PART 1

Comprehensive Assessment

CHAPTER 1

Creating a Complete Language-Testing Program

James Dean Brown

INTRODUCTION

The perfect language test for many administrators and teachers would be an aptitude-proficiency-placement-diagnostic-progress-achievement test that does not take too long to administer and is easy to score. For such people, the dream test would be one that does everything and does it painlessly. As I have explained elsewhere (e.g., Brown, 1989, 1990, 1995, 1996b), such a test is simply not possible because the purposes of the different types of tests vary in many systematic ways. In general, language tests can be designed for either norm-referenced or criterion-referenced tests.

Norm-referenced tests are designed to compare the performance of each student with the performances of all the other students in a norm group and do so across a continuum of general language abilities. Thus, norm-referenced tests are not designed to determine how much students have learned in a specific setting, but rather to compare students with other students (usually from many different institutions), even if they all know a great deal or very little. Norm-referenced tests are typically used for three different purposes, which differ in the degree of specificity of what they are measuring:

1. Language aptitude purposes: determining each student's relative standing in terms of his or her overall ability to learn languages (usually for purposes of deciding who would be the best investment for language study); measuring very broadly but not in any particular language
2. Language proficiency purposes: determining each student's relative standing in terms of his or her overall ability in a particular language (usually for purposes of moving students from one institution to another as in university admissions decisions); still measuring broadly but at least within a specific language
3. Language placement purposes: determining each student's relative standing in terms of the appropriate level of study in a particular language program (usually for purposes of creating classrooms of students with relatively homogeneous ability levels); measuring only at the specific levels of ability involved in the placement decisions

Because norm-referenced tests are referenced to a norm group, the three purposes usually cannot be accomplished with the same test. A language aptitude test would typically be designed for a general norm group without reference to a particular language (e.g., the Defense Language Aptitude Test is normed to all new military personnel, who presumably differ considerably in their aptitudes to learn languages). A language proficiency test would also typically be designed for a general norm group, but in this case, only people who have learned a particular language (e.g., the Test of English as a Foreign Language [TOEFL] is normed to a wide range of abilities from virtually no English language proficiency to nativelike proficiency). A language placement test is more specific in that the norm group would be students in the range of abilities found at a particular institution (e.g., the English Language Institute Placement Test at the University of Hawaii at Manoa is normed to the rather restricted range of students who fall between 500 and 600 on the TOEFL). In short, aptitude, proficiency, and placement tests are designed to compare students with each other for purposes of grouping them into those who (a) will or will not be a good investment for language study (aptitude); (b) will or will not be admitted to a particular institution (proficiency); or (c) will study at the elementary, intermediate, or advanced levels or be exempted altogether (placement).

Instead of being designed to compare students with each other for purposes of grouping them, criterion-referenced tests are designed to measure the percentage of material each student has learned or acquired, most often based on the specific objectives of a particular language course. Criterion-referenced tests can be used for three different purposes, which differ in terms of when in the learning process they are administered:

1. Diagnostic purposes: determining each student's strengths and weaknesses vis-à-vis the objectives at the beginning of the course (usually for purposes of giving feedback to students and teachers on where they should focus their energies)
2. Progress purposes: determining how much progress each student has made in learning each of the objectives in the middle of the course (usually for purposes of giving feedback to students on how they are progressing and how they should refocus their energies)
3. Achievement purposes: determining how much each student has learned across all the objectives at the end of the course (usually for grading, promotion, or graduation purposes)

Once such a criterion-referenced test is designed, the three functions can be accomplished with the same test administered at different times (i.e., in the beginning, middle, and end of a course). However, for security reasons, different forms of the test are typically used for these three functions, or better yet, different forms of the test are created and then administered in a counterbalanced manner (i.e., at the diagnostic testing point, some students take Form A, others take Form B, and still others take Form C; at the progress testing point, each of the three groups takes a different form of the test; then at the achievement testing point, each of the three groups takes the third form of the test that they have not previously seen). In short, diagnostic, progress, and achievement tests are all designed to measure how much of a particular set of objectives the students have mastered, but they differ in their purposes and timing.

