A COMPREHENSIVE CRITERION-REFERENCED LANGUAGE TESTING PROJECT

J.D. BROWN
University of Hawai'i

The English Language Institute (ELI) at the University of Hawai'i regularly offers seven courses in academic listening, reading, and writing. The curriculum for each course has been extensively revised including thorough needs analysis, development of objectives, criterion-referenced tests, and materials, as well as improvements in teaching practices and regularly conducted formative evaluation procedures. This paper reports on the criterion-referenced test development portion of the curriculum.

Each of the seven ELI courses has two forms of a criterion-referenced test designed expressly to measure the objectives of that course. The two forms are administered at the beginning and end of instruction in a counterbalanced design. Hence this testing project is large in scale including 14 different tests administered before and after instruction for about 500–600 students per year. While the objectives and resulting tests differ in organization and form across the seven courses, the processes involved in putting the tests in place are quite similar. The initial item development, piloting and revision processes are described in general terms. Details are provided about the results of the administrations of these CRTs during fall 1989. Descriptive and item statistics are presented (including the difference index, item ϕ, B-index, and item agreement index) for each test. Dependability estimates [phi and phi(lambda)] are given, and evidence for the content and construct validity of the tests is also provided.

The discussion centers on the problems encountered in developing such a comprehensive testing program, then turns to the benefits which CRTs can provide for overall curriculum development.

INTRODUCTION

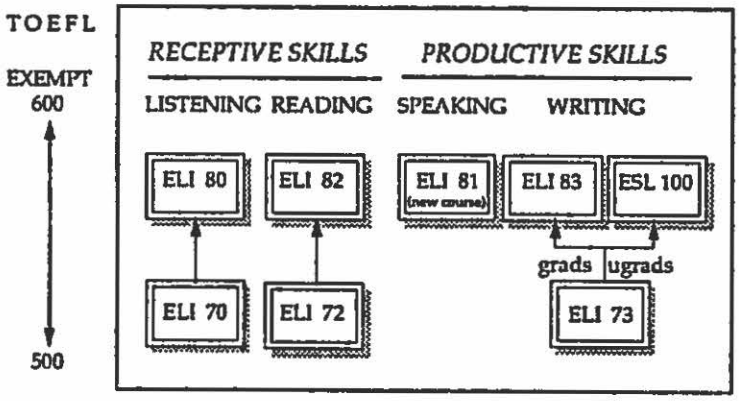
IMMEDIATELY UPON ARRIVAL, all foreign students who have been admitted to the University of Hawai'i at Manoa (UHM) are required to report to the English Language Institute (ELI) for clearance. The purpose of this clearance process is to determine the amount of ESL training that students must undergo,

University of Hawai'i Working Papers in ESL, Vol. 10, No. 1, Spring 1991, pp. 95–125.

if any. Thus, students may be entirely exempted from ESL courses or be required to take between one and six three-unit courses during the first year or two of their stay at UHM. These classes may be taken concurrently with other courses at the university, but, according to University policy, ELI courses take precedence over all other course work.

The ELI regularly offers seven different classes in academic listening, reading, and writing. In addition, a new course in Speaking for Foreign Teaching Assistants is also being implemented in the Fall semester 1990. These courses are shown in Figure 1. Notice that a TOEFL range of between 500 and 600 is indicated down the left side of the figure and that the courses are clearly organized into four skill areas and two levels (which roughly correspond to TOEFL ranges of 500–549 for the courses numbered in the 70s and 550–599 for those numbered in the 80s or higher).

Figure 1
ELI Courses

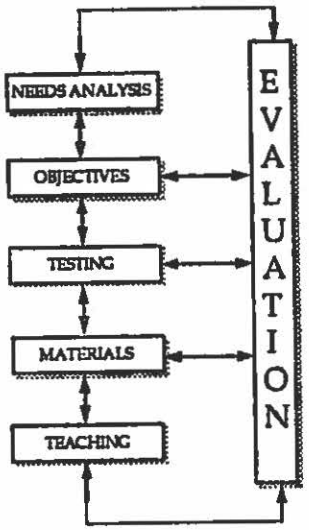


Curriculum Context

Recently, the curriculum for the courses shown in Figure 1 has been extensively revised. The revisions have included 1) thorough needs analysis, 2) development of objectives, 3) design and implementation of tests, 4) materials development, 5) improvements in teaching practices, and 6) regularly conducted formative evaluation procedures. Figure 2 illustrates how these elements are related in our curriculum. Notice the central position of testing in the model as well as the fact that program evaluation is depicted as formative

and constantly interacting with all of the other elements of the curriculum development process. (For more complete descriptions of this model see Brown 1989b & in preparation.)

Figure 2
Systematic Approach to Curriculum Developmemt in the ELI
(adapted from Brown 1989b)



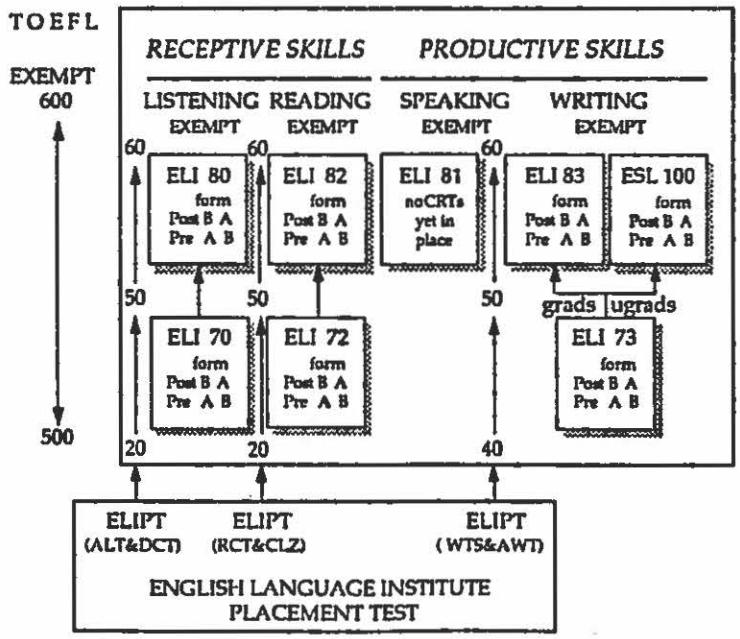
This paper reports on the testing facet of the new curriculum and will almost immediately narrow the focus to the development and implementation of criterion-referenced tests for individual courses. Notice in the model in Figure 2 that testing comes into play primarily after clear program and course objectives have been established. As described in the next section, tests serve a number of purposes in our program. However, it is the development of course-level criterion-referenced tests that will be of primary interest here. As indicated by the arrows in Figure 2, the development of such tests is viewed as interacting with objectives, materials development and teaching, so that each can be used to improve the others (sometimes through the program evaluation processes). Such interactions among objectives, criterion-referenced tests,

materials, and teaching are taken to be essential to the success of our testing program and, indeed, to the growth of the entire curriculum.

