Articles

Using a Commercially Produced Proficiency Test in a One-Year Core EFL Curriculum in Japan for Placement Purposes

Brent Culligan
Seigakuin University

Greta Gorsuch
Mejiro University

EFL program administrators have two general testing options for placement of students: commercially produced proficiency tests or locally developed tests. This study focuses on the use of a commercially produced proficiency test (the Secondary Level English Proficiency® test) for student placement in a core EFL program at a private junior college and university in Tokyo. The research was conducted to judge the degree to which the use of the SLEP® test was appropriate for student placement purposes. Pre- and post-test results for 538 students were analyzed for item facility, item discrimination, and item difference indices. It was found that the test did not appear to "fit" the students nor the program. The authors urge the adoption of supplemental placement procedures as well as the development of more program-sensitive tests.

学習者の能力に応じた英語のクラス分けを行う際には、一般に普及している英語能力テストを実施してその得点をもとに行う場合と、各英語コースが独自に開発したテストを実施してその得点をもとに行う場合とが考えられる。本論では、日本の私立短期大学・大学において Second Level English Proficiency Test (SLEP)を用いて英語のクラス分けを行ったケースについて論じる。

本研究の目的は、プレースメント目的でのSLEPの有用性を検証することである。538名に事前・事後テストを実施し、項目分析を行った結果、SLEP®本研究被験者・英語コースのいずれにも適合しないことが判明した。この結果をもとに、より妥当なプレースメントを行うためにテストに併せて実施すべき補足的な事柄と、当該英語コースにより適合したプレースメント・テストの開発について言及する。

JALT Journal, Vol. 21, No. 1, May, 1999

EFL program administrators have two general testing options for placement of students: commercially produced proficiency tests or locally developed tests. However, surprisingly little research has been published on the use of commercially produced proficiency tests for student placement in such programs and only a few researchers have published accounts of local placement test development in ESL programs for which the test has been written, piloted, and/or revised by on-site developers (Brown, 1989; Wall, Clapham & Alderson, 1994). This study will describe the use of one commercial test, the Secondary Level English Proficiency® for student placement in a core EFL program at a private junior college and university in Tokyo. The main focus of the research is to assess the degree to which the use of the SLEP® test is appropriate for placement purposes in the program. We seek to determine how appropriately it places students and how well the test "matches" the program goals and objectives. A second interest is to suggest methodology for other researchers to investigate the appropriateness of commercially produced proficiency tests used for student placement in their programs.

"Locally" Developed Placement Tests

"Local" placement tests, if developed along the lines of sound testing principles, have two important advantages. First, such placement tests can be piloted, analyzed, and then revised freely—the type and length of the test need only be limited by the skills of the local test development team and the teachers in the program. Second, such a test can be linked with the curriculum. This second advantage is strongly desirable. In Brown's words, "a placement test must be . . . specifically related to a given program, particularly in terms of the relatively narrow range of abilities assessed and the content of the curriculum" (1996, p. 12). This aspect of test validity is known as content validity. It is the notion that the test content should reflect the content of the curriculum or course it is being used in (Alderson, Clapham, & Wall, 1995; Bachman, 1990; Brown, 1990; Brown, 1995; Brown, 1996; Oller, 1979).

However, these advantages only hold if tests are developed using sound testing principles, including creating test item specifications and item banks, piloting the test, analyzing the test items and the statistical parameters of the test, and then revising the test to improve it on a continuous basis (Alderson et al., 1995; Brown, 1996; Davies, 1990; Henning, 1987). The local test developers would also have to estimate the reliability of the test, determining whether the test was measuring students' traits consistently (Alderson et al., 1995; Brown, 1996; Heywood,

1989; Hughes, 1989; Weir, 1993). Finally, the test developers would have to develop various arguments for the validity of the test. For example, placement decisions could be correlated with students' later achievement in their classes or with the appropriateness of the students' initial placement (Hughes, 1989; Wall et al., 1994).

Developing any sort of test is an arduous process requiring time and adequate knowledge of testing principles. Weir (1993, p. 19) notes that local test development requires group effort. However, having a group of informed and committed test developers in a program is sometimes not possible and administrators and/or teachers in ESL/EFL programs often elect to purchase commercially produced proficiency tests for placement purposes.

Commercially Produced Proficiency Tests

Using commercially produced proficiency tests in a language program has several advantages, the foremost being convenience. As many local test developers will attest, it may take months of committed, enlightened effort to produce a minimally reliable test (Griffee, 1995). Another advantage is economy. For a reasonable sum, programs can purchase testing packages such as the SLEP®. Such packages also include evidence supporting the reliability of the test (Gorsuch, 1995), since testing companies have the resources to make generally reliable tests and to offer well-organized information regarding the valid use of their tests.