This case study is of a comprehensive language-testing program in the English Language Institute (ELI), University of Hawaii at Manoa (UHM), that was implemented during my tenure as director. This comprehensive testing program combined and integrated norm-referenced tests (proficiency and placement) and criterion-referenced tests (diagnostic and achievement) into a single language-testing program.[1]

CONTEXT

As shown in Figure 1, the ELI regularly offered seven courses in three skill areas: (a) academic listening (ELI 70 and ELI 80), (b) reading (ELI 72 and ELI 82), and (c) writing (ELI 73, and ELI 83 for graduate students; English as a second language [ESL] 100 for undergraduate students). All of these courses (except ESL 100) were credit equivalent (i.e., students received credit for enrollment and financial aid purposes only). ESL 100 was the credit-bearing equivalent of the regular English composition course.

The curriculum for each course was developed following the model shown in Figure 2, which included six interrelated curriculum components: (a) needs analysis; (b) goals and objectives specification; (c) test development and improvement; (d) materials adoption, adaptation, or development; (e) teacher support; and (f) regularly conducted formative evaluation (for more information on these topics, see Brown, 1995).

After several years, we had completed each of the components shown in Figure 2 for each of the courses, so we set about revisiting the components on a

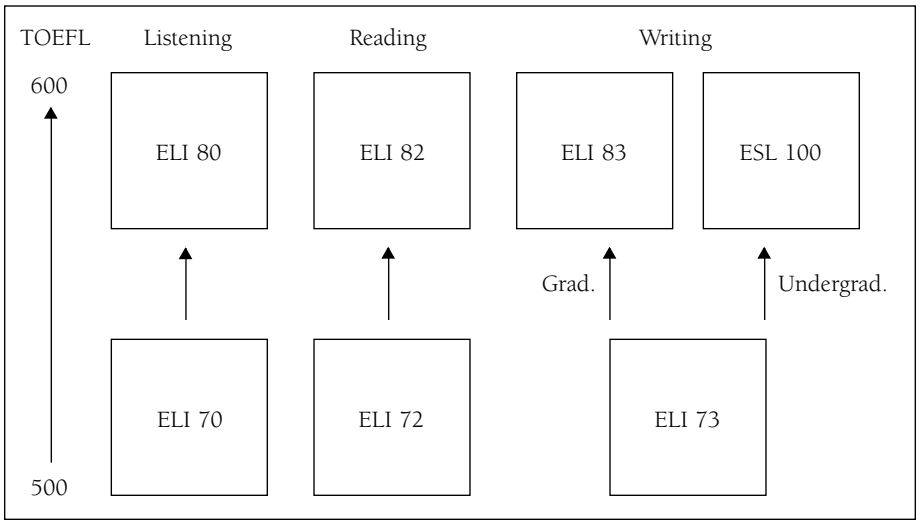

Figure 1. English Language Institute Courses

(From *Testing in Language Programs*, p. 282, by J. D. Brown, 1996, Upper Saddle River, NJ: Prentice Hall Regents. Copyright © 2001 by J. D. Brown. Adapted with permission.)

[1] Note that aptitude testing was not a political option in a program that must accept all students who are admitted to the university, and progress testing was left entirely up to the individual teachers.

