IMPROVING ESL PLACEMENT TESTS USING TWO PERSPECTIVES[1]

JAMES DEAN BROWN
University of Hawai'i at Manoa

The first contact that many students have with an ESL or EFL program is the relatively cold, detached and "objective" experience of taking some form of placement examination. This is an important element in most programs because of the necessity for sorting students into relatively homogeneous proficiency groupings, sometimes within specific skill areas. It would seem, if we are going to put students through this experience, that we should do the best job possible of making responsible placement decisions based on the results of their efforts.

Very often, the tests that are used for such placement decisions are bought from commercial publishing houses, adopted from other ESL programs or pulled straight from the current textbook. Given the wide diversity and variation in the nationalities and levels involved in the various ESL/EFL programs around the world, it is reasonable to assume that many of these tests are being used with populations quite different from those envisioned when the tests were originally normed. One result of such practices could be that many placement decisions, ones that can dramatically affect the lives of the students, may be irresponsibly based on tests made up of test questions quite unrelated to the needs of the particular students in a given language program or to the curriculum being taught in that program.

In rare situations, sufficient expertise is available within a program so that placement tests can be developed and normed on the basis of the population actually studying in the center. This would seem to be an ideal situation, but Brown (1981) found that even when a relatively sophisticated test like the English as a Second Language Placement Test (ESLPE) is developed specifically for a given program, in this case the service courses at UCLA, a serious mismatch can occur between what is being tested by the placement examination and what is being taught/learned in the program into which the students have been placed.

In that 1981 study, two groups who sat side-by-side in the UCLA 33C

[1] This paper is a much revised version of a paper presented at the 21st Annual TESOL Convention in Miami, Florida in April 1987.

University of Hawai'i Working Papers in ESL, Vol. 7, Number 1, March 1988, pp. 239-260.

(highest level) service course were compared: students placed directly into the course (Placed Ss) and others who were continuing from lower level courses (Continuing Ss). These two types of students were compared for three successive quarters across three dependent variables: course grade, final examination scores and cloze test scores. Multivariate analyses (see Table 1) indicated that the Placed Ss significantly ($p < .05$) outperformed the Continuing Ss on all three measures. Thus the two groups of students—those arriving in the course because of there performance on the placement test and those arriving in the course because of successful completion of lower level courses—were found to be different in overall ESL proficiency as indicated by the cloze test and in ESL achievement as indicated by their performance on the final examination and course grades. These results suggest a serious mismatch between what was being measured by the placement procedures and what was being taught in the lower level courses.

In the English Language Institute (ELI) of the University of Hawai'i at Manoa (UHM), we have recently recognized problems analogous to those found at UCLA. Our goal soon became the development of a placement battery that would somehow be related to the curriculum of our institute—a proposal that struck us as strangely novel at the time. The purpose of this project, then, was to develop placement tests related to what our students learn while enrolled in our language courses.

As a starting point, we chose to focus on the reading component of the ELI. This is one of three skills taught by us. The others are listening and writing. Reading was chosen because the subtests used for reading placement where found to be in most serious need of revision, and because the decisions based on the reading test scores were ultimately the simplest, i.e., students are either placed into a single level of reading or exempted altogether from training in the reading skill. The plan was to develop a new, workable strategy for constructing a program-related reading placement test then use the same strategy in developing/revising the other skill tests in our battery.

Table 1: Descriptive Statistics and Mean Differences (Brown 1981)

		SAMPLE			
MEASURE	GROUP	Fall 1977 (n=164)	Winter 1978 (n=82)	Spring 1978 (n=73)	TOTAL (N=319)
		MEAN(SD)	MEAN(SD)	MEAN(SD)	MEAN
COURSE	Placed	2.99(.62)	3.21(.50)	3.35(.50)	3.13
GRADE	Continuing	2.04(1.04)	2.88(.51)	2.83(.27)	2.44
	Difference	.95*	.33*	.52*	.69
FINAL	Placed	67.83(7.89)	78.33(8.33)	78.15(7.03)	72.89
EXAM	Continuing	55.31(9.66)	73.36(7.01)	68.97(6.01)	63.07
	Difference	12.52*	4.97**	9.18*	9.82
CLOZE	Placed	22.97(4.56)	24.22(4.26)	23.12(4.72)	23.32
	Continuing	15.87(4.57)	18.56(4.92)	16.09(5.91)	16.61
	Difference	7.10*	5.66*	7.03*	6.71