Testing Program

As Director of the ELI, it is my duty to insure that decision making mechanisms are in place to insure that students are working at the correct level and progressing satisfactorily through our program. To those ends, we have designed four sets of procedures: 1) initial screening procedures, 2) placement procedures, 3) second week assessment procedures and 4) achievement procedures. Brief discussion of each of these sets should help to clarify our testing program as it is shown in Figure 3. This section is meant to clarify the decision-making context in which our criterion-referenced tests operate.

Figure 3
ELI Testing Program



Initial screening procedures. Before students are admitted to UHM, they are screened by the Office of Admissions. The students' previous academic records, letters of recommendation, TOEFL scores, financial situations, etc. are carefully reviewed. From our perspective, one of the most important pieces of

information is the students' TOEFL results because only students with total scores of 500 or more are accepted for admissions to UHM. This information, including each student's TOEFL subtest and total scores, is immediately sent to the ELI. Students with scores above 600 are automatically exempted from any further ELI requirement and are notified of that fact before arriving. Those students who scored in the range between 500 and 599 are informed that they must clear the ELI immediately upon arrival in Hawai'i. The initial screening procedures clearly serve the beneficial purpose of narrowing the range (see arrows to the left of Figure 3) of overall English proficiency with which we must concern ourselves in the ELI.

At any stage of this process, any student may request an interview with the Director in order to have the particular case reconsidered. This permits some flexibility and a chance to identify students who may easily be exempted from ELI training without any further testing (e.g., students who were born in foreign countries but did K–12 in Honolulu, or students from India who did all of their education in English medium schools). In Hawai'i, we encounter many different and interesting situations, particularly with immigrants, which can only be decided on a one-to-one basis.

Placement procedures. However, a majority of the students who score between 500 and 599 on the TOEFL are required to take the ELI Placement Test (ELIPT) as soon as they arrive on campus. This test serves three purposes: 1) it gives us more detailed information than that provided by TOEFL scores; 2) it yields information that is more recent than the TOEFL scores (which can be up to two years old); and 3) it provides information about how the students will fit into our particular language program (in terms of their level in each skill area). Placement procedures are particularly important in our program because we have different tracks and levels (as shown in Figure 1). Recall that we have four tracks, each of which is focused on one skill (reading, writing, listening or speaking), and that, within the tracks, there are up to two levels. As a result, the placement tests must be as focused as possible on the skills and levels of ability that are found in the ELI.

The ELIPT is a three hour test battery made up of six subtests: the Academic Listening Test (ALT), Dictation (DCT), Reading Comprehension Test

(RCT), Cloze Procedure (CLZ), Writing Sample (WTS), and Academic Writing Test (AWT). The ALT and DCT are used to place students into our listening skill courses (see arrows just to the left of ELI 70 and 80 in Figure 3). The RCT and CLZ are used for the reading skill courses (see arrows to the left of ELI 72 and 82). The AWT and WTS are employed for placing students in writing track (arrows to the left of ELI 83, ESL 100 and ELI 73). Notice that we have two test scores for placement into each of the three primary skill areas that we teach. This arrangement provides us with two different views of each student's abilities within a particular skill area.[1]

However, our placement decisions are based on much more than the students' ELIPT scores. The each student's actual placement occurs during an individual interview conducted by an ELI instructor. The interviewers are provided with the student's file and test scores and are told to base their placement decisions for each skill area not only on the two ELIPT subtest scores for that skill, but also on any other pertinent information in the student's records (e.g., the length of English study, amount of time since that study, TOEFL subtest scores, spoken production in the interview, academic records, and any other information available at the time). In cases where the instructor cannot decide, or when a student disputes the decision, the ELI Director (or Assistant Director) interviews the student and makes a final decision.

Second week assessment procedures. During the second week of classes, teachers administer a criterion-referenced test designed to test the course objectives. These tests are currently used in two forms (forms A and B). They are administered in a counterbalanced design such that half of the students in each course take Form A at the beginning and half take Form B; at the end of the course, all students take the opposite form. This is illustrated within each of the course boxes in Figure 3.

There are three purposes for this first week test administration: 1) it helps teachers to determine which students have been misplaced; 2) it provides teachers with an opportunity to diagnose any weak students who may need special help; and, 3) it allows the curriculum committee to take a hard look at

[1] The Speaking course is only available for international teaching assistants. The testing is therefore handled separately from the mainstream program.

the degree to which students across sections of a course actually need to learn (i.e., score low on) each objective. Thus this pretest administration is an integral part of our placement, diagnosis and curriculum development processes.

Achievement procedures. The same criterion-referenced test (the opposite form) is administered to each student at the end of the course. Based on this test and the student's classroom performance, the teachers must decide whether to pass, fail, or suggest exemption from any further study in that particular skill area. Again, in cases where it is necessary, interviews with the ELI Director are set up, and the students are advised on what we feel is most appropriate for them. In all cases, a student performance report is filled out by the teacher for each student. On that form, the teacher is asked to grade the students, specify what level of ELI course the student should take next, rate the student on six different scales (e.g., attendance, participation, content mastery, etc.), and provide a prose description of their content mastery and conduct in class. Copies of these reports are then sent to the students' academic departments so that their advisors will know how they performed. In this way, all students can be treated fairly, while those who have learned more than their peers can be identified and adjustments in their subsequent placement can be made.

This whole system of procedures is enhanced by (but not limited to) information provided by tests. The initial screening procedures rely primarily on the norm-referenced overall proficiency scores provided by the TOEFL. The placement procedures depend, in large part, on the norm-referenced placement results provided by the ELIPT. The second week assessment procedures are based partially on the criterion-referenced diagnostic test given at the beginning of each course, and the achievement procedures are largely based on criterion-referenced posttest scores.

Why Criterion-referenced Tests?