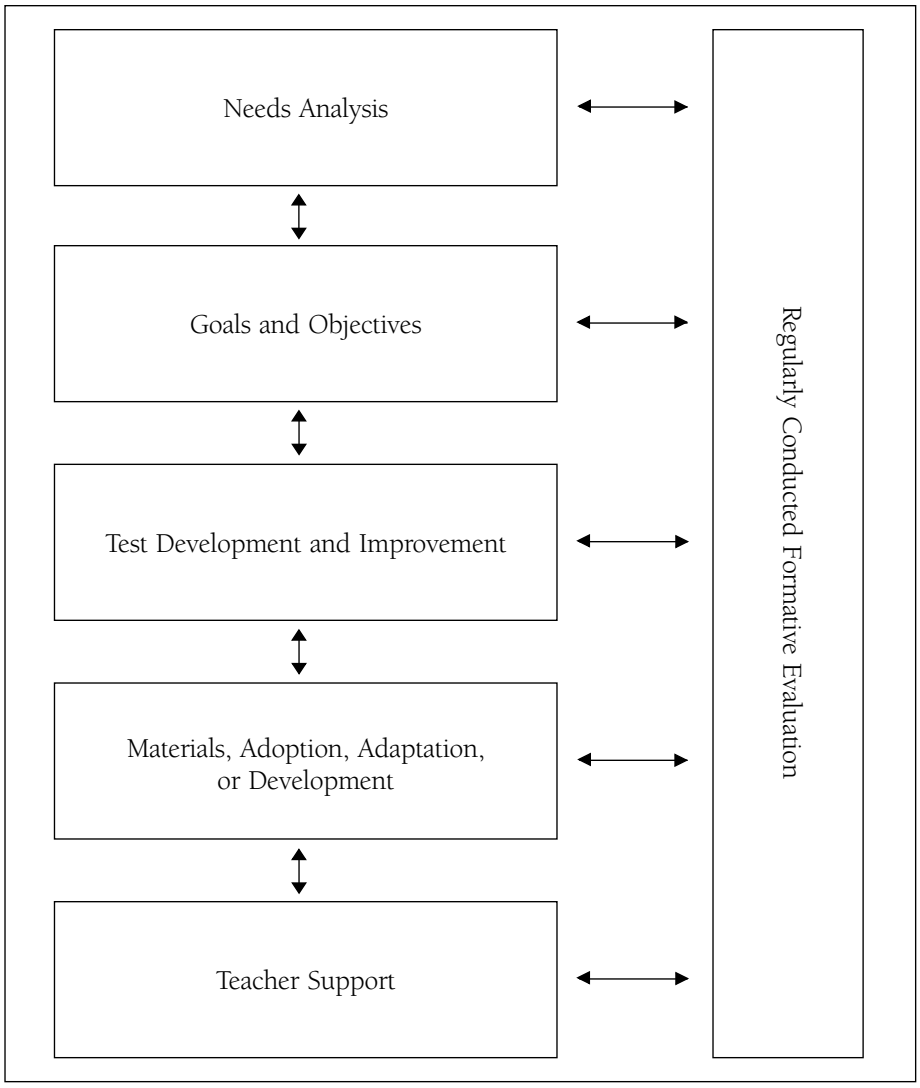

FIGURE 2. Systematic Approach to Curriculum Design

(From *Testing in Language Programs*, p. 27, by J. D. Brown, 1996, Upper Saddle River, NJ: Prentice Hall Regents. Copyright © 2001 by J. D. Brown. Adapted with permission.)

cyclical basis in what we called our ongoing formative evaluation. Every autumn, the curriculum committee (made up of the director, assistant director, and lead teachers for listening, reading, writing, and testing), working with the teachers in each skill area, would revisit the needs analysis, objectives, and criterion-referenced tests (i.e., classroom assessment procedures) to determine the degree to which they still matched the needs of our teachers, students, and administrators. If revision was needed, the needs analysis, objectives, and criterion-referenced tests were revised together as a unit for implementation during the following semester. Each spring, the

same was done for the materials adoption, adaptation, or development component as well as for the teacher support component of the curriculum. In addition, every spring, the ELI director analyzed all of the norm-referenced ELI Placement Test (ELIPT) results from the previous year (usually for about 800 students) and presented an extended report to the curriculum committee with an abbreviated report to the ELI faculty. Both reports covered item analysis, descriptive statistics, reliability analysis, and arguments for the validity of our placement testing procedures, as well as suggestions for revising the test for the following autumn. During the summer, the director also revised and updated the teacher orientation materials for the yearly teacher orientation meeting.

This case study will focus on the test development and improvement portion of the curriculum development processes and show how the proficiency, placement, diagnostic, and achievement testing became an interlocking system of information useful for both teachers and administrators.

DESCRIPTION

An Institutional Point of View

As director of the ELI, it was my duty to make sure all students were working at the appropriate level in each skill area (i.e., listening, reading, or writing) and that they were progressing satisfactorily through the courses. To those ends, a testing program (shown in Figure 3) evolved to include four distinct stages. They include the initial screening stage, the placement stage, the 2nd-week assessment stage, and the end-of-course achievement procedures.

Initial Screening Stage

Before being admitted to UHM, the admissions staff screened all international students by examining their previous academic records, letters of recommendation, TOEFL scores, and financial statements. To be accepted into any degree program at UHM, students were required to have a score of at least 500 on the TOEFL.

International students admitted to UHM were informed of that fact in a letter that also directed them to report to the ELI as soon as they arrived at UHM. They also received a separate letter from the ELI explaining how this requirement worked: We could either give them clearance (if they had 600 or higher on the TOEFL) or assess their need for further ELI courses (if they fell in the range between 500 and 599 on the TOEFL). Thus, students could be entirely exempted from ELI courses or be required to take between one and six 3-unit courses during their first 2 years at UHM. The students were allowed to take these ELI courses in tandem with their other content-area courses at the university. According to UHM policy, ELI courses took precedence over other course work. However, as a matter of ELI policy, we did not usually require students to take more than two ELI courses in any given semester.