* $p < .01$
** $p < .05$

In order to clearly explain what took place in this test development process and how to apply it to other teaching situations, there are a few very basic concepts and terms that should be clarified. The first is the distinction between norm-referenced tests and criterion-referenced tests. Next, three very simple item analysis statistics will be discussed: the item facility, item discrimination and item difference indices. Since the overall purpose of this paper is to suggest a model to help others develop placement tests that reflect what is being taught and learned in their programs, an effort will be made to ensure that all technical jargon is defined and explained in such a way that all readers will understand what is going on in the study.

Norm-referenced versus Criterion-referenced Tests

One useful distinction that helps ESL teachers understand the different purposes of language tests is that between norm-referenced tests and criterion-references ones. These are two terms which may not be entirely familiar to readers because the distinction is relatively new in our field (see Cziko 1983, Brown 1984a and Hudson & Lynch 1984) though it has been around in educational testing circles for years (see Popham & Husek 1969, Popham 1978, 1981, and Berk 1980 for much more on criterion-referenced testing and its background). In fact, the notion of criterion-referenced, as distinct from norm-referenced, testing dates back to Glaser 1963.

In general terms, **norm-referenced tests** (NRTs) are designed to measure global langauge abilities or proficiencies that a student may have developed (e.g., overall English language proficiency, academic listening ability, reading comprehension, etc.). Each student's score on an NRT is interpreted relative to the scores of all of the other students who took the test. This is done with reference to the statistical concept of normal distribution (familiarly known as the "bell curve") of scores dispersed around a mean, or average. The purpose of an NRT is to spread students out along a continuum of scores so that those with "low" abilities in a general area such as reading comprehension end up on one end of the normal distribution, while those with "high" abilities are found at the other (with the bulk of the students found in between the extremes clustered fairly closely around the mean). Another characteristic of NRTs is that, even though the students may know the general form that the questions will take on an examination (e.g., multiple-choice, true-false, etc.), they typically have no idea what specific content or skills will be tested by those questions.

Criterion-referenced tests (CRTs), on the other hand, are produced to measure well-defined and fairly specific instructional objectives. Often these objectives are specific to a particular program, school district, or state. It is therefore important, in most cases, that the students and teachers know exactly what those objectives are so that time and attention can be focused on them during the appropriate course(s). The interpretation of CRTs is considered absolute in the sense that each student's score is meaningful unto itself without reference to the other students' scores. In other words, a student's score on a particular objective indicates the percentage of the skill or knowledge in that objective which has been learned or acquired. Moreover, the distribution of scores on a CRT need not necessarily be normal. If all of the students know 100% of the material on all of the objectives, it follows that all of the students would receive the same score with no variation at all. The purpose of CRTs, then, is to measure the amount of knowledge or skill that the students have developed on a specific set of objectives. In most cases, the students would know in advance what types of questions, tasks and/or content to expect for each objective on such a test because it would be implied (if not explicitly stated) in the objectives of the course.

The discussion of NRTs and CRTs has centered, to this point, on practical and important differences in the type of measurement involved, the way scores are interpreted and distributed, the purpose for giving each type of test and the students' knowledge of question content. This is summarized in Table 2. There are also numerous differences between NRTs and CRTs in the ways that they are viewed empirically and treated statistically (see Hudson & Lynch 1984), but for the purposes of this paper, a basic understanding of how they differ in item characteristics is of prime importance.

Item Statistics

When considering item characteristics, the unit of focus is the individual test question (also known as an item). The item characteristics of NRTs are most often described in terms of item facility and item discrimination, whereas CRTs are more appropriately characterized by the item facility (usually pretest and posttest) and item difference indices. Each of these will be defined in turn then discussed with a focus on how they are used differently in developing each of the two types of tests.