An additional reason is ease of administration and scoring. In very large programs such as the one discussed in this study (748 students), it may be impossible to administer tests in which students are interviewed and rated or in which students' writing samples are rated. In such large programs, the number of students may necessitate the use of a paper-and-pencil test, which is the form taken by commercially produced proficiency tests. Finally, such tests may have high face validity in the eyes of students and administrators; commercially produced tests are characterized by professionally laid out and printed pages and high quality tape recordings. The SLEP® test offers an additional advantage. The makers of the test, ETS®, have developed a chart that test administrators can use to estimate students' TOEFL® scores based on their SLEP® scores. That can be valuable in programs in which administrators and/or teachers are anxious to "prove" the value of the program to other interested parties.

However, the literature regarding the use of various kinds of tests for student placement indicates that proficiency tests are a second choice,

and even then only in specific kinds of situations. For example, Bachman (1990) suggests the use of proficiency tests for placement when:

1. the students to be tested vary widely in terms of background and language ability;
2. the learning objectives of a program are not clearly specified; and
3. levels of students are known to vary widely from year to year, making the use of a locally developed test normed on one sample of students problematic.

Brown partially agrees: "If a particular program is designed with levels that include beginners as well as very advanced learners, a general proficiency test *might* (italics in the original) adequately serve as a placement instrument." Brown also cautions, "However, such a wide range of abilities is not common . . . in programs" (1996, p. 13).

Yet in most tertiary level EFL programs in Japan the students' second language learning experiences and abilities do not vary widely. Students in these programs have had six years of formal EFL education using similar textbooks and instructional practices. Furthermore, many colleges and universities in Japan are revising their EFL curricula, and have developed program-specific learning goals and objectives. Is the use of commercially produced proficiency tests for placement purposes appropriate for such schools?

As noted, administrators in ESL/EFL programs often choose to use commercially produced proficiency tests for student placement, yet this decision may be problematic. In Brown's words, "Each [placement] test must be examined in terms of how well it fits the abilities of the students and how well it matches what is actually taught in the classrooms" (1996, p. 13). Otherwise students may be placed in class levels based on a test that makes no comment on the curriculum in which the students are enrolled (Brown, 1990). The potential for inappropriate placement can become all too real in such a situation. (For additional cautions concerning the use of proficiency tests for placement, see Brown, 1995; Henning, 1987; and Hughes, 1989.)

Program administrators thus have the difficult choice of using a commercially produced proficiency test which may not be appropriate for placement of their students or they can expend a massive amount of effort writing their own tests. In the end, however, locally written tests may be no more appropriate or reliable than a commercially produced proficiency test. Another option may be to use a commercially produced proficiency test as a stepping stone towards developing a locally written placement test, as will be described below.

Research Focus

This study estimates the extent to which the SLEP® proficiency test is suitable as a placement test for a core English program at a Japanese university. We will address three questions. First, how well does the SLEP® test "fit" the students in the program? Second, how well does the SLEP® test "fit" the goals and objectives of the program? And third, what steps can be taken to improve placement decisions in the program? In answering these questions, we will outline the minimal steps that should be taken to determine the validity of such tests for student placement in tertiary level EFL programs, if reliable and valid "local" tests cannot be developed.

Research Questions

1. What items on the SLEP® test discriminate effectively between high and low scoring students?
2. Will selective scoring of the SLEP® test produce more effective placement of students?
3. To what extent will items from the first and second test administration with high difference index values match the stated goals, objectives, and syllabus of the program[1]

Method

Subjects

The majority of the 748 first-year students enrolled in the university and junior college divisions of the English program during the year of the study were recent graduates from Japanese high schools and were approximately 18 years of age. The students were predominantly of Japanese nationality, with the exception of three Korean students and one Chinese student in the university division. There were 310 males and 87 females in the university division of the program, while the 380 students in the junior college division were all female. In addition, there were seven second-year students in the program who were repeating their first-year English requirements.

The university students were drawn from three majors: Political Science and Economics (268 first-year students), American and European Culture (65), and Early Childhood Education (64). Students in the junior college division majored either in English Literature (180 first-year students and three second-year students) or Japanese Literature (200 first-year and four second-year students).

Material

Two sets of materials were used in this study: the SLEP® test and the core English program goals, objectives, and syllabuses (see Appendix).

SLEP®

The SLEP® test was developed by the Educational Testing Service (ETS®) in 1980, using over 6,000 non-native English speaking secondary school students in the US and in "foreign countries" as its norming population (ETS®, 1991, p. 8). In the words of ETS®, it is a proficiency test and "a measure of ability in two primary areas: understanding spoken English and understanding written English" (ETS®, 1991, p. 7). Further, it is "helpful in evaluating ESL teaching programs and making placement decisions" (ETS®, 1991, p. 7). It is not an aptitude or achievement test.

The SLEP® test currently has three equivalent forms. Students taking the test have a test book and an answer sheet for marking answers. The reported reliability coefficient of the SLEP® is .94 for the listening subtest, .93 for the reading subtest, and .96 for the entire test (ETS®, 1991, p. 9). The SLEP® test is designed to be locally scored, either using a two-ply pressure-sensitive answer form, or an optical recognition form. Scoring here was done using the optical recognition forms and a scoring machine.