At any point in this initial screening stage, students could request an interview with the ELI director for further consideration. This allowed a modicum of flexibility and an opportunity to identify students who could be exempted from ELI training without any further testing. For example, students who were born in foreign countries but did all of their K-12 education in Hawaii or students from India who did all of their education in English-medium schools could be exempted in an

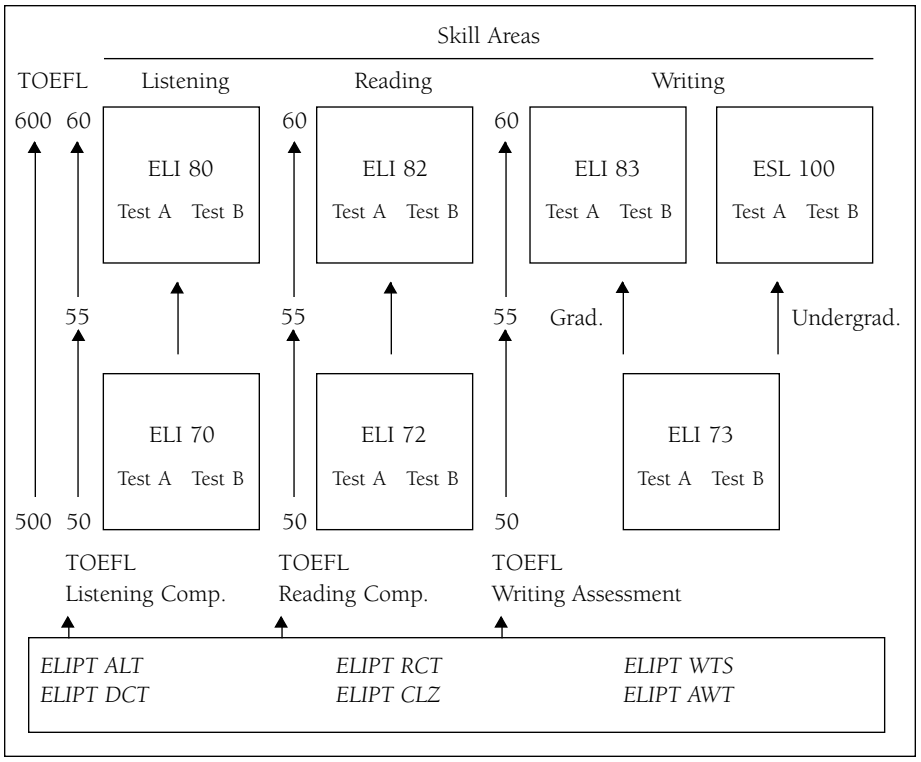

FIGURE 3. English Language Institute Placement Test

(From *Testing in Language Programs*, p. 133, by J. D. Brown, 1996, Upper Saddle River, NJ: Prentice Hall Regents. Copyright © 2001 by J. D. Brown. Adapted with permission.)

interview after examining their high school transcripts. In Hawaii, many different and interesting language background situations arose, and the decisions in these cases could only be made on a one-to-one basis in an interview with the student. In cases where I was in doubt or felt that I did not have enough information, I would have the students take the ELIPT to obtain more information.

Placement Stage

In most cases, students who scored between 500 and 599 on the TOEFL were required to take the ELIPT as soon as they arrived because the ELIPT

- gave us more precise and comprehensive information than the TOEFL scores did
- provided more recent information than the TOEFL scores did (TOEFL scores could be as much as 2 years old)
- pinpointed how the students fit into our particular ESL program in terms of their level of ability in each of the skill areas

The placement stage was particularly important to the ELI program because the ELIPT focused specifically on the skills and levels of ability that were found in the ELI courses.

The ELIPT was a 3-hour battery of six tests, including the:

1. Academic Listening Test (ALT)
2. Dictation (DCT)
3. Reading Comprehension Test (RCT)
4. Cloze Procedure (CLZ)
5. Writing Sample (WTS)
6. Academic Writing Test (AWT)

The ALT and DCT scores served as the basis for placing students into the listening courses (as indicated by the arrows just to the left of ELI 70 and ELI 80 in Figure 3) with the TOEFL listening comprehension subtest scores used as backup information. The RCT and CLZ served as the basis for our reading placement decisions (as indicated by the arrows to the left of ELI 72 and ELI 82 in Figure 3) with the TOEFL reading comprehension subtest scores used as a backup. Finally, the AWT and WTS served as the basis for placing students into writing courses (as indicated by the arrows to the left of ELI 73, ELI 83, and ESL 100 in Figure 3) with the TOEFL writing and analysis subtest scores used as a backup. Note that placement into each skill area was done independently of the other skills. A student, therefore, could end up in ELI 80 for advanced listening, ELI 72 for intermediate reading, and be exempt in writing, though it was much more common for students to place into the same level in all three skills. Note also that each skill-area placement was based on two ELIPT test scores with a TOEFL subtest score as a backup. Thus we had information from three points of view on each student's abilities within each skill area.