Table 2: Differences between Norm-referenced and Criterion-referenced Tests (adapted from Brown 1984)

CHARACTERISTIC	NORM-REFERENCED	CRITERION-REFERENCED
1. Type of Measurement	To measure general language abilities or proficiencies.	To measure specific objectives-based language points.
2. Type of Interpretation	Relative (a student's performance is compared to that of all other students).	Absolute (a student's performance is compared only to the amount, or percent, of material learned).
3. Score Distribution	Normal distribution of scores around a mean.	If all students know all of the material, all should score 100%.
4. Purpose of Testing	Spread students out along a continuum of general abilities or proficiencies.	Assess the amount of material known, or learned, by each student.
5. Knowledge of Questions	Students have little or no idea what content to expect in the questions.	Students know exactly what content to expect in test questions.

Norm-referenced Item Analysis

Item facility (also called item difficulty, item easiness or simply IF) is the percentage of students who answered a given item correctly. This index is calculated by adding up the number of students who responded correctly to a question and dividing that sum by the total number who attempted it. This yields an index which ranges from 0 to 1.00. The index can be interpreted as the percentage of correct answers for that question by moving the decimal point two places to the right. For example, a correct interpretation of an IF index of .27 would be that 27 percent of the students correctly answered the item. In most cases, this would be a very difficult question because many more students missed it than answered it correctly. Conversely, an IF of .96 would indicate that 96 percent of the students answered correctly—a very easy test item because almost everyone answered correctly.

Item discrimination is an index of the degree to which an item separates the "good" students from the "bad" ones.[2] It is often calculated by contrasting the performance of the upper third of the students on the test with that for the lower third. This is done by determining which students had scores in the top third of the group on the whole test and which had scores in the bottom third. The IF for each item is then calculated for the two groups separately; then the IF for the lower third of the students on the whole test is subtracted from the IF for the top third. ID indices can range from -1.00 (if all of the low students answer correctly and all of the high ones answer incorrectly) to +1.00 (if all of the high students answer correctly and all of the low ones answer incorrectly) and, of course, can be everything in between as well.

If, for instance, those students who scored in the top third on a test had an IF of .90 for item 4 and those in the lower third had an IF of .20, the item discrimination index for that item would be .90–.20 = .70. This would indicate that the item was "discriminating," or distinguishing, very well between the "high" students and "low" students on the whole test. On the other hand, an item for which the upper 1/3 had an IF of .10 and the lower 1/3 an IF of .71 would have an ID of .10–.71 = -.61. This would indicate that the item was somehow testing something quite different from the rest of the test because those who scored low on the whole test managed to correctly answer this item while those who scored high on the total test were answering it incorrectly. Since the multiple observations of the whole test are generally considered to be a better estimate of the students' actual knowledge or skills than any single item, there is good reason to question the contribution being made to a norm-referenced test by items that have low or negative ID indices.

Another statistic that is often used to examine item discrimination is the point–biserial correlation coefficient. [Indeed, it was used in this study.]

[2] "High" and "low" achievers or "high" and "low" proficiency students are phrases that might equally well be substituted here depending on the testing situation.

This statistic is usually lower in magnitude for a given item when compared directly with the ID given above but is analogous in interpretation (for more on this, see Guilford & Fruchter 1973). Thus no further distinction will be made between the two approaches to item discrimination in this paper.

Norm-referenced test development or revision projects are usually designed to 1) pilot a relatively large number of test items on a group of students similar to the group which will ultimately be examined with the test, 2) analyze the items and 3) select the best items to make up a smaller, more effective revised version of the test.

Ideal items in such a project for development of an NRT would be those with an IF of .50. These would be well-centered items (i.e., 50 percent answer correctly and 50 percent wrong). In reality however, items are generally chosen which fall in a range of IF between .30 and .70. Once it is determined which of the items fall within that acceptable range of IF, those items with the highest ID would be selected so that the test would not only be centered but also discriminate well between the low and the high students. Ebel (1979) has suggested the following guidelines for making decisions based on ID:

.40 and up	Very good items
.30 to .39	Reasonably good but possibly subject to improvement
.20 to .29	Marginal items, usually needing and being subject to improvement
Below .19	Poor items, to be rejected or improved by revision

Of course, these are not meant to be used as hard and fast "rules" but rather should be used as aids in making decisions about which items to keep and which to discard until a sufficient number of items have been found to make up whatever norm-referenced test is being developed. This is a process which is far less scientific than many neophytes would wish it to be.