The test is made up of a listening section and a reading section, each with 75 multiple choice items. The listening section has four subsections, made up of four different types of multiple choice items. In Form 1, the first listening subsection ("1Pic") asks the students to look at a photograph in the test book and then listen to four sentences on a tape. On their answer sheet the students mark the sentence best describing the photograph. There are 25 items in the "1Pic" subsection. The second listening subsection ("Dict") asks the students to read four sentences in the test book and listen to a sentence recorded on the tape. The students mark the sentence in the test book that is the same as the one on the tape. There are 20 items in the "Dict" subsection.

The third listening subsection ("Map") has 12 items based on an illustration representing a bird's-eye view of a small town. The students identify the buildings and streets on the map and the locations of four cars on the streets. The students then hear short conversations between various adult North Americans on the tape and must surmise in which car the conversation is taking place. The "Map" subsection assumes the cars in the illustration are driven on the right hand side of the road.

The fourth listening subsection ("Conv") has 18 items regarding a North American high school. The students hear several short conversations between adult and teen-age North Americans on the tape. After

Table 1: Summary of Sections and Subsections of SLEP® (Form 1)

Listening Section Subsections	Number of Items	Time Allowed
1Pic	25	
Dict	20	
Map	12	
Conv	18	
		45 minutes
Reading Section Subsections	**Number of Items**	**Time Allowed**
Cart	12	
4Pics	15	
Cloze	22	
RP1	18	
RP2	8	
		45 minutes

each conversation, the students hear one or two questions about the conversation and select the correct answer from written items in the test book. The entire listening test with the four subsections takes approximately 45 minutes to complete.

The reading section, which ETS® claims tests grammar and vocabulary, also contains four subsections with four types of multiple choice items. The first reading subsection ("Cart") presents a cartoon illustration in which several people have "thought bubbles" above their heads, each illustrating a different point of view of a particular event. For each item, students read two or three sentences and then match the item to the "thought bubble" of one of the people in the illustration. There are 12 items of this type. The second reading subsection ("4Pics") asks the students to read a sentence, then match it to one of four illustrations which best describe it. There are 15 items of this type.

The third subsection is a short modified cloze reading passage ("Cloze"). For each missing word the students choose one of four possible answers. There are 22 items. The fourth reading subsection ("RP1") contains questions about the preceding passage; the students choose the best answer to the question from four choices. There are 18 items. There are three such modified cloze passages with three sets of questions. Finally, the fifth reading subsection ("RP2") presents a reading passage (without cloze) and eight multiple choice questions about it (eight items).

The students are given 45 minutes to complete the reading test.

See Table 1 for a summary of the tests and subsections of Form 1 of the SLEP® test.

Program Curriculum

In early 1993 two special committees at the university were formed to revise the EFL curriculum. The goal was the creation of a multi-level core EFL program for all first-year university and junior college students, to be implemented at the start of the 1996 academic year. The curriculum design process included administration of a Japanese-language needs analysis questionnaire to 2,067 lower and upper class students at the school in early 1995, numerous in-service lectures conducted by faculty and non-faculty expert/informants over a three year period, readings from the *ACTFL Proficiency Guidelines* (Buck, 1989), and individual study and reflection on the part of the committees' members.

During the period of this study, the program had three levels: A level (high), B level (intermediate) and C level (remedial), corresponding to intermediate/high, intermediate/mid, and intermediate/low levels on the speaking portion of the *ACTFL Proficiency Guidelines* (Buck, 1989). First-year students in the university division attended two 90-minute classes per week for 26 weeks in the core English program, amounting to 78 hours of instruction in one academic year. English Literature majors in the junior college division also received 78 hours of instruction in one academic year, while Japanese Literature majors received 39 hours of instruction given only in the first semester.

Within each level, general goals concerning English proficiency and vocabulary were set, as were objectives describing more precise learn-

Table 2: Recommended Textbooks

Level A
Atlas II (Nunan, 1996)
Level B
Atlas I (Nunan, 1996)
Interchange I (Richards, Hull & Proctor, 1990)
New Person to Person Book 2 (Richards, Bycina & Kisslinger, 1996b)
Level C
New Person to Person Book 1 (Richards et al., 1996a)
First Impact (Ellis, Helgesen, Browne, Gorsuch & Schwab, 1996)

ing outcomes (see Appendix). These goals and objectives resulted in a series of notional/functional syllabuses stressing a communicative approach to language learning. Although objectives for developing students' communicative reading and writing skills were articulated, the program was mainly designed to promote oral/aural skills development.

Based on the program objectives, a selection of textbooks was made for teachers to choose from for use in their classes. (See Table 2.)