One definition for a *criterion-referenced test* (CRT) is provided by Richards et al (1985):

> a test which measures a student's performance according to a particular standard or criterion which has been agreed upon. The student must reach this level of performance to pass the test, and a student's score is therefore interpreted with reference to the criterion score, rather than to the scores of other students.

This is markedly different from the definition for a *norm-referenced* test (NRT) which is taken from the same source:

> a test which is designed to measure how the performance of a particular student or group of students compares with the performance of another student or group of students whose scores are given as the norm. A student's score is therefore interpreted with reference to the scores of other students or groups of students, rather than to an agreed criterion score.

While these may not be the most comprehensive definitions available, they do point to the most important difference between these two types of tests: the performance of each student on a CRT is compared to a particular standard called a criterion level (e.g., if the passing score on a test were set at 60 percent, a student who answered 66 percent of the questions correctly would pass), whereas on an NRT, each student's performance is compared to the performances of other students in the group that has been designated as the norm (e.g., if a pupil scored in the 98th percentile, that performance was better than 98 out of 100 people who took the test, without reference to the actual number, or percentage, of items correctly answered).

The key to understanding the difference between CRTs and NRTs lies in the distinction between the terms percentage and percentile. In administering a CRT, the primary focus is on the amount of material that the students know. As a result, it makes sense to report the results in the form of a percentage, i.e., the percent of the questions that the students can answer correctly in relation to the material taught in the course and in relation to a previously established criterion level for passing.

On an NRT, the concerns are entirely different. Instead, the focus is on how each student's score is related to the scores of the other students who took the test. Thus the central issue is the student's position in the distribution of

scores. This can be expressed in terms of a percentile score because such scores reveal the proportion of students who scored above and below a given student.

In short, CRTs are generally designed to assess the amount of material known by each individual student while NRTs examine the relationship of each student's performance to the scores of all of the other students. These definitions cover the primary difference between the two types of tests, that is, that the scores are interpreted differently. However, as a result of this primary distinction, there are other differences that arise in practice. The two types of tests also differ in five other ways: 1) the kinds of things that they are used to measure, 2) the testing purposes involved, 3) the distributions of scores that will result, 4) the testing formats, and 5) the degree to which students know what content to expect (for more information on these differences, see Brown 1989a & 1990).

The separation of tests into the norm-referenced and criterion-referenced interpretations is becoming increasingly important in the language testing literature (e.g., Cartier 1968, Cziko 1982 & 1983; Hudson and Lynch 1984; Delamere 1985; Henning 1987; Bachman 1989 & 1990; Brown 1984a, 1989a, 1989b, 1989c & 1990). It has surely been an important issue for years (beginning with Glaser 1963) in educational testing circles. For example, almost any recent volume of the *Journal of Educational Measurement* or *Applied Psychological Measurement* will contain at least one article on criterion-referenced testing issues. More importantly to us, the NRT/CRT distinction is becoming increasingly useful at UHM for developing and analyzing the various kinds of tests that we need for admissions, placement, diagnosis and achievement decisions.

This paper is the first to describe the criterion-referenced side of our testing program. To that end, the following research questions have been framed to help give shape to the description of our results:

1) What are the descriptive characteristics of criterion-referenced tests when used in a variety of courses? How do they differ across skills, levels, and courses?

2) What item statistics are most useful for revising criterion-referenced tests in such a context? How does the usefulness of NRT, CRT and IRT (Item Response Theory) approaches compare?

3) To what degree are these criterion-referenced tests consistent in what they test? How do NRT reliability and CRT dependability approaches compare in usefulness? How do they differ?

4) To what degree are these criterion-referenced tests valid? What strategies can best be used to investigate the validity of CRTs in a practical situation?

METHOD

Subjects

The students involved in this study were the 294 who were enrolled in the Fall semester 1989 in the ELI at UHM. This group included 29 percent graduate students, 58 percent undergraduates, 9 percent unclassified, and 4 percent with other classifications. They came mostly from countries in Asia with 26 percent from the People's Republic of China, 14 percent from Hong Kong, 11 percent from Korea, 9 percent from Japan, 8 percent from the Philippines, 6 percent from Vietnam, 6 percent from Taiwan, 4 percent from Indonesia, 3 percent from Thailand, 2 percent from Macao, 2 percent from Malaysia, and the remaining 9 percent from 18 other countries. Of these students, 85 percent were new to the ELI, while 15 percent had taken previous course work with us.

Of the total number of ELI course enrollments, 9 percent were in ELI 70, 20 percent in ELI 80, 12 percent in ELI 72, 25 percent in ELI 82, 13 percent in ELI 73, 7 percent in ELI 83, and 14 percent in ESL 100.

Materials and Procedures

While the objectives and resulting tests differed in organization and form across the seven courses, the processes involved in putting the tests in place

have been quite similar. The initial item development was done collectively by the teachers in each skill area as part of their overall curriculum development commitment. Once thorough-going needs analyses had been performed for each course and tentative sets of objectives were established, the work of actually writing items to measure those objectives began. The piloting and revision processes have been ongoing for nearly three years with various tests at different stages of development. The tests were administered in the students' classrooms during the second week of class and again during final examination week.

It is important to recognize that the CRTs created by our teachers differed considerably in organization and form across the skill areas and levels. Since all decisions about test content and methods were made by consensus among the teachers, the test methods ranged considerably from multiple-choice format to open-ended writing tasks depending on the skill being tested. For instance, a typical multiple-choice item might be the following "inference" item, which was used in the directions on the lower-level reading course test:

> Out of the darkness of the cold, wintry night came the clatter of a toppled garbage can lid. Startled, Peter dropped his book and ran to the back door.
>
> Ex.1 What was Peter doing before he heard the noise?
> A. singing C. washing
> B. reading D. sleeping

Naturally, the reading passages in the test itself were considerably longer and more academic in nature.

The writing tasks assigned to the students also tended to be academic in focus. One such task, meant to simulate an in-class essay, required the students to read a five-page selection on genetic engineering and then answer an essay question on the ethics of genetic engineering in 60 minutes (with no notes). They were rated using a scoring grid developed specifically to reflect the ELI objectives (similar to one shown in Brown and Bailey 1984). A similar strategy was used by our listening teachers in ELI 80 to score in-class presentations.