Criterion-referenced Item Analysis

Notice that the revision process for NRTs is described as being based on a single pilot administration of the test. This is fine because the purpose of an NRT is usually a one-shot determination of the language placement or proficiency of the students in a single group. The piloting of items in a

criterion-referenced test development project is quite different because the purpose for selecting those items is so fundamentally different. Recall that the purpose of a CRT is to assess how much of an objective or set of objectives has been learned by each student. In order to measure such learning, it seems logical that the students should be measured before and after studying the concepts or skills (or whatever it was that was being taught) to determine whether there was any gain or rise in scores. Hence, the piloting of a CRT logically involves using it as a pretest and posttest and comparing results. To limit the practice effect due to taking exactly the same test twice, two forms can be developed with half of the students taking each form on the pretest then the other form on the posttest.

Item analysis of a CRT can then be conducted on the bases of these results. As with NRT item analysis, **item facility** plays an important role. However, there are now two possible item facilities for each item: one for the pretest and one for the posttest. In CRT development, the goal is to find items that reflect what is being learned, if anything. As a result, an ideal item for CRT purposes might be one that had an IF (for the whole group) of .00 at the beginning of instruction and another IF of 1.00 at the end. This would indicate that everyone had missed the item at the beginning of instruction (i.e., they had desperately needed to study this objective) and everyone answered it correctly at the end of the instruction (i.e., they completely absorbed whatever it was that was being taught). Of course, this example is of an ideal item, in an ideal world, with ideal students and an infallible teacher.

Reality may be a bit different. Students arrive in most teaching situations with differing amounts of knowledge. Thus it is unlikely that there will be an IF of .00 for any CRT item that measures any realistic objective, even at the very beginning of instruction. Similarly, students differ in ability and in the speed with which they learn so it is possible, if not probable, that the students will not learn each and every objective to an equal degree. This would mitigate against the possibility that many CRT items will have an IF of 1.00 at the end of instruction. Since it is unreasonable to expect items to be so clear cut, the **difference index** (DI) is used to analyze the degree to which an item is reflecting gain in knowledge or skill. In contrast to item discrimination, which shows the degree to which an NRT item separates the upper 1/3 of students from the lower 1/3 on a given test administration, the difference index indicates the degree to which a CRT item is distinguishing between the students who know the material or have the skill being taught (sometimes called masters) and those who do not (termed nonmasters). To calculate the difference index, the IF for the pretest results (presumably nonmasters) is subtracted from the posttest results (hopefully masters). For example, if the posttest IF for item 10 on a test was .77 and the pretest IF was .22, it would indicate that only 22 percent knew the concept or skill at the beginning of instruction wile 77 percent knew it by the end. That would be an encouraging trend further supported by the relatively high DI difference index for that item of .77–.22 = .55. Other examples of calculations for the DI are shown in Table 4. Note that it can range from -1.00 (indicating that students knew but somehow unlearned the objective in question) to +1.00 (showing that the students went from knowing nothing about the objective to knowing it completely)—and everything in between as well.

Purpose of this Study

To date, the NRT and CRT statistics have been viewed as separate but equal (as in Hudson & Lynch 1984; Brown 1984a). The development of NRTs was seen as a completely different process from the development of CRTs. But as Popham & Husek (1969) pointed out, NRTs and CRTs are indistinguishable to look at; it is only by knowing about the purpose and the analyses that they can be distinguished by most educators. Why then must they be treated so separately?

Brown (1981) provided tangible evidence for a longtime frustration with NRT placement tests. They generally seem to spread students out fairly well, but do not necessarily match what is going on in the ESL program in question. The purpose of this project was to develop and use a new strategy that combines the best qualities of CRTs with those of NRTs to create placement tests which not only spread students out along a continuum of language abilities, but do so on the basis of items that are somehow related to what the students will be taught once they are placed into the program. Having developed such an hybrid test for the ESL reading program at UHM, it was then necessary to examine the following research questions:

1) What are the descriptive statistics for the original and revised versions of the program-related ESL reading comprehension test?
2) What are the item statistics for the original and revised versions of this reading comprehension test?
3) To what degree are the original and revised versions of the test reliable?

4) To what degree are they valid as tests of ESL reading comprehension as it is taught in the ELI?