In line with goals concerning vocabulary development, a number of learning objectives were specified (see Appendix). After considering materials such as the *Longman Language Activator* (1994), *A General Service List of English Words* (West, 1953) and *A University Word List* (Nation, 1990), a "master vocabulary list" of 3,000 words was compiled using the *Cambridge English Lexicon* (Hindmarsh, 1990), the *Longman Dictionary of Contemporary English* (1995), and the *Cambridge International Dictionary of English* (1995). Vocabulary was broadly sequenced according to frequency to correspond to Levels A, B, and C.

Twenty-five words per week were integrated into the syllabus. Program teachers created weekly vocabulary worksheets based on the 25 words, including crossword puzzles, definition matching, and cloze exercises. The teachers collected the worksheets periodically for correction and comment as formative assessment. Lead teachers assigned to the levels wrote vocabulary quizzes which were given every three weeks to test the students' progress. The vocabulary quizzes contained 25 items taken from the 75 words the students had been studying for the previous three weeks.

Procedure

At the beginning of the 1996 academic year 748 junior college and university students in the program took the SLEP® test Form 1, both listening and reading, for placement purposes. This administration will be referred to as the "pre-test." Nine months later, in January, 1997, 487 students were administered the same Form 1 test for purposes of program evaluation. This is termed the "post-test." The 210 students in the Japanese Literature program did not take the post-test at the same time as the other students because of different degree requirements. Therefore, their scores were not included in this study, nor were those of the 51 university students who failed to take the post-test. Thus, pre-test and post-test scores of only 487 students were used in the analysis.

Data Analyses

To determine which test items discriminated effectively between high and low scoring students (the first research question), the pre-test scores for 487 students on all items of the SLEP® test were entered into a

spreadsheet program and were subjected to an item discrimination analysis (ID), a norm-referenced item statistic. According to Brown (1996, p. 66), ID analysis of test items "indicates the degree to which an item separates the students who performed well from those who performed poorly." The ID was calculated for each test item by subtracting the item facility (IF_{lower}) of the students scoring in the lowest third of the test overall from the item facility (IF_{upper}) of the students scoring in the highest third of the test overall. Item facility (IF) is the proportion of students who answered a particular item correctly. For example, if six out of ten students correctly answered an item, the IF would be .60.

Generally speaking, test administrators expect students who score highly on the test overall to also score highly on individual test items. Conversely, administrators expect students with low scores on the test overall to score poorly on most of the individual items. However, the opposite may happen; students who score highly overall may do poorly on individual items. Such items may be poorly constructed, ambiguously worded, or simply too difficult for the students. It is those items that are thought not to discriminate effectively between high and low scoring students and are thus likely to have low item discrimination (ID) values. According to Ebel (as cited in Brown, 1996, p. 70), test items with ID values of .40 and above are considered "very good" items, those with ID values of .30 to .39 are thought to be "reasonably good," and those with ID values of .20 to .29 are "marginal" items, usually "needing improvement." For this study, we looked for items with ID values of .20 and over.

To address the second research question, the high ID items were identified and were taken out of the rest of the data, creating a "high ID" data set. Thus two data sets were analyzed, the original data set with all the items included, and the "high ID" data set,in order to calculate the means, standard deviations, reliability estimates, and standard errors of measurement. This was done to see which data set yielded the more reliable information for placing students appropriately.

To answer the third research question, pre-test scores on individual test items for 487 students were compared to their matching post-test scores using a criterion-referenced test statistic, the difference index (DI) (Brown, 1996, p. 80). DI was calculated by subtracting pre-test item facility (IF) for each item from post-test IF for each matching item. Thus, if students did better on particular items on the post-test, the DI for those items had a positive value. Items with DI values of .10 or over were examined in light of the stated goals, objectives, and syllabuses of the program. In particular, we looked for any patterns in students' improvement in terms of SLEP® tests (listening and reading) and subtests ("1Pic," "Dict," "Map," etc.). We wanted to see the extent to which the SLEP® test "matched" the

program goals, objectives, and syllabus statements.

We would like to note here that although we used the goals, objectives, and syllabuses of the program to gauge the degree of fit between the program curriculum and the SLEP®, the implementation of the goals and objectives was not investigated. This issue is central to the whole question of defining what a curriculum is and what it does (i.e., program evaluation) (Holliday, 1992; Snyder, Bolin & Zumwalt, 1992; White, 1988). Our study, we feel, constitutes only one part of such a program evaluation. However, in Brown's (1995) model of curriculum development the establishment of objectives is followed by testing, and is then subject to evaluation. This first step is the limited scope of our study.

Results

Upon analysis of the pre-test data, we found that less than half of the items had an ID of .20 or higher, the minimum level thought acceptable for effective discrimination (Ebel cited in Brown, 1996). See Table 3 below.