Even with all that information, the listening, reading, and writing skill placement decisions were not simple mechanical judgments based on the students' test scores. Each student's placement decisions were made in an individual interview with an ELI teacher or administrator. The interviewers had the student's file and test scores in front of them. They were told to base each skill-area placement decision on the appropriate ELIPT and TOEFL scores as well as to consider any other applicable information (e.g., the length of English language study, speaking ability during the interview, grades in high school). In those cases where an instructor could not make a decision or the student disputed the decision, the instructor sent the student to talk to the ELI director or assistant director for further evaluation and a final decision.

Second-Week Assessment Stage

After the final course withdrawal date in the 2nd week of classes, the teachers were asked to administer two forms of a criterion-referenced test designed to test the objectives of their particular course. These criterion-referenced tests were developed in two forms (cleverly labeled Forms A and B) so they could be administered in a counterbalanced manner. This pattern of counterbalancing is shown inside each of the course boxes in Figure 3. We administered these criterion-referenced tests to

- determine which students, if any, had been misplaced by the ELIPT
- identify students who already knew the material covered in the course objectives so well that the course would do them no good (such students were either promoted to the next level or exempted from further study in that skill area)

- provide the teachers with diagnostic information that would help them to identify the strengths and weaknesses of each student
- emphasize to the students what the objectives of the course would be and how they would be assessed
- permit the curriculum committee to examine the degree to which the students actually needed to learn each of the course objectives

At this point, our placement and 2nd-week assessment procedures had provided the teachers with groups of students who (a) were fairly similar in overall English language proficiency and (b) needed to learn at least some of the objectives of the course.

Achievement Procedures

At the end of the course, to complete the counterbalancing, the teachers administered the opposite form of the criterion-referenced test to each student. We found that putting their names on the forms in advance (instead of trying to decide on the spot who had taken which form and giving them the other one) helped a great deal to avoid confusion.

The teachers' final grades in the courses for each student were based on this test score, as well as on the student's attendance, classroom performance, and other tests and quizzes given by the teacher. In all of the courses except ESL 100 (where students received letter grades), the students were graded "pass," "fail," or "pass and exempt from the next level of study" in that particular skill area. In cases where students failed the final achievement test or otherwise were failing the course, it was my job as ELI director to arrange interviews with the students to advise them that they were failing and to inform them that they would need to repeat the course during the next semester. This took some of the pressure off the teachers.

In all cases, teachers filled out a performance report form for each student. On that form, the teacher filled in the student's grade, a recommendation for what level of ELI course the student should take next, ratings of the student's performance on six different scales (attendance, participation, content mastery, etc.), and a prose description of the student's achievements in the course. These performance reports were copied and sent to the students' academic departments so that their content-area advisors would know how they had performed. In this way, we treated all students equally, whereas those who had achieved more than their fellow students were identified so that adjustments could be made in their subsequent placement.

This system of four assessment stages depended on the information provided by our various types of tests. The initial screening stage depended primarily on the overall norm-referenced proficiency scores provided by the TOEFL. The placement stage relied principally on the norm-referenced placement results of the ELIPT. The 2nd-week assessment stage was based mainly on the criterion-referenced diagnostic test given in the 2nd week of each course, and the achievement stage was based partially on the students' scores of the end-of-term criterion-referenced test.

A Student's Point of View

To review the testing program from a different perspective, I would like to take the point of view of the ultimate consumer of our language teaching services, the

student. Consider for a moment the case of Xiao Lao (a hypothetical student), who wanted to study for a business degree at UHM.

Xiao Lao began the process by filling out an application form, statement of purpose, and financial statement. She also had to get three letters of reference; have transcripts of her previous studies in China officially translated, copied, and sent to UHM; and take the TOEFL test. Once she completed her application and all the other parts had been received (including her TOEFL score of 571), the admissions office decided to admit Xiao Lao.