METHOD

Subjects

All of the subjects in this study were foreign students who were required to take the English Language Institute Placement Test at the University of Hawai'i at Manoa. They were therefore fully admitted students, which meant that they had attained a total score of at least 500 on the Test of English as a Foreign Langauge (Educational Testing Service 1987), also known as TOEFL. If they had scored 600 or higher they were exempted from training in the English Language Institute (ELI). Hence, it is safe to say that the TOEFL scores of these subjects ranged from about 500 to 600.

The groups of students studied here included the entire population of foreign students who took the Fall 1987 ELI Placement Test at UHM (N = 194) as well as the subset of those students who were placed into the reading course (N = 61).

Because of our geographical location, the overall population of 194 students was predominantly Asian with 21 percent from the People's Republic of China, 19 percent from Hong Kong, 11 percent each from Japan and Korea, 9 percent each from Taiwan and Viet Nam, 4 percent each from Indonesia, the Philippines and Thailand. The remaining 8 percent came from a variety of other predominantly Asian countries. They included a mixture of graduate (38 percent), undergraduate (48 percent), unclassified (10 percent) and auditors/visiting faculty (4 percent). The students in this group of 194 were 43 percent females and 57 percent males and had a wide variety of majors with a majority in the sciences.

Materials

At the UHM, placement of students in terms of reading ability has been done primarily on the basis of the Reading Comprehension Test, which was one of the five subtests on the overall ELI Placement Test (ELIPT). The reading comprehension subtest was chosen for this study because it was most urgently in need of revision (last revised in 1983), but also because the objectives for the reading skill were better defined (at the time) than the other three skills.

The "original version" of the UHM Reading Comprehension Test was made up of 60 multiple-choice reading comprehension questions. These included questions on vocabulary, factual questions, inference questions, etc., which were based on 10 academic English reading passages. The passages varied in length from 72 to 290 words.

Procedures

The original version of the Reading Comprehension test was administered as a placement test in Fall 1986. There were 194 students who took this subtest at that time. This administration took place in a large auditorium. The students were allowed 50 minutes to finish the 60 items. It was administered again 16 weeks later as a posttest to the 61 ESL students, who had been placed into the four sections of the reading course. This administration took place in the students' classrooms during their final examination period. They were once again allowed 50 minutes to finish the test.

Analyses

Item selection. This study differs fundamentally from other test revision projects primarily in that the item selection part of the test revision process was not based solely on the NRT item analysis approach using the item facility and discrimination indices, nor solely on the CRT item analysis approach using the mastery/nonmastery item facility and item difference indices. It was based instead on a strategy which combined both of these approaches such that an item was selected for retention in the revised version of the test on the basis of its item facility and item discrimination (when used for placement) as well as its difference index (when viewed as a pretest and posttest for the students who took the course).

The first two criteria helped us select sound NRT items, i.e., those which were effectively spreading students out along a continuum of abilities in reading. The last criterion helped us to select that subset of effective NRT items which was most closely related to something being learned during the 16 weeks that the students had been in our reading course. Of course, the students were doing many other things in their lives, including other ELI courses, that would affect the gains observed here. Nevertheless, we felt that it would be preferable to select items which were somehow related to the ESL reading experience that UHM offered the students than to continue in ignorance of the relationship, if any, between our placement test items and the students' experiences.

Overall design. To that end, the reading comprehension test was administered to all incoming students as a placement test (for IF and ID) and viewed as a pretest-posttest study for the reading course students (DI). Each of these sets of observations was analyzed in terms of descriptive test statistics. The items were then individually analyzed and the goal was to select items for the revised version of the test. Only those which fell approximately within a range of .30 to .70 in IF and had the highest ID and DI indices were to be kept in the revised test. This new version was then reanalyzed for descriptive test statistics and item characteristics in order to determine the degree to which all of this was worth doing at all.