Table 3: Pretest Items with ID of .20 and Above

Section	Subsection	Items with ID of .20 and Above	Total Items in Subsection
Listening	1Pic	16	25
Listening	Dict	20	20
Listening	Map	5	12
Listening	Conv	1	18
Reading	Cart	10	12
Reading	4Pics	6	15
Reading	cloze	4	22
Reading	RP1	2	18
Reading	RP2	2	8
Totals		66	150

The first research question asked which items on the SLEP® test discriminated effectively between high and low scoring students. Of the 66 items with "acceptable" IDs, 42 were listening section items and 24 were reading section items. The test thus appears to have discriminated better for listening than for reading. The remaining 84 items had an ID of .19 or below and, by Ebel's standards (as cited in Brown, 1996), were not useful for discriminating between high and low scoring students.

In answering the second research question, two data sets were created to see whether selective scoring of the SLEP® test would result in more effective placement of students. The "original data set" included data for all 150 items in the SLEP® test, whereas the "high ID data set" included data for only those 66 items that were found to have an ID of .20 or over (see Table 3 above). Comparisons of descriptive statistics on the two data sets are given in Table 4. Also included are KR-20 internal consistency estimates for the two data sets.

Table 4: Comparisons of Original Data Set and High ID Data Set

	Original Data Set	High ID Data Set
K	150	66
M	69.36	39.60
SD	12.38	9.05
high	107	61
low	32	11
range	76	51
KR-20	0.81	0.84
SEM	5.46	3.62

The standard error of measure (SEM) of the high ID data set is substantially lower than that of the original data set, whereas the KR-20 internal consistency estimate is somewhat higher for the high ID data set. These results indicate that selective scoring of the SLEP® test would most likely result in more effective placement of students in the program.[2]

Finally, to answer the third research question, regarding whether items from the first and second test administration with high difference index values match the goals and objectives of the program, pre-test and post-test data were compared to calculate the difference index (DI) for each item, thus estimating students' gain scores on particular items. Items with a DI of .10 or better by SLEP® test subsection are shown in Table 5.

Thirty-one of the "high DI" items were in the listening section and 16 were in the reading section. Four subsections had six or more items with high DIs, four subsections had items with low DIs, and one subsection had items with DIs of zero. Each of the subsections will be analyzed below and compared to the goals, objectives, and syllabuses of the core English program in order to understand the extent to which the items in the subsections "fit" the curriculum.

Table 5: Items with DI of .10 and Above

Section	Subsection	Number of High DI Items	Total Items in Subsection
Listening	1Pic	13	25
Listening	Dict	15	20
Listening	Map	2	12
Listening	Conv	1	18
Section Total		31	75
Reading	Cart	2	12
Reading	4Pics	0	15
Reading	Cloze	6	22
Reading	RP1	6	18
Reading	RP2	2	8
Section Total		16	75
Total		47	150

As shown in Table 5, students showed gain scores on 13 out of 25 items in the "1Pic" subsection, which focuses primarily on meaning; students see a picture, hear four statements, and then decide which statement matches the picture. While the goals and objectives for the core English curriculum cannot be explicitly matched with the subsection in terms of content, the goals and objectives statements for Programs A, B, and C (see Appendix) calls for students to learn how to "ask and answer questions" in a variety of settings. The goals and objectives statement for Program A mentions that students should learn to "understand and respond to extended discourse." If teachers created classroom activities based on these goals and objectives, perhaps these activities gave the students meaning-focused listening practice, either through pair work, completing listening activities in textbooks, or listening to extended lectures in English.

On the "Dict" listening subsection of the test, students showed high gain scores on 15 out of 20 items (see Table 5). Items in this subsection were more oriented to form than meaning. Students had to listen to a statement and match it with one of four written statements in the textbook. The connection between items of this type and the core curriculum is more tenuous and indirect. Only the Program A goals and objectives statements concerning the improvement of students' note-taking ability can be directly related to this subsection. Note-taking practice requires accuracy in listening. In addition, all the textbooks listed in Table 2

utilize tape-recorded listening activities which focus on accuracy in listening. We speculate that activities designed to meet the meaning-focused goals and objectives for listening had a "spill over" effect which improved students' accuracy in hearing and identifying English forms. Another possibility is that activities designed to fulfill the goals and objectives related to improving students' reading helped students to improve their scores in this listening subsection. Such test items require more reading skill than would at first seem apparent. In order to answer the items, students must "race ahead" of the tape and read the four answer statements quickly and accurately before the test statement is played on the tape. After the statement is played, the students must quickly read the answers again to evaluate which one is being said. It may be that students' reading practice in the core English program helped them read the answer choices on this subsection of the test more efficiently.

On the "cloze" reading items in the test (see Table 5), students showed gain scores on only 6 out of 22 items. While some of the cloze items tested vocabulary, many of them seemed to test the students' judgments of correct word morphology. Students were given four versions of the same verb or adjective and had to choose the most appropriate one. Of these six items, two indicated an increase in vocabulary knowledge, two showed gains in students' morphological discrimination, and two showed an increase in students' ability to choose correct function words, such as referents. The students' relative improvement on the six items may be partly due to the program's weekly vocabulary worksheets mentioned above. The vocabulary worksheets took a variety of forms, including cloze exercises and definition matching games, but presented the vocabulary items in the morphological form required for the correct answer. We speculate that students received input that promoted an inductive understanding of correct word morphology and syntactic structure on the relevant items in the SLEP® test.