Once Xiao Lao was admitted, the admissions office immediately (in the bureaucratic sense of that word) sent her TOEFL proficiency test results to the ELI office. If she had scored over 600, she would have automatically been exempted from ELI training. However, her score was 571, so we had to consider whether she needed further ESL instruction. In her letter of acceptance, Xiao Lao had been informed that she would have to report to the ELI office before she could register. She also had received a letter from the ELI telling her very much the same thing but with more details. Unfortunately, she only read the acceptance letter as far as the words "Congratulations, you have been admitted. . . ." Similarly, she glanced at the letter from the ELI but then proceeded to forget all about it.

When Xiao Lao tried to register after standing in a long line, she found that she could not do so because something called the ELI had put a hold on her registration with the message to report to the ELI. Needless to say, she was hot, tired, and irritated when she finally located and reported to the ELI. At that point, she found herself talking to an overworked and busy secretary who asked her to talk to the ELI director. When she did so, she discovered that she would have to take the ELIPT 3 days later and would not be able to register until after that examination—by which time she was sure all of the courses she wanted to take would be closed. Against her wishes, she went back to the ELI secretary and signed up for the ELIPT. At that time, the secretary gave her a pamphlet describing the various parts of the ELIPT. Xiao Lao read this information carefully so that she would know what to expect when she took the ELIPT.

Xiao Lao went to the designated language laboratory to take the placement examination at 7:30 a.m. and took all six tests. ELI teachers were proctoring the ELIPT and this was Xiao Lao's first contact with these instructors. She finished the test battery about 11:00 a.m. and went to lunch. She had to return to the ELI office area at 1:30 p.m. for a placement interview conducted by one of the instructors. In that interview, the instructor told Xiao Lao that she was exempt from any further reading courses (as indicated by her ELIPT scores) but that she would have to take the advanced listening course (ELI 80) and the advanced writing course for graduate students (ELI 83). She reluctantly agreed to sign up for the courses. She was reluctant because she was in a hurry to finish her degree due to financial constraints and these ELI course requirements would slow her down. However, because she had no choice, she agreed to sign up for the two ELI courses at the same time as she registered for some of her other degree courses.

During the 2nd week of each of her ELI classes, she found herself taking another test in each class. Xiao Lao noticed that she was taking Form B in the listening class and wondered what that meant. In the directions for the tests, the teachers told her that any students who did very well would be exempted from the course, so Xiao Lao did her best on the test but did not do well enough on either test to be exempted.

However, she did notice that one student in her listening class and two in her writing class were exempted. Based on this 2nd-week diagnostic test, both teachers told Xiao Lao about her specific strengths and weaknesses in the skill area involved. For instance, in her writing course, she was told that her grammar was excellent but that she should concentrate on improving her organization and mechanics. Xiao Lao took the advanced listening and writing courses for 15 weeks. In the listening course, she learned about listening skills, did a great deal of listening to English, and gave a speech. In her writing course, she learned word-processing skills and practiced writing, proofreading, and revising her written English.

In the 16th week, at the end of her courses, Xiao Lao had to take another test to measure her achievement in each course. Xiao Lao noticed that the test was similar to the test she had taken at the beginning of the course but was not exactly the same. She noticed that the one in her listening course was labeled Form A. She did very well on these achievement tests, which was good because she knew that she had to pass these achievement tests to pass the courses. In addition, these achievement test scores would account for 20% of her grade in each course. Xiao Lao's overall performance in the course was recorded by the teacher and reported to her Business School advisor. Thus, Xiao Lao had completed all of her ELI requirements.

DISTINGUISHING FEATURES

One central message in this case study is that, from the ELI's point of view, testing was an important and integral part of the curriculum. In addition, all program participants (students, teachers, and administrators,) benefited from the testing.

Because the placement test was essentially norm-referenced and used for administrative decision making, I took primary responsibility (in my role as ELI director) for developing and revising it.[2] As mentioned above, I reported annually on the ELIPT to the ELI staff as a means of quality assurance. In this report, I examined the item characteristics; descriptive statistics; and reliability, validity, and practicality of the various tests and suggested ways the test could be revised and improved. At the same time, I sought feedback from the teachers on any aspects of the tests they thought were problematic.

The development of the criterion-referenced diagnostic and achievement tests was even more cooperative. The lead teacher for testing worked with the teachers in each of the skill areas to write these tests. As a result, each and every test item came directly from the teachers. Thus, they had a large stake in the testing process.