Table 3: Descriptive Statistics

	TOTAL PLACEMENT POPULATION		READING STUDENTS ONLY	
STATISTIC	Original Test	Revised	Pretest	Posttest
N	194.00	194.00	61.00	61.00
k	60.00	35.00	60.00	60.00
$\bar{X}$	36.96	21.44	33.84*	40.93*
SD	10.78	7.10	6.62	7.19
K-R20	.89	.85	.72	.79
SEM	3.51	2.76	3.52	3.28

* Difference between Pretest and Posttest means significant at $p < .01$ (t observed = 6.87; t critical = 2.66 with 60 df)

RESULTS

The descriptive statistics for the original and revised versions of the test, when analyzed separately for the total placement population and for the reading students alone (pretest and posttest), are shown in Table 3. This table includes the number of students in each analysis (N), the number of items (k),the mean ($\overline{X}$), the standard deviation (SD), the Kuder-Richardson formula 20 (K-R20) reliability coefficient and the standard error of measurement (SEM).

Table 4 illustrates the steps involved in calculating each of the difference indices by subtracting the pretest IF for the reading students from there posttest IFs. This indicates the degree to which they have gained (or lost) knowledge or skill on each of the items. A sound CRT item is one that reflects that which is being taught. Hence, those items with the highest DIs would be those selected for a CRT. By extension, they can also be said to be the items that best reflect what was being learned while the students were in our reading course.

Table 4: Calculating the Difference Index

Item	Posttest	–	Pretest	=	DI	Item	Posttest	–	Pretest	=	DI
#	IF		IF			#	IF		IF		
1.	.951	–	.918	=	.033	31.	.820	–	.705	=	.115
2.	.623	–	.656	=	-.033	32.	.607	–	.443	=	.164
3.	.885	–	.918	=	-.033	33.	.574	–	.557	=	.017
4.	.738	–	.770	=	-.032	34.	.803	–	.656	=	.147
5.	.541	–	.393	=	.148	35.	.885	–	.754	=	.131
6.	.852	–	.754	=	.098	36.	.410	–	.334	=	.066
7.	.885	–	.672	=	.213	37.	.754	–	.623	=	.131
8.	.607	–	.443	=	.164	38.	.738	–	.590	=	.148
9.	.443	–	.393	=	.050	39.	.787	–	.623	=	.164
10.	.262	–	.262	=	.000	40.	.574	–	.492	=	.082
11.	.951	–	.902	=	.049	41.	.770	–	.574	=	.196
12.	.820	–	.738	=	.082	42.	.623	–	.492	=	.131
13.	.656	–	.623	=	.033	43.	.836	–	.689	=	.147
14.	.820	–	.803	=	.017	44.	.787	–	.639	=	.148
15.	.639	–	.639	=	.000	45.	.738	–	.656	=	.082
16.	.705	–	.557	=	.148	46.	.328	–	.246	=	.082
17.	.869	–	.754	=	.115	47.	.869	–	.574	=	.295
18.	.590	–	.508	=	.082	48.	.689	–	.344	=	.345
19.	.492	–	.311	=	.181	49.	.623	–	.311	=	.312
20.	.607	–	.475	=	.132	50.	.557	–	.262	=	.295
21.	.852	–	.787	=	.065	51.	.820	–	.639	=	.181
22.	.705	–	.557	=	.148	52.	.262	–	.246	=	.016
23.	.689	–	.721	=	-.032	53.	.754	–	.623	=	.131
24.	.508	–	.328	=	.180	54.	.639	–	.508	=	.131
25.	.934	–	.885	=	.049	55.	.689	–	.541	=	.148
26.	.770	–	.557	=	.213	56.	.508	–	.426	=	.082
27.	.754	–	.672	=	.082	57.	.656	–	.492	=	.164
28.	.803	–	.738	=	.065	58.	.426	–	.361	=	.065
29.	.787	–	.525	=	.262	59.	.492	–	.311	=	.181
30.	.541	–	.410	=	.131	60.	.639	–	.443	=	.196

The actual decisions about which items to keep and which to discard were made on the basis of more information than just the DI. Recall that we wanted not only a test that would reflect what was being taught in the courses but also function well as a norm-referenced placement tool. In other words, we wanted an instrument that would help us make sound placement decisions, but insofar as possible, on the basis of things that we had to offer in terms of language learning—particularly reading.