The students showed an improvement on 6 out of 18 items (see Table 5) on the "RP1" subsection, and this seemed to have an indirect relationship to the goals and objectives of the program. The items in this subsection required the students to infer meaning. It is possible that through meaning-focused listening and reading activities, designed and used in accordance with the goals and objectives of the program (i.e., "understanding extended discourse," "reading written materials for information," "carrying on simple face to face conversations"), the students' ability to answer meaning-focused test questions improved.

As shown in Table 5, students showed little or no gain on five subsections: "Map," "Conv," "Cart," "4Pics," and "RP2." There are several explanations for this. Students already had fairly high scores on the "Cart" and

"4Pics" subsections on the pre-test. Thus, there was not much room for improvement. The "Cart" subsection pre-test item facilities (IFs) for 10 out of 12 items were .60 or over. In the "4Pics" section, 10 out of 15 items had pre-test IFs of .60 or over. These high values suggest that the items in the two subsections were generally easy for the students.

The small gains shown by students in the "Map" and "Conv" subsections probably have different causes. The students' pre-test IFs for most of the items in these subsections were low and remained so in the post-test. We feel that the two subsections were simply too difficult for these students because they were culturally inappropriate. Both the "Map" and "Conv" subsections assumed experiences that first-year Japanese college students are unlikely to have had. For example, the "Map" subsection assumed that the testees had done extensive car travel, or could drive, particularly on the right side of the road. However, most young Japanese do not get driver's licenses until they are 20 years old and then drive on the left hand side of the road.

Similarly, the "Conv" section assumes students are familiar with the duties of administrative personnel in American high schools. However, there is no guarantee that administrative counterparts in Japan handled the same duties, or even that there are such administrators in Japanese high schools. We feel that regardless of the language learning support students received in the program, the "Map" and "Conv" subsections presented unfamiliar concepts. Thus, students could not effectively demonstrate their learning through these two subsections.

The modest gains shown on the final subsection, "RP2" may have been due to students' unfamiliarity with the genre of fictional short reading. Many students are familiar with expository written English since this makes up the bulk of the reading presented in high school textbooks. However, they may be less familiar with stylistic devices and imagery used in fiction. The goals and objectives statements for program levels A, B, and C (see the Appendix) allude to reading in functional terms. In level A for example, students are asked to read easy "academic" materials. Students in levels B and C are asked to read "public transport schedules," "newspaper articles," and "notes from the teacher." The program is not intended to promote students' reading of literary works in English. Thus, this particular subsection is not really connected to the program, either in content or in terms of what activities students are asked to do.

Discussion

According to Bachman (1990, p. 238), test validity is not an abstract notion. Rather, test validity must be considered in the context of the infer-

ences that teachers or program administrators plan to make from the students' test results. Thus, in a situation where a commercially produced proficiency test is used to place students in different levels in a program, we need to answer the question of whether the test is valid for this purpose, i.e., whether the test "fits" the students and "fits" the program.

There are a number of reasons why the SLEP® test does not appear to be valid when used for placement of students in the core EFL program described in this study. First, we found that only 66 out of a total of 150 items on the test discriminated between high and low scoring students. The result was a standard error of measure of 5.46 (see Table 4), indicating a good deal of "looseness" around the cutoff points used to decide whether students should be placed in the A, B, or C levels of the program.

Second, the SLEP® test does not estimate oral ability, although an aim of the program is to increase students' oral skills. This alone constitutes a mismatch between the test and the program. We were able to make only indirect comparisons between the program's listening and reading goals and objectives and various SLEP® subsections, but these comparisons were at best speculative. The SLEP® test, therefore, does not seem to "fit" this particular program.

However, as discussed, administrators and/or teachers often elect to use commercially produced proficiency tests for placement in a program with defined goals and objectives. In our particular situation, the large number of students (748) made oral testing for placement purposes prohibitively difficult. Also, as this was the first year the core EFL program was in place, there was no possibility of developing a local paper-and-pencil test more suited to the students and to the program. We strongly hope that as the program continues the administrators and teachers will consider developing a reliable and valid local test or will develop placement procedures to supplement the SLEP® test. The data that we have gathered through this study can be of some assistance. For example, item types from the SLEP® test that consistently produce high gains and/or high discrimination can be used as models for item writing for the local placement test.