Unfortunately, because of the time constraints for scoring (especially during the final examinations), we have tended to favor the machine scorable test formats. However, as we continue to gain experience and confidence in criterion-referenced testing, we are becoming increasingly willing to experiment with more imaginative test types. For instance, we are currently focusing on development of task-based subtests. These tasks will be assigned during the last week of classes and scored in conjunction with the students' final examinations. An example taken from the upper level reading course will involve a set of tasks wherein the students are required to go to the library, retrieve specific information, and report it back to the teacher on open-ended forms. Their answers will then be scored for accuracy and completeness, and the scores will be included in their overall final examination scores.

Analyses

Because we were breaking new ground, a variety of testing statistics were used in our analyses. Techniques were borrowed from classical (NRT) theory approaches, from the CRT literature including generalizability theory, and from item response theory (IRT). The analyses were performed entirely on an IBM AT desktop computer using the QuattroPro (Borland 1989) spreadsheet program and a test analysis program called TESTAT (SYSTAT 1987). Thus the technology required is well within the resources of many language programs.

Descriptive statistics include the mean, standard deviation, range, number of items and number of subjects. *Item statistics* include traditional NRT statistics (item facility and discrimination indexes), CRT estimates (difference index, item ϕ, B-index, and item agreement index), as well as IRT (item difficulty and discrimination estimates). *Consistency estimates* include NRT approaches (Cronbach alpha, split-half adjusted, and Guttman estimates), and CRT methods (phi domain score dependability index and phi(lambda) squared-error loss agreement coefficient). The NRT standard error of measurement is reported, as well as the analogous CRT confidence intervals. *Validity* is discussed in terms of content validity, and the construct validity strategy is also considered from the intervention and differential groups perspectives.

RESULTS

Descriptive Statistics

The descriptive statistics for this study are presented in Table 1 (p. 108). They include the number of students (N) who took each version, the number of items involved (k), the mean, the standard deviation (SD), minimum score (Min.), maximum score (Max.) and range. These statistics are given for each of the forms (A and B) when administered at the beginning of the course (Pre) as well as at the end (Post). Where no results are shown, the test was either not ready at the time (e.g., ELI 70 PreA and PreB) or inadvertently omitted (e.g., ELI 73 PostA).

Item Statistics

The mean item statistics for this project are shown in Table 2 (p. 109). They include NRT, IRT and CRT estimates, all of which are being used in our thinking about the item selection and test revision. Naturally, we are much more interested in the statistics for each individual item. However, mean item statistics are the only practical way to provide readers with at least an overview of the present state of these tests.

Table 1
Descriptive Statistics

	COURSE	STATISTIC						
	Test	N	k	Mean	SD	Min.	Max.	Range
R	ELI 72							
E	PreA	35	46	31.11	5.18	15	39	25
A	PreB	29	46	30.90	5.47	16	41	26
D	PostA	26	46	34.73	4.86	20	42	23
I	PostB	35	46	33.57	3.81	28	41	14
N	ELI 82							
G	PreA	87	34	21.05	3.95	13	31	19
	PreB	65	34	21.26	3.92	10	30	21
	PostA	63	34	23.44	3.94	14	31	18
	PostB	67	34	23.12	3.90	14	31	18
W	ELI 73							
R	PreA	41	50	33.71	3.49	25	41	17
I	PreB	23	50	32.35	5.55	18	41	24
T	PostA	--	--	---	--	--	--	--
I	PostB	64	50	33.78	5.81	16	45	30
N	ELI 83 (In-class essay not included)							
G	PreA	--	--	---	--	--	--	--
.	PreB	--	--	---	--	--	--	--
.	PostA	37	9	4.27	2.18	0	8	9
.	PostB	--	--	---	--	--	--	--
.	ESL 100 (In-class essay not included)							
.	PreA	47	32	24.89	3.24	16	30	15
	PreB	--	--	---	--	--	--	--
	PostA	67	32	27.90	3.54	11	32	22
	PostB	--	--	---	--	--	--	--
L	ELI 70							
I	PreA	--	--	---	--	--	--	--
S	PreB	--	--	---	--	--	--	--
T	PostA	122	24	16.07	3.07	5	22	18
E	PostB	122	24	16.30	3.07	9	22	14
N	ELI 80 (In-class presentation not included)*							
I	PreA	112	24	14.80	4.15	5	24	20
N	PreB	117	24	14.27	3.41	4	23	20
G	PostA	95	24	15.71	2.91	7	22	15
	PostB	116	24	15.36	3.31	6	23	18

Table 2
Mean Item Statistics

	COURSE	STATISTIC									
		NRT		IRT				CRT			
	Test	IF	ID	P	Dif.	Dis.	Used I(P)	DI	Item ϕ (cut-point)**	B	A
R	ELI 72										
E	PreA	.68	.23	.67	-1.46	.38	45(35)*	.08	.07	.18	.34
A	PreB	.67	.25	.67	-1.31	.44	46(29)*	.06	.09	.23	.36
D	PostA	.76	.20	.74	-2.01	.42	44(26)*	.08	.18	.33	.76
I	PostB	.73	.19	.71	-1.96	.38	43(35)*	.06	.00	.00	.73
N	ELI 82										
G	PreA	.62	.29	.62	-0.99	.36	34(87)	.07	.07	.30	.39
	PreB	.63	.28	.63	-1.14	.37	34(65)	.05	.07	.26	.39
	PostA	.69	.25	.69	-1.54	.40	34(63)	.07	.20	.21	.68
	PostB	.68	.25	.68	-1.45	.38	34(67)	.05	.19	.17	.64
W	ELI 73										
R	PreA	.67	.00	.67	-2.47	.24	50(41)	NA	.04	.11	.33
I	PreB	.65	.24	.65	-1.27	.35	50(23)	.03	.09	.18	.38
T	PostA	--	--	--	--	--	---	--	--	--	--
I	PostB	.68	.24	.67	-1.46	.41	50(64)	.03	.21	.22	.68
N	ELI 83										
G	PreA	--	--	--	--	--	---	--	--	--	--
.	PreB	--	--	--	--	--	---	--	--	--	--
.	PostA	.47	.52	.52	-0.03	.85	9(34)*	NA	.43	.39	.69
.	PostB	--	--	--	--	--	---	--	--	--	--
.	ESL 100										
.	PreA	.78	.07	.78	-2.12	.50	32(47)	.09	.06	.10	.28
	PreB	--	--	--	--	--	---	--	--	--	--
	PostA	.87	.19	.87	-2.32	.68	31(66)*	.09	.30	.39	.87
	PostB	--	--	--	--	--	---	--	--	--	--
L	ELI 70										
I	PreA	--	--	--	--	--	---	--	--	--	--
S	PreB	--	--	--	--	--	---	--	--	--	--
T	PostA	.67	.27	.66	-1.12	.44	23(122)*	NA	.23	.23	.66
E	PostB	.68	.28	.68	-1.38	.44	24(122)	NA	.25	.24	.67
N	ELI 80										
I	PreA	.62	.39	.61	-0.64	.56	24(111)*	.03	.17	.35	.43
N	PreB	.60	.31	.60	-0.67	.47	24(117)	.05	.07	.37	.41
G	PostA	.65	.26	.65	-1.24	.37	24(95)	.03	.23	.22	.65
	PostB	.64	.26	.64	-0.98	.48	24(116)	.05	.24	.23	.64

* Either an item or person (or both) was deleted because 0% or 100%.