As a result, the items for the revised version were selected on the basis of not only the difference indices but also the IF and ID statistics found in the placement administration. These are shown in Table 5. Notice that the items with an asterisk after them are the ones that were selected for the revised version of the reading test. Comparing those that were selected to those which were not, it should become clear that most of the items had an IF between .30 and .70, an ID near or in excess of .30 on the placement administration, and a DI higher than .10 in the pretest-posttest analysis to be selected for our revised version. From an NRT point of view, such an item would be reasonably well centered (IF between .30 and .70) and maximally separate the "high" proficiency ESL readers from the "low" proficiency ones (high ID). From a CRT point of view, such an item would be at least somewhat related to the learning that was going on in the reading course (high DI). Choices were also tempered by the fact that these items were based on passages which had to be treated as units. We ended up keeping all but one passage with fewer items per passage.

Once the items were selected, the results of the placement test were reanalyzed as though only the 35 remaining items had been administered. The item statistics that resulted from this reanalysis are reported in Table 6, while the overall descriptive statistics can be found in the second column of Table 3. This analysis gives a rough estimate of what will happen when we actually use this version.

Table 5: Selecting Norm-referenced Items Related to the Program

#	IF	ID	DI	#	IF	ID	DI
1.*	.951	.918	.033	31.*	.696	.347	.115
2.	.649	.303	-.033	32.*	.454	.222	.164
3.	.871	.304	-.033	33.	.582	.228	.017
4.	.747	.273	-.032	34.*	.727	.440	.147
5.*	.407	.357	.148	35.*	.789	.476	.131
6.	.799	.470	.098	36.	.392	.253	.066
7.*	.649	.355	.213	37.*	.686	.446	.131
8.*	.541	.302	.164	38.*	.644	.473	.148
9.	.500	.279	.050	39.*	.722	.489	.164
10.	.340	.268	.000	40.*	.552	.425	.082
11.	.897	.490	.049	41.*	.624	.569	.196
12.	.742	.443	.082	42.*	.521	.305	.131
13.	.577	.406	.033	43.*	.711	.385	.147
14.	.809	.387	.017	44.*	.696	.465	.148
15.	.629	.170	.000	45.	.660	.374	.082
16.*	.644	.344	.148	46.	.309	.236	.082
17.*	.763	.329	.115	47.*	.680	.543	.295
18.	.536	.305	.082	48.*	.552	.465	.345
19.*	.479	.348	.181	49.*	.443	.406	.312
20.*	.567	.308	.132	50.*	.490	.349	.295
21.	.845	.455	.065	51.*	.686	.401	.181
22.*	.593	.310	.148	52.	.278	.264	.016
23.	.711	.325	-.032	53.*	.732	.462	.131
24.	.423	.290	.180	54.*	.577	.411	.131
25.	.881	.425	.049	55.*	.665	.480	.148
26.*	.629	.437	.213	56.	.541	.544	.082
27.	.691	.351	.082	57.*	.536	.480	.164
28.	.722	.325	.065	58.	.407	.275	.065
29.*	.629	.329	.262	59.*	.381	.323	.181
30.*	.510	.299	.131	60.*	.531	.310	.196

* Items selected for the revised version.

Table 6: Revised Version

#	IF	ID	DI
1.	.951	.384	.033
5.	.407	.337	.148
7.	.649	.342	.213
8.	.541	.309	.164
16.	.644	.319	.148
17.	.763	.311	.115
19.	.479	.331	.181
20.	.567	.284	.132
22.	.593	.282	.148
26.	.629	.398	.213
29.	.629	.326	.262
30.	.510	.272	.131
31.	.696	.331	.115
32.	.454	.232	.164
34.	.727	.424	.147
35.	.789	.472	.131
37.	.686	.458	.131
38.	.644	.486	.148
39.	.722	.504	.164
40.	.552	.454	.082
41.	.624	.584	.196
42.	.521	.278	.131
43.	.711	.383	.147
44.	.696	.497	.148
47.	.680	.600	.295
48.	.552	.537	.345
49.	.443	.435	.312
50.	.490	.383	.295
51.	.686	.418	.181
53.	.732	.503	.131
54.	.577	.435	.131
55.	.665	.519	.148
57.	.536	.506	.164
59.	.381	.344	.181
60.	.531	.305	.196

DISCUSSION

In response to the first and third research questions, then, it appears that the revised version of the ELIPT reading test will function reasonably well as an NRT reading placement test in that it is well centered (mean), produces a wide spread of scores (SD) and is reasonably reliable—especially in view of its new shorter length.[3]

The fourth research question concerned validity. Validity is defined as the degree to which a test is measuring what it claims to measure. The original test could be defended from a content validity point of view in the sense that it was judged by "experts" to tap various reading comprehension skills. However, this is not a very strong or satisfying argument, especially since no other types of validity studies had been conducted on this particular test.