We suggest that the SLEP® test, if scored with all 150 items, is problematic for placement of the students in the program described above. We therefore recommend that the test be scored selectively, using only the 66 high ID items. By selectively scoring the SLEP® test, the program administrators may be able to obtain more effective placement of students by reducing error variance. Although the number of test items counted toward the total score would be reduced, the reliability of that score would increase. By scoring only the 66 items with high IDs, the SEM dropped from 5.46 to 3.62. The SEM is best conceived as "a band

around a student's score within which that student's score would probably fall if the test were administered to him or her repeatedly" (Brown, 1996, p. 206). We interpret this to mean that on the total test, the true score of a student who got a raw score of 70 could actually range from plus one SEM to minus one SEM 68% of the time, from 65 to 75. For the remaining 32%, the measurement error could be greater. This can result in the misplacement of "borderline" students. Reducing the SEM by selectively scoring the pre-test would reduce misplacement.

Continual assessment of the test items, such as we did in this study, will provide much needed "tuning" for educational institutions using proficiency tests, whether locally developed or commercially produced. With this in mind, we must assert that the results of this study cannot be used as justification for using portions of the SLEP® test in any other Japanese institutional setting. Only with continual monitoring of the results on an item-by-item basis can valid inferences be made using the SLEP®, or any other test, for a particular setting. As testing situations change, so must the assessment of the validity of the tests used.

Acknowledgments

The authors would like to thank J.D. Brown for his instruction and encouragement, and Dale T. Griffee, William Kroehler, and the three anonymous JALT Journal *reviewers for their insightful comments. Thanks are also due to the administrators, teachers, and students in the core English program at Seigakuin University.*

Brent Culligan is a full time instructor at Seigakuin University. A doctoral candidate at Temple University, Tokyo, he is interested in second language vocabulary acquisition and testing.

Greta Gorsuch teaches full time at Mejiro University, and is former editor of *The Language Teacher*. She is a doctoral candidate at Temple University, Tokyo, and is interested in testing, instruction processes, and education policy evaluation.

Notes

1. One of the reviewers objected to our use of this research question. She/he felt quite rightly that a multiple choice listening and reading test (such as the SLEP®) could not be considered appropriate for use in a program designed to promote students' oral/aural skills. However, we felt we needed to retain this research question. As stated earlier, one of our purposes is to suggest a method for readers to judge commercially-produced proficiency tests used for placement in their own programs. We feel that research question three presents a useful tool for relating the test to the program.
2. One reviewer suggested that in order to confirm our claim we would have to assess the students' progress over a semester to gauge the appropriateness of their placement using the high ID data set. While we feel this is a

cogent point, we also feel that in practical terms this would be difficult to carry out. Such an assessment would require comparing a control group (students placed using the original data set) to an experimental group (students placed using the high ID data set). Even if this or a time series study had been done, we would have to consider that these students' progress could be due to a multitude of factors and could not necessarily be attributed to appropriateness of student placement.

References

Alderson, J.C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

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

Brown, J.D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly,* 23 (1), 65-83.

Brown, J.D. (1990). Where do tests fit into language programs? *JALT Journal,* 12 (1), 121-140.

Brown, J.D. (1995). *The elements of language curriculum: A systematic approach to program development*. Boston: Heinle & Heinle.

Brown, J.D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Buck, K. (Ed.). (1989). *The ACTFL oral proficiency interview: Tester training manual*. New York: The American Council on the Teaching of Foreign Languages.

Cambridge international dictionary of English. (1995). Cambridge: Cambridge University Press.

Davies, A. (1990). *Principles of language testing*. Oxford: Blackwell.

Educational Testing Service®. (1991). *SLEP® test manual*. Princeton, NJ: Author.

Ellis, R., Helgesen, M., Browne, C., Gorsuch, G. & Schwab, J. (1996). *First impact*. Hong Kong: Longman Asia ELT.

Gorsuch, G.J. (1995). Tests, testing companies, educators, and students. *The Language Teacher,* 19 (10), 37, 39, 41.

Griffee, D.T. (1995). Criterion-referenced test construction and evaluation. In J.D. Brown & S.O. Yamashita (Eds.). *JALT applied materials: Language testing in Japan* (pp. 20-28). Tokyo: Japan Association for Language Teaching.

Henning, G. (1987). *A guide to language testing*. Boston: Heinle & Heinle.

Heywood, J. (1989). *Assessment in higher education* (2nd ed.). Chichester, MA: John Wiley and Sons.

Hindmarsh, R. (1990). *Cambridge English lexicon*. Cambridge: Cambridge University Press.

Holliday, A. (1992). Tissue rejection and informal orders in ELT projects: Collecting the right information. *Applied Linguistics,* 13, 403-424.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Longman dictionary of contemporary English (3rd. ed.). (1995). London: Longman.