** Cut-points were set at .90 for pretest decisions and .60 for posttests.

The means for the *norm-referenced test* estimates include traditional item facility and item discrimination indexes, which suggest that we have created tests that look statistically very much like norm-referenced tests for placement. Indeed, if we were to proceed in selecting items on the basis of these norm-referenced statistics, the tests would probably become increasingly powerful as NRTs. Instead, we have chosen to use the two other types of item analyses to tailor our tests for criterion-referenced purposes.

The *item response theory* item estimates were calculated using a one-parameter model. Our primary purpose in using IRT was to include the item difficulty estimates in our thinking. Notice that, in all cases, the mean difficulty estimates (Dif.) are negative indicating that the items are on average relatively easy for the students, more so on the posttests than pretests, but nevertheless somewhat easy. Note also that the discrimination estimate reported is the slope (which is held constant across all items in a one-parameter analysis).

Caution must be used in thinking about these IRT results because, in a number of cases, our sample sizes are too small to be appropriate for even the one-parameter model. We would be much more comfortable if we had at least 100 students in each sampling.

Our principal motivation in using IRT analyses was that we wanted to be able to use the individual student ability estimates for examining appropriate cut-points for pass/fail decisions. They are not given here because the mean ability estimates were zero in all cases. In the long run, we would also like to be able to set up an item bank for each of these courses—a task for which IRT is particularly well-suited.

The *criterion-referenced test* item statistics include the difference index (DI), item ϕ, the B-index and the agreement index (A) as described in Shannon and Cliver (1987) and Berk (1984b). The DI is calculated for each item by subtracting its item facility on the pretest from the facility for the same item on the posttest. Item ϕ is an estimate of the degree to which the students' item performances (right or wrong) are related to whether or not they passed the test. The B-index is the difference between proportions of correct answers on each item and the proportions of students passing and failing. The agreement statistic is defined "as the proportion of consistent item-test outcomes" with

regard to those students who correctly answered the item and passed the test, and those who missed the item and failed the test. Thus the agreement statistic is similar at the item level to the agreement coefficient used to explore the overall consistency, or dependability, of tests in decision making (see Cohen 1960; Subkoviak 1980, 1988).

It is important to recall that we are using these CRT statistics not as the averages summarized in Table 2, but rather on an item-by-item basis. It is also important to note that we calculated each of them for .50, .60, .70, .80, and .90 decision levels. This has proven very useful in thinking about item selection in terms of the kinds of decisions that we make with the tests, as well as the relative appropriateness of various cut-points for our decision making. We have two types of decisions that we must make on the basis of these tests. The pretest results are used, among other things, for finding those students who were misplaced and should be moved up to the next level or be exempted; the posttest administrations are used primarily to decide whether or not students should fail the course. We have tentatively set our decision levels at about .90 for pretest exemption from the course, and at about .60 for posttest pass/fail determinations. The values reported in Table 2 are therefore based on .90 for pretests and .60 for posttests. Ultimately, we want to select those items which are strong for both types of decisions.

Consistency Estimates

Table 3 (p. 112) presents both NRT reliability statistics and CRT dependability estimates. The NRT *reliability estimates* include Cronbach alpha, the split-half method (adjusted by the Spearman-Brown prophecy formula), and the Guttman coefficient. These NRT coefficients appear to be fairly low. However, it is important to remember that the ranges of talent in these courses have been severely restricted by previous NRT selection procedures for admissions and placement. As demonstrated in Brown (1984b), Ebel (1979), and elsewhere, even a good test may appear to be unreliable if the range of talent is depressed. Given that information, the reliability estimates produced by most of these tests are fairly respectable, even from an NRT perspective.

Table 3
Reliability and Dependability

	COURSE Test	STATISTIC NRT Alpha	Odd-even	Guttman	SEM	CRT Phi	Phi(lambda)*	CI
R	ELI 72							
E	PreA	.704	.814	.809	2.234	.674	.928	.068
A	PreB	.750	.822	.816	1.879	.713	.932	.068
D	PostA	.713	.785	.784	2.253	.691	.892	.062
I	PostB	.573	.661	.660	2.218	.497	.823	.065
N	ELI 82							
G	PreA	.575	.651	.651	2.334	.541	.927	.082
	PreB	.586	.722	.722	2.068	.546	.925	.082
	PostA	.617	.531	.529	2.695	.584	.719	.078
	PostB	.587	.650	.649	2.305	.562	.686	.079
W	ELI 73							
R	PreA	.276	.555	.554	2.326	.239	.921	.215
I	PreB	.703	.687	.674	3.107	.683	.943	.066
T	PostA	—	—	—	—	—	—	—
I	PostB	.750	.789	.786	2.669	.714	.784	.065
N	ELI 83							
G	PreA	—	—	—	—	—	—	—
.	PreB	—	—	—	—	—	—	—
.	PostA	.712	.739	.738	1.112	.650	.688	.154
.	PostB	—	—	—	—	—	—	—
.	ESL 100							
.	PreA	.630	.690	.689	1.802	.609	.811	.072
	PreB	—	—	—	—	—	—	—
	PostA	.793	.863	.861	1.319	.757	.963	.057
	PostB	—	—	—	—	—	—	—
L	ELI 70							
I	PreA	—	—	—	—	—	—	—
S	PreB	—	—	—	—	—	—	—
T	PostA	.554	.583	.581	1.984	.509	.582	.094
E	PostB	.551	.497	.497	2.179	.512	.615	.094
N	ELI 80							
I	PreA	.725	.814	.812	1.788	.711	.919	.095
N	PreB	.600	.734	.730	1.757	.562	.916	.098
G	PostA	.428	.535	.533	1.981	.411	.483	.095
	PostB	.594	.727	.727	1.731	.556	.559	.096

* Cut-points were set at .90 for pretest decisions and .60 for posttests.