The present study has demonstrated another kind of validity. This is called construct validity, i.e., showing that a test is measuring what it purports to test through an experiment. One way that construct validity can be demonstrated is by showing that the test is assessing a particular construct through a pretest-posttest experimental design. In this study, the construct in question would be reading proficiency. The strategy here was to test the students before they have the construct, teach the construct, and test them again to see if the test taps what they have gained in learning the construct. Since this test is for placement into our course, it seems logical that the construct, reading proficiency, should be defined by those skills and knowledge areas taught in the course. Thus in a very fundamental way, we should be able to argue not only that the test has a high degree of content validity but also that these results demonstrate its construct validity in the form of the 21 percent gain shown in Table 3 for the pretest-posttest results on the original version.

It can also be argued that we have improved the construct validity of the test, as reflected by the larger gains which result when the revised 35 item version is analyzed for pretest-posttest differences. The pretest mean for those students taking our reading course was 18.90 and their posttest mean was 24.87—a gain of nearly six points, or approximately 32 percent. While this is

[3] In general, if all other factors are held constant, longer tests tend to be more reliable than short ones. See Brown 1984b for example and Ebel 1979 for more explanation.

still not a gain of staggering magnitude, we take it to indicate that our revised reading comprehension test is not only more efficient than the original version but also more valid for our purposes in reading placement.

There is always a possibility that observed differences like these are due to chance alone. However, the overall results reported in Table 3 indicate a difference in means that is not only statistically significant (i.e., we can be 99 percent sure that the observed difference in means is due to other than chance factors) but also large. While the practice effect (i.e., having already taken a test affects the results of a subsequent administration of that test) is one possible explanation for these gains, it is not likely to account for any large proportion of the gains because there was a 16 week interval between the administrations and the students had no warning, except in very general terms, about what their "final examination" would be like.

It is also important to remind the reader that these gains cannot be attributed solely to the reading instruction received in our ELI course. Students were also enrolled in other courses, in the ELI and elsewhere on campus. Hopefully, they also had a good deal of English language input in other nonacademic aspects of their lives. So it would be erroneous to claim that the gains were entirely due to our marvelous course.

Nevertheless, the outcomes here are encouraging from our point of view because they indicate that we have managed to revise the test such that it is more fully assessing reading comprehension skills that are somehow related to what we are teaching and what the students are learning while in our course. This is much more satisfying than the previous situation wherein we had no idea whatsoever how the test was related to the actual learning that was going on.

Because of the success of this project, implementation of the revised version of the Reading Comprehension Test is now planned for the Spring semester 1988. At that time, 25 additional items will be piloted with those that have been selected in this project. These 25 will be items that are written to be very similar to those that worked well from both NRT and CRT points of view. Thus another version of the test can be further refined and administered for use in Fall 1988.

In addition, based on this model, a lead teacher for each skill area has been given release time and primary responsibility for marshalling the available resources and personnel to generate tests for each skill. These separate, but related projects are now underway. Thus an ongoing process of placement test generation, analysis and revision has been set in motion. The difference here is not that we will systematically generate norm-referenced placement tests for the various skill areas, but that we will do so with items that function well as NRT placement items and assess content and skills related to what the students are learning in the ELI and at the University of Hawai'i.

Received February 15, 1988.

Author's address for correspondence:

James D. Brown
Department of English as a Second Language
University of Hawai'i at Manoa
1890 East-West Road
Honolulu, HI 96822

REFERENCES

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

Brown, J.D. (1981). Newly placed versus continuing students: comparing proficiency. In J.C. Fisher, M.A. Clarke & J. Schachter (Eds.) *On TESOL '80 building bridges: research and practice in teaching English as a second language.* Washington, D.C.: 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, D.C.: TESOL.

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

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

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

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

Guilford, J.P. & B. Fruchter (1973). *Fundamental statistics in psychology and education* (5th ed.). New York: McGraw-Hill.

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

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

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

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