testing Service.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519-521.

Hudson, T. (1989). Mastery decisions in program evaluation. In K. Johnson (Ed.), *The second language curriculum* (pp. 259-269). Cambridge: Cambridge University Press.

Hudson, T., & Lynch, B. (1984). A criterion-referenced approach to ESL achievement testing. *Language Testing, 1,* 171-201.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

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

Popham, W.J., & Husek, T.R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6,* 1-9.