From the CRT viewpoint, the phi coefficients are domain score estimates of the dependability of these tests, while the phi(lambda) coefficients are decision consistency estimates based on the squared-error loss agreement approach (see Berk 1980, 1984a). Both phi and phi(lambda) are based on the short-cut formulas presented in Brown (1989c). Like the CRT item statistics, the phi(lambda) estimates are for .90 cut-points on the pretests and .60 cut-points on the posttests.

Notice that the standard error of measurement (SEM) is presented just to the right of the NRT reliability estimates. In this case, the SEM is based on the odd-even, or split-half (adjusted), coefficients. Notice also that a statistic called the confidence interval (CI) is presented in the column furthest to the right. This confidence interval (which ranged from .057 to .215 in these data) is analogous to the SEM, but is appropriate for use with CRTs. It should be interpreted as the proportion of error that would be accounted for with 68 percent confidence around an individuals proportion score. For example, the CI in the lower-right corner of Table 3 would indicate that a person receiving a proportion score of .80 (or 80 percent) would score within plus or minus one CI, or a band from .704 (.80 – .096 = .704) to .896 (.80 + .096 = .896) 68 percent of the time. In percent score terms, this would simply be a band between 70.4 percent and 89.6 percent. The CI is derived from a statistic called the absolute error variance component in generalizability theory (see Bolus, Hinofotis & Bailey 1982; Brennan 1980, 1984; Brown 1984c; Brown & Bailey 1984).

Validity

Essentially two strategies are practical and appropriate for investigating the validity criterion-referenced tests: the content and construct approaches.

Content validity involves the systematic study of the degree to which the items on a test match the content that the test was designed to measure. Content validity has become an integral part of the item development process at UHM in the sense that items are always written to closely match the objectives of our courses. Thus we are constantly considering content validity on an item-by-item basis and subtest-by-subtest. Since the items are written by the teachers of the courses and carefully reviewed by the lead teachers and director of the ELI, the items are not only expected to match the objectives, but

also to match those objectives as they are actually addressed in the classrooms. When we become reasonably comfortable with the tests in terms of dependability and validity, we will no doubt turn to outside "experts" in order to get independent judgments of the degree to which the items match our objectives, and obtain their insights as to the degree to which our objectives match the students' needs (see Popham 1978, 1981 for more on such strategies).

Construct validity involves the experimental demonstration of the degree to which a test is measuring the psychological construct it claims to be measuring. Such demonstrations can take many forms, but for CRTs the intervention and differential groups methods are the most practical and appropriate strategies.

A typical *intervention study* is one in which a test is administered, the students are taught whatever construct is involved, then they are tested again. If the test is actually measuring the construct, the students should score significantly higher on the posttest than they did on the pretest. In this way, one argument can be built for the construct validity of the test.

The differences found in the present study between pretest and posttest means indicate that, in every case, there was some effect on the scores due to instruction. These gains range from three to nine percent as indicated by the difference indexes (DI) reported in Table 2. It is expected that the actual gains experienced by the students who took our courses are somewhat higher for two reasons:

1) The results reported here include all students who took the pretest and posttest administrations. Thus those students who scored high on the pretest and were exempted from the courses did not take the posttest. This fact would have the effect of diminishing the observed differences. In future analyses, these exempted students will be separated out of the pretest results. As a result, only those students who actually received instruction will be included in the analysis.

2) In most cases, these tests have not yet been substantially revised to select those items which are most sensitive to instruction. When this occurs, i.e., when those items with the largest difference indexes are selected and the tests are further strengthened using the other item

> statistics, it is expected that much larger gains will be reflected in the tests. This does not necessarily mean that the students will be learning more, but rather that the tests will become more sensitive to that which they do learn.

It is also important to note that we cannot attribute these gains solely to the effects of our courses because students were simultaneously being exposed to English from many other angles in their daily lives, and because many of them were also taking other ELI courses which could have affected their English. Nevertheless, we can interpret the differences as reflecting gains due to the total English language experience that students had during that semester at UHM.

For all of the above reasons, it was felt that it would be premature to perform statistical analyses of the current differences, especially before addressing the issues explained in 1) and 2) above. Nevertheless, the intervention study approach to CRT construct validity is very much on our minds. At the test level, we are considering the mean gains. At the item level, we are choosing the items that will remain on future revised versions of the tests on the basis of the difference index. Thus construct validity is particularly important for insuring that our CRTs have a fairly strong relationship with the learning that is occurring in our courses.

Another approach to construct validity also figures into our thinking. The *differential groups approach* usually involves administering a test to a group of students who can be said to possess the construct in question (masters), as well as to another group (nonmasters), who lack it (see Brown 1984c for an example of this approach). We have used this differential groups approach in two ways. First, we have compared the performances of students who passed the courses (masters) with that of students who failed (nonmasters). Naturally, there were large differences between those groups because passing or failing was partially determined by the test itself. Second and more important, we have examined the individual and mean item ϕ, the B-index and A. They generally indicate that a fairly strong relationship exists between the accuracy of the students' answers on individual items, and whether or not they pass the course. This

type of validity is especially important in thinking about the degree to which our pass/fail decisions are fair.

We have found one aspect of CRT validity to be particularly satisfying: the fact that content, intervention and differential groups strategies are built directly into the item development and item analysis processes. Thus item selection and test revision are integrally related to analyzing and improving test validity. As always with issues of validity, the goal is to marshall evidence from a variety of sources so that, collectively, they can be used to investigate (and perhaps support) the validity of the test in question.

DISCUSSION

TO SUMMARIZE BRIEFLY and return to the original research questions, we have found that the CRTs developed as of Fall semester 1989 are functioning reasonably well. From a norm-referenced point of view, the tests appear to be functioning about the same across skills, levels, and courses. They are reasonably well-centered and dispersing students. From a criterion-referenced point of view, the tests generally appear to be too easy at the beginning of the course and too difficult at the end. The item selection processes and revisions that we make will be aimed at improving this situation so that: 1) the tests better reflect any learning that is occurring and 2) the tests help us to make fair decisions with regard to passing or failing courses.