Another central message in this case study was that the student's point of view was considered important. The ELI administrators and teachers never forgot that students were affected by the tests and the decisions based on those tests. Even though there was a great deal of testing, to Xiao Lao the testing appeared to be a natural part of the way things were done in the ELI. I think this was true only because the testing was set up as an important and integral part of the overall curriculum.

[2] Teachers were in no sense excluded from the norm-referenced testing. Indeed, they often needed to explain the ELI admissions and placement procedures to students and were a crucial part of the proctoring, scoring, and interviewing/placing processes during the actual ELIPT admissions.

The following characteristics of this testing program distinguished it from the testing that goes on in most language programs:

- Both norm-referenced (TOEFL and ELIPT) and criterion-referenced tests (diagnostic and achievement) were included in the testing program.
- The different types of tests were integrated into one continuous assessment and decision-making process that included admissions, placement, diagnostic, and achievement decisions.
- Teachers were involved in the placement testing (as proctors, scorers, interviewers, and critics).
- Each year, one teacher was given release time as lead teacher for testing to coordinate the development, administration, analysis, and revision of the criterion-referenced tests for diagnosis and achievement.
- Administrators were supportive of test development for classroom purposes as evidenced by the marshalling of resources (e.g., release time for a lead teacher for testing).
- Teachers were integrally involved in developing the items for the diagnostic and achievement tests in each course (as item writers, proctors, scorers, interviewers, and critics) and understood that to be part of their jobs.
- All placement tests were regularly and appropriately analyzed (for norm-referenced item characteristics, reliability, and validity) and revised (e.g., see Brown, 1989).
- All diagnostic and achievement tests were regularly and appropriately analyzed (for criterion-referenced item characteristics, dependability, and validity) and revised (e.g., see Brown, 1993).
- Students were not forgotten in the testing process. The students' points of view on the testing were considered important.
- Testing was important, but every effort was made to keep it from becoming intrusive.

PRACTICAL IDEAS

To avoid creating a blissful picture of curriculum perfection, in this final section, I will discuss some of the problems we had in developing our comprehensive language-testing program and how we overcame them.

Get the Work Done

Testing is often an afterthought in language teaching. Such an attitude is too prevalent in language programs and, from my point of view, irresponsible. Tests are too important in terms of the significant decisions we make with them to be treated so cavalierly. Unfortunately, an integrated testing program (complete with a six-test placement battery and seven different classroom-oriented criterion-referenced tests in two forms each) is not easy to develop, implement, and maintain. Indeed, such a project across many courses requires an extraordinary amount of work. Nonetheless,

a group of administrators, lead teachers, and teachers did create just such a testing program.

Can we count on teachers to do all that work on their own? Indeed, can we count on teachers to independently do any sort of coherent curriculum development across courses? No, we probably cannot, unless their work is coordinated and supported at the institutional level. Fortunately, the cost of such support need not be very high. The sole costs of our coherent testing program were constant attention from one ELI director, the appointment of a lead teacher for testing, and the cooperation of all teachers.

Appoint a Lead Teacher for Testing

The lead teacher for testing was given 50 percent release time (amounting to a one course reduction) to help the administrators and teachers develop, administer, score, analyze, and revise all the ELI tests. At the beginning of each semester, during the norm-referenced placement testing, the lead teacher for testing took responsibility for administering the tests, while the assistant director took charge of logistical matters and the director did the scoring and analysis of the results as each test was completed.

During the remainder of the semester, when our attention turned to the diagnostic and achievement tests, the lead teacher was in charge of the testing logistics, which included:

- getting teachers to write the items
- putting them together into a test
- proofreading the items
- getting the ELI director to proofread the items
- revising and compiling the pilot items into two forms
- photocopying the tests
- distributing the tests to teachers for in-class administration
- collecting the tests after they were administered
- scoring the tests
- getting the results back to the teachers (within 24 hours if possible)
- analyzing the items for revision purposes
- analyzing the tests for dependability and validity

As for the importance of fancy statistics, I would like to emphasize that the lead teacher for testing needed only the knowledge about language-testing programs provided in a book such as Brown (1996b). With basic understanding of item analysis techniques, any language teaching professional should be able to get a good start on developing such tests.

Get All Teachers on Board

At first, getting teachers to cooperate in this testing effort was difficult. Many teachers felt they worked too hard and did not want to do anything extra. A number of

strategies were used to overcome this attitude, but the primary and most successful strategy involved clearly outlining the teachers' duties.