Longman language activator. (1994). Harlow, Essex: Longman House.
Nation, I.S.P. (1990). A university word list. In I.S.P. Nation (Ed.). *Teaching and learning vocabulary* (pp. 235-239). New York: Newbury House.
Nunan, D. (1996). *Atlas I.* Boston: Heinle & Heinle.
Nunan, D. (1996). *Atlas II.* Boston: Heinle & Heinle.
Oller, J.W., Jr. (1979). *Language tests at school.* London: Longman.
Richards, J., Bycina, D. & Kisslinger, E. (1996a). *New person to person, book 1.* Oxford: Oxford University Press.
Richards, J., Bycina, D. & Kisslinger, E. (1996b). *New person to person, book 2.* Oxford: Oxford University Press.
Richards, J., Hull, J. & Proctor, S. (1990) *Interchange I.* Cambridge: Cambridge University Press.
Snyder, J., Bolin, F. & Zumwalt, K. (1992). Curriculum implementation. In P. W. Jackson (Ed.). *Handbook of curriculum research* (pp. 402-435). New York: MacMillan.
Wall, D., Clapham, C. & Alderson, J.C. (1994). Evaluating a placement test. *Language Testing,* 11(3), 321-344.
Weir, C. (1993). *Understanding and developing language tests.* New York: Prentice Hall.
West, M. (1953). *A general service list of English words.* London: Longman, Green and Co.
White, R.V. (1988). *The ELT curriculum: Design, innovation and management.* Oxford: Blackwell.

(Received January 11, 1998; revised June 13, 1998)

Appendix

Goals and Objectives for Levels A, B, and C

Goals and Objectives for Program A (intermediate-high)

Course Overview: The purpose of this course is to prepare students to understand and to respond to extended discourse such as lectures, TV and radio talks, to make simple presentations, and to narrate in the past.

Goals	Objectives
1. Increase mastery of vocabulary and idioms in order to expand the range of situations in which students can function in English, and in order to gain competency in academic pursuits.	Be able to score at least 80% on a vocabulary test on approximately 3500+ words including the *University Vocabulary* and other high frequency vocabulary items. Be able to score at least 80% on a test of 700 high frequency idioms (including the 500 in Program B).
2. Understand extended discourse.	Listen to and understand simple lectures and speeches in general and academic settings.
3. Ask questions regarding extended discourse; narrate in the past.	Be able to ask pertinent questions regarding lectures and speeches; be able to make presentations such as a report in a seminar; be able to narrate events and experiences in the past.
4. Read written materials of increasing difficulty for gathering information for personal and academic purposes.	Be able to understand simple academic writing and an increasing number of newspaper and magazine articles.
5. Note-taking and academic writing.	Take notes on lectures, write simple reports based on reading materials, taking into consideration citation and bibliographical protocols.

Goals and Objectives for Program B (intermediate-mid)

Course Overview: The purpose of this course is to prepare students to participate in simple conversations about their personal history, leisure time activities, etc., to recognize different registers (politeness, etc.), to listen to simple announcements and use the telephone, to read descriptions of persons, places and events, and to write simple letters or compositions on assigned themes.

Note: Goals and Objectives for Program C are assumed, and if necessary some review of goals and objectives for Program C will be included in Program B.

Goals	Objectives
1. Increase mastery of essential vocabulary and idioms to increase overall mastery of English, and in order to be able to effectively use an English/ English dictionary designed for ESL learners.	Be able to score at least 80% on a vocabulary test on 2,500+ word level expanded from the vocabulary list in Program C from such lists as the Key Concepts in the *Longman's Activator Dictionary*; be able to score at least 80% on 500 high frequency idioms (including the 300 in Program C).
2. Be able to ask and answer questions and carry on face-to-face conversations when traveling overseas and in a setting such as a homestay in an English-speaking family.	Ask and give information about travel plans; offer, accept and refuse invitations; explain aspects of one's culture; describe health problems, etc.
3. Be able to read a widening range of written materials for essential information and for enjoyment.	Be able to understand and read public transport schedules, notices and advertisements, and simple newspaper and magazine articles.
4. Be able to convey increasingly complex ideas and information through written English.	Write letters and expanded compositions about daily activities and social activities; write more detailed book reports.

Goals and Objectives for Program C (intermediate-low)

Course Overview: The purpose of this course is to prepare students to be able to introduce themselves, ask and answer simple questions and successfully handle a limited number of interactive, task-oriented and social situations, and to convey and gather basic information through writing.

Goals	Objectives
1. Increase mastery of essential vocabulary and idioms in order to increase overall English ability, and in order to be able to begin using an English/English dictionary designed for ESL learners.	Be able to score at least 80% on a vocabulary test on the 2,000+ word level developed in-house from *West's General Service List, Longman Defining* vocabulary; be able to score at least 80% on 300 high frequency idioms.
2. Be able to ask and answer questions, and carry on simple face-to-face conversations such as self-introductions, ordering a meal, asking directions, making purchases.	Participate in role plays, greet and carry on minimal conversations with native speakers on campus, understand and respond to classroom instructions in appropriate ways.
3. Be able to gather basic information from simple written English instructions.	Become familiar with written English instructions in order to take tests without resorting to the use of Japanese. Be able to read class notices and notes from the teacher. Read simplified graded readers.
4. Be able to convey simple messages through written English.	Write simple answers to questions. Write simple short passages such as self-introductions, everyday activities, plans.