In this effort, it seems that all of the item statistics are proving to be useful. The NRT statistics are helping us to examine our tests in terms with which we have long been familiar. In addition, items that do not turn out to be useful in the criterion-referenced tests may later serve as new items for our NRT placement tests. Thus NRT item statistics may continue to prove useful in the future. Similarly, the increasing use of IRT approaches to item analysis is expected to help us in setting up item banks, and in making better pass/fail decisions in the future.

The tests also appear to be at least moderately reliable from the NRT perspective on that concept, especially given the restrictions of range in these

courses. From a CRT viewpoint, the tests also appear to be moderately consistent in terms of domain score dependability (as indicated by phi). However, the dependability of these tests, as estimated by phi(lambda), seems to be more uneven. The estimates range from very high to very low depending on the test and cut-point. Further analysis of these related issues must be considered when we are making the actual pass/fail decisions. In addition, we must pay careful attention to the confidence interval statistics and obtain additional information about students who fall close to our cut-points—at least for those students who fall within one such CI, plus or minus, of the cut-point. Thus for pass/fail decisions, the CRT dependability approaches and the CIs will be much more useful than the analogous NRT reliability estimates.

The validity issue will also be an ongoing one. We can say with some pride that there was no item in this study that was not been carefully scrutinized for content validity by the appropriate ELI teachers. In addition, all tests in this study showed some sensitivity to instruction. We should nevertheless learn from our experiences here so that future studies will exclude from the analysis those students who are exempted on the basis of their pretest scores.

CONCLUSIONS

THIS SECTION briefly touches on the problems encountered in developing a comprehensive CRT program like the one described in this paper, then turns to the benefits which CRTs can provide for overall curriculum development.

Problems

Teacher cooperation. This process of test development was made relatively efficient, indeed was made possible, by the appointment of a lead teacher, who was given 50 percent release time to help in administering, scoring, developing, analyzing, and revising ELI tests. During the norm-referenced placement testing, this lead teacher takes responsibility for administering the tests, while

the director does the scoring and analysis of the results as each subtest is completed.

During the remainder of the semester, when our attention turns to the CRT diagnostic and achievement decisions, the lead teacher is essential in rallying the teachers to write, review, and revise items. This lead teacher is also responsible for getting the tests to the teachers for administration in their classes, for scoring the tests and for getting the results back to the teachers within 24 hours. Such promptness has made the results particularly useful and has helped foster teacher support for the testing program.

Magnitude of project. In our present situation, it is useful that the director is a language testing specialist and that the lead teacher for testing is typically a graduate assistant who excelled in our language testing course. The sheer number of tests involved in this project along with the wide variety of statistics that are necessary make such a project fairly laborious. The central message seems to be that adequate resources in terms of expertise, time, money and computer equipment must be allocated before any such project can succeed on this scale. For smaller programs with fewer courses, a modified version of this project starting out with fewer tests and more select statistics would seem to be feasible. Naturally, these issues should be considered long before any such testing project is initiated.

Objectives that fail. It is also important to recognize that we do not always learn what we set out to learn. For instance, the first versions of the ELI 72 reading test were developed about three years ago. The first administration of these items indicated that the students already knew virtually all of the material—at the beginning of the course. We found ourselves in the uncomfortable position of realizing that our objectives (ones that had been used for years) were aimed far too low for the abilities of our students. As a result, we had to throw out much of the test and revise our objectives considerably. In this reading course, we were able to do so by using similar objectives but applying them to much more difficult textual material (college level texts). It initially hurt to realize that all of our efforts in developing that early test were

for nothing, but in retrospect, it is clear that this early attempt at a CRT and the subsequent failure have served to change our views of our students' language needs. This effect has benefitted not only the reading curriculum, but the other skills as well.

Benefits of a Criterion-Referenced Testing Program

Criterion-referenced tests are not easy to develop, implement, analyze and revise. In truth, such a project requires a prodigious amount of work. However, a determined group of teachers and administrators did create these tests, and did so in multiple forms. The pay-off is that the information that we derive from them is directly applicable to what we are doing in the ESL classroom AND helps us to improve all of the elements of the curriculum design process (needs, objectives, materials, tests, teaching and program evaluation).

First, the criterion-referenced tests helped us to closely re-examine our perceptions of the students *needs*. Objective by objective, we can now consider how the students perform at the beginning and end of each course on each objective, and the degree to which we have defined our objectives in clear and observable terms. Sometimes our initial perceptions turn out to have been wrong and, as described in the previous section, major portions of tests must be revamped. However, we feel that this is better than blithely continuing to teach material that our students do not need to learn.

Second, knowing which *objectives* are (and which are not) working allows us to streamline and concentrate on objectives that reflect the students' needs while adding others that are designed to meet other needs. This strategy enables us to avoid the waste of time and effort involved in teaching material that the students already know. We can instead focus on that which the students actually need to learn, and do so much more efficiently. Perhaps we are succumbing to what Tumposky (1984) sarcastically labeled the "cult of efficiency." However, we are defining the objectives in so many different ways (ranging from experiential to instructional objectives) that her complaints no longer seem applicable. Frankly, we see no problem with attempting to be relatively efficient in the delivery of ESL instruction to students who pay good money for it. In our view, we are simply trying to foster as much language

learning as we can during the short period of time that we have with our students.

Third, having criterion-referenced *tests* in place allows us the luxury of working together as groups of teachers, rather than in isolation, to build, implement, analyze and revise classroom tests that are relatively effective. An additional effect of having criterion-referenced tests in place is that information gained from this type of testing can be utilized in improving other types of tests. For instance, information gained from CRT achievement tests can prove useful in revising the placement procedures in order to overcome the mismatch that sometimes occurs between placement batteries and the courses to which they are supposed to be related (as noted in Brown 1984b). One such process of modifying placement procedures to more closely align them with courses is described in Brown (1989b), other strategies are currently being explored at UHM.

Fourth, modifying the objectives based on what we learn from our diagnostic and achievement tests naturally leads to rethinking our *materials* so that they better match the newly perceived needs of the students. Sometimes, these have proven to be large changes, but more often they have taken the form of incremental modifications in materials, teaching techniques and practice exercises. In all cases, the tests help us to make choices in gauging the correct level and objectives for the textbooks that we adopt, the materials modules that we develop, and the lessons that we teach.