The curriculum committee began this process by first outlining their duties and responsibilities. These descriptions were discussed with the teachers and revised based on their comments. The committee was then in a good position to develop a description of the teachers' duties and responsibilities. Using the graduate division's description of the duties of graduate assistants (all ELI teachers fell in this category at that time), the curriculum committee proposed a draft set of duties, which included the graduate division's requirement that they work 20 hours a week. We divided the 20 hours as follows:

- 6 hours of classroom teaching
- 6 hours for preparation time
- 2 office hours
- 1 hour for meetings
- 5 hours for curriculum development

Thus, curriculum development (and therefore testing) was included as an integral part of the teachers' job description. In the discussions that followed, some teachers were unhappy about what they (correctly) perceived as a trick to get them to do more work. But that unrest soon settled down, when everybody realized that they were seldom asked to do any more than an hour or two of curriculum work each week. In addition, the teachers realized that they received a great deal from these curriculum development efforts.

Anticipate Surprise Problems

As in so many human endeavors, we often did not learn what we set out to learn in the process of developing our tests. For example, we chose to develop two forms of the ELI 72 reading test as our first criterion-referenced diagnostic and achievement tests. We carefully developed them to match the course objectives and administered them with some excitement in the 2nd week of classes. Unfortunately, the scores of the students for each and every objective indicated that the students already knew virtually all of the material we were planning to teach them—at the beginning of the course. We found ourselves facing the uncomfortable fact that our objectives, which had been used for many years, were targeted at a level far too low for the abilities of our students.

As a result, we had no choice but to discard our tests and revise our objectives considerably. In this particular case, we were able to use similar objectives by applying the existing objectives to considerably more difficult reading texts. Initially, we had been devastated to realize that our first criterion-referenced test development efforts were in vain, but, in retrospect, this early attempt at developing diagnostic and achievement tests (even though it failed) helped us to change our views of the students' language abilities and learning needs. This failure benefited not only the single reading course but all the other courses as well.

Thereafter, we never had such a wholesale failure of our curriculum, though in virtually every course, we did get detailed information from our criterion-referenced tests about objectives that (a) were not being tested well, (b) were not being taught

at all, (c) were not being taught well, or (d) were not teachable at the level of proficiency involved. All of which, on an objective-by-objective basis was useful information for curriculum-revision purposes.

Rely on Multiple Measures of Assessment

Although we did everything possible to ensure that our testing program appropriately admitted, placed, diagnosed, and promoted most of the students served by the ELI, we recognized that such decisions are made by human beings and that, even when based on something as seemingly scientific as test scores, the human judgments involved are hardly perfect. One way a comprehensive testing program like that described here helps to safeguard against errors in judgment is by relying on multiple sources of information. Admissions decisions were never based solely on the proficiency test scores but rather on multiple sources of information including grades, letters of recommendation, statements of purpose, financial statements, and TOEFL scores. Similarly, placement decisions for each skill area were based on two ELIPT test scores with a TOEFL subtest score as backup and other information gathered during a personal interview with each student. The classroom diagnostic and achievement criterion-referenced decisions were each based on a single test score. However, at the achievement decision point, the diagnostic and achievement test results were examined together and teachers were encouraged to gather information of their own including homework, quizzes, midterm exams, attendance, and additional final examination testing.

CONCLUSION

In sum, we viewed our admissions, placement, diagnostics, and achievement decisions as important: An incorrect decision could cost a student a great deal in extra time and tuition for studying in the ELI. Thus, we took those decisions seriously and based them on the best available information taken from a variety of sources including test scores.

Such a testing program certainly entails considerable effort on the part of the administrators, lead teachers, and teachers, but the benefits derived from effective and fair testing procedures accrue to the entire curriculum and to everyone involved in the language learning process.

ACKNOWLEDGMENTS

I would like to personally thank for their dedication and hard work the three testing lead teachers who served while I was director of the ELI: Charles Lockhardt, who is a well-known author and professor in Hong Kong; Gary Cook, who recently completed a doctorate in educational testing at the University of Michigan; and Yuichi Watanabe, who just completed his dissertation defense at UHM and is working as a professor at the Kanda Institute of Foreign Studies in Japan.

CONTRIBUTOR

James Dean ("JD") Brown is a professor in the Department of Second Language Studies at the University of Hawaii at Manoa. He has published 13 books and numerous articles in the areas of language testing, curriculum design, and research methods.