Fifth, the goal of all of our curriculum activities is to support *teaching* so that the teachers can do what they do best—teach. One perspective that we take on our criterion-referenced tests is that they provide a way of helping teachers do rational and well-designed achievement testing (which is mandated in all undergraduate courses at UHM). While the teachers are welcome to add sections of their own devising to the final examination for their individual courses, the core test is essentially provided for them. In addition, these tests were jointly developed by the teachers in each course over a period of three years. The tests must constantly be reviewed and revised to insure that they match the objectives of the courses as they are currently taught. Because the tests are so important to the students and teachers, they are reviewed and

revised by the groups of teachers most directly involved. As such, they form an important focus on which the teachers can concentrate, while working constructively together toward a common goal.

The sixth and last benefit derived from our testing program has to do with program *evaluation*. In the formative sense of program evaluation, the tests clearly help us to modify our curriculum as it continues to develop. However, if called upon to perform program evaluation in the summative sense, the tests will also put us in a very strong position. When we do need to focus on summative program evaluation (in about two years for our program review), we will have a staggering amount of information ready to be presented. This will include norm-referenced information about the overall proficiency of our students in terms of their TOEFL scores for admissions, as well as information about their placement based on the six subtests of the ELIPT. In addition, the criterion-referenced tests will supply data about the students' knowledge at the beginning (diagnostic) and end (achievement) of each course, as well as about what and how much the students have learned in our courses. At the very least, we will clearly be in a position to fashion a summary report that describes our program in terms of student needs, program goals and objectives, materials, and teaching. We will also be in a strong position for suggesting clear-cut changes in the program in an on-going process of curriculum development.

ELI testing program. Naturally, we hope that a majority of the students who are served by the procedures discussed above are correctly admitted, placed, diagnosed and promoted. However, decisions are made by human beings and, even when they are based on seemingly scientific test scores, human judgments can go awry. The problem is that an incorrect decision may cost a student a great deal in the form of extra tuition or extra, unnecessary time spent studying ESL. Hence the decisions that we make about our students' lives are taken very seriously and based on the best available information—information from a variety of sources including criterion-referenced tests.

Certainly, all of this requires more effort on the part of the administrators and teachers, but the benefits gained from effective and humane testing procedures accrue to all—students, teachers and administrators alike. It is

hoped that the strategies, which we find so useful, can be generalized and adapted to other language programs as well.

Received September 15, 1990.

Author's address for correspondence:
J.D. Brown
Department of English as a Second Language
University of Hawai'i
1890 East-West Road
Honolulu, HI 96822

REFERENCES

Bachman, L.F. (1989). The development and use of criterion-referenced tests of language ability in language program evaluation. In K. Johnson (Ed.). *Program design and evaluation in language teaching*. London: Cambridge University.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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

Berk, R.A. (Ed.). (1984a). *A guide to criterion-referenced test construction*. Baltimore: Johns Hopkins University Press.

Berk, R.A. (1984b). Selecting the index of reliability. In R.A. Berk(Ed.) *A guide to criterion-referenced test construction*. Baltimore: Johns Hopkins University Press.

Bolus, R.E., F.B. Hinofotis & K.M. Bailey. (1982). An introduction to generalizability theory in second language research. *Language Learning, 32*, 245–258.

Borland (1989). *QuattroPro*. Scotts Valley, CA: Borland International.

Brennan, R.L. (1980). Applications of generalizability theory. In R.A.

Berk (Ed.) *Criterion-referenced measurement: the state of the art*. Baltimore: Johns Hopkins University Press.

Brennan, R.L. (1984). Estimating the dependability of the scores. In R.A. Berk (Ed.) *A guide to criterion-referenced test construction*. 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 building bridges: research and practice in teaching English as a second language*. Washington, DC: TESOL.

Brown, J.D. (1984a). Criterion-referenced language tests: what, how and why? *Gulf Area TESOL Bi-annual, 1*, 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: the question of control*. Washington, DC: TESOL.

Brown, J.D. (1984c). A norm-referenced engineering reading test. In A.K. Pugh & J.M. Ulijn (Eds.) *Reading for professional purposes: studies and practices in native and foreign languages*. London: Heinemann Educational Books.

Brown, J.D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. London: Cambridge University.

Brown, J.D. (1989a). Improving ESL placement tests using two perspectives. *TESOL Quarterly*, 23, 1.

Brown, J.D. (1989b). Language program evaluation: a synthesis of existing possibilities. In K. Johnson (Ed.). *Program design and evaluation in language teaching*. London: Cambridge University.

Brown, J.D. (1989c). Short-cut estimates of criterion-referenced test consistency. *University of Hawai'i Working Papers in English as a Second Language, 8*, 1.

Brown, J.D. (1990). Where do tests fit into language programs? *JALT Journal, 12*, 1.

Brown, J.D. (in preparation). *The systematic development of language curriculum.* Honolulu, HI: University of Hawai'i at Manoa.

Brown, J.D. & K.M. Bailey. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*, 4, 21–42.

Cartier, F.A. (1968). Criterion-referenced testing of language skills. *TESOL Quarterly*, 2, 27–32.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20. 37–46.

Cziko, G.A. (1982). Improving the psychometric, criterion-referenced, and practical qualities of integrative language tests. *TESOL Quarterly*, 16, 367–379.

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

Delamere, T. (1985). Notional-functional syllabi and criterion-referenced tests: the missing link. *System*, 13, 43–47.

Ebel, R.L. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

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

Henning, G. (1987). *A guide to language testing: Development,evaluation, research.* Cambridge, MA: Newbury House.

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

Jacobs, H.L., S.A. Zinkgraf, D.R. Wormuth, V.F. Hartfiel & J.B. Hughey. (1981). *Testing ESL composition: a practical approach.* Rowley, MA: Newbury House.

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

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

Richards, J.C., J. Platt and H. Weber. (1985). *Longman dictionary of applied linguistics*. London: Longman.

Shannon, G.A. & B.A. Cliver. (1987). An application of item response theory in the comparison of four conventional item discrimination indices for criterion-referenced tests. *Journal of Educational Measurement*, 24, 347–356.

Subkoviak, M.J. (1980). Decision-consistency approaches. In R.A. Berk, (Ed.). *Criterion-referenced measurement: the state of the art*. Baltimore: Johns Hopkins University Press, 129–185.

Subkoviak, M.J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, 25, 47–55.

SYSTAT (1987). *TESTAT*. Evanston, IL: SYSTAT.

Tumposky, N.R. (1984). Behavioral objectives, the cult of efficiency and foreign language learning: Are they compatible? *TESOL Quarterly*, 18, 2: 295–310.