The *Eiken* Test: An Investigation

Laura MacGregor
Sapporo International University

The *Eiken* tests, first administered in 1963 by the Society for Testing English Proficiency (STEP), are highly respected in social, educational, and employment circles and taken by millions each year. However, upon closer scrutiny, it appears that the *Eigo Kentei Kyokai* (*Eikyo*) operates on its own terms. Unlike TOEFL and TOEIC, *Eikyo* does not make information about the tests' reliability or validity available to the public. Therefore, important questions remain unanswered: Are the *Eiken* tests reliable and valid instrument? Do the *Eiken* tests really function as tests of English proficiency? This paper examines the *Eiken* pre-second level test from June 1994. The test was administered to 168 first-year Japanese college students. The results provided data for reliability and validity studies, in an effort to shed light on the value of the *Eiken* pre-second level test as a reputable test instrument of English proficiency. The results of the studies conducted here are far from encouraging with regards to the *Eiken* pre-second level test's reliability and validity.

英検は1963年から英語検定協会によって実施されており、社会的、教育的に、また企業によっても高く評価され、毎年数千人の受験者がいる。しかしながら詳しく調査してみると、英検は独自の方式で実施されており、TOEFLやTOEICと違って信頼性や妥当性についての情報が公開されていない。従って、以下のような重要な疑問は答えのないままである。英検は信頼性と妥当性のあるテストであるのか。英検は本当に英語能力のテストとして機能しているのか。本研究は、英語能力を測定するテストとしての英検の価値を知るために、1994年6月に実施された英検2級のテストを168人の日本人大学一年生に実施し、その信頼性と妥当性を調査した。その結果、英検の信頼性と妥当性ははなはだ心細いものであることがわかった。

Japan is a country whose people thrive on tests, from *kendo* to calligraphy, flower arranging to gift wrapping; tests which evaluate almost every skill imaginable are available for the taking, so to speak. Walk into any bookstore or culture center and you'll see an array of posters and pamphlets advertising such tests. In the academic world, tests are in abundance as well. Entrance examinations which determine students' future high school and post-secondary careers are a fact of life for virtually every family.

By far, the oldest and best established English language tests in Japan are the *Eiken* tests (*Eigo Kentei*), produced by *Nihon Eigo Kentei Kyokai* (*Eikyo*), or in English, STEP: the Society for Testing English Proficiency. Since it began offering the *Eiken* tests in 1963, *Eikyo* has enjoyed a long period of unprecedented success: The Ministry of Education endorses the *Eiken* tests,[1] and recommends students take them. Some schools even offer courses dedicated to *Eiken* test preparation. In the working world, it has made its mark as well. Many employers regard *Eiken* test qualification as a valuable asset and look for it on prospective employees' resumes (MacGregor, 1995). However, despite the test's wide acceptance and use, one important question has been overlooked: Are the Eiken tests reliable and valid instruments to measure English proficiency? Eikyo has failed to give a direct answer. Other language tests, such as TOEFL and TOEIC, publish regular reports with statistical analyses of reliability and validity (TOEFL, 1995; TOEIC, 1995; Woodford, 1980, 1992). Why doesn't *Eikyo*? This question has been raised by other language educators concerned about Eikyo's position as well (Bostwick, 1995; Brown, 1995; Gorsuch, 1995).

This paper seeks answers to the above questions of reliability and validity by conducting analyses on a set of data collected by the author. Because certain reliability analyses are best suited to certain types of tests, it is necessary to begin with the question, is the *Eiken* a criterion-referenced test (CRT) or a norm-referenced test (NRT)? To help answer this question, various aspects of the *Eiken* test will be compared with those of TOEFL and TOEIC. This discussion is followed by an investigation of the *Eiken* pre-second level's[2] reliability and validity, which focuses on four questions: 1) Is the *Eiken* test appropriate for the group *Eikyo* claims to evaluate? 2) Do the test items reflect the practical English found in daily life that *Eikyo* claims to test? 3) Does the test measure the abilities that *Eikyo* claims it does? and 4) Are there any poorly constructed test items? An examination of how the *Eiken* test is scored and how the scores are reported follows. Finally, recommendations are made as to how the *Eiken* test can better serve students and teachers.

Background

The Society for Testing English Proficiency (STEP) was established over 30 years ago as part of a plan by Japan's Ministry of Education to develop education across the nation. Specifically, the goals of the STEP were to popularize and improve the level of practical English in Japan (*Nihon Eigo Kentei Kyokai*, 1994a). In 1963, the same year that the TOEFL

was first administered, the *Eiken* tests were given for the first time to about 37,000 people at three levels (first, second, and third). Five years later, the *Eiken* tests received official approval from the Ministry of Education, which began actively promoting the *Eiken* tests as important tests of English proficiency. With this stamp of approval, the number of test-takers soared, and continues to increase each year.[3] The original three-level test has grown to seven levels: first, pre-first, second, pre-second, third, fourth, and fifth, the first level being the most difficult. The most recent addition to this series was the pre-second level in 1994, introduced to bridge what was felt to be a wide gap in difficulty between the third and second levels.

For levels one to three, the tests are given in two stages. The first stage is a written test (reading and listening comprehension) and the second stage is a speaking (interview) test. Both stages are offered twice a year, in June and October.[4] The focus of this paper is on the first stage of the pre-second level *Eiken* test of June, 1994 (Nihon Eigo Kentei Kyokai, 1994b).

What kind of tests are the Eiken tests?

In the literature on language testing, two types of tests are found (Brown & Yamashita, 1995; Henning, 1987; Hughes, 1989), criterion-referenced tests (CRTs) and norm-referenced tests (NRTs). The differences between the two types lie, not in the actual items themselves, but in the purpose of the tests, how the tests are scored, and how the test scores are used. Therefore, just looking at the test instrument is not enough to determine what kind of test it is.

The purpose of a criterion-referenced test is to evaluate how well the test-taker can perform a specific set of tasks. For example, classroom and term tests evaluate how well a student has learned a defined set of material over a specific period of time. Brown (1996) explained that, "the interpretation of scores on a CRT is considered absolute in the sense that each student's score is meaningful without reference to other students' scores" (p. 2). Therefore, CRT scores do not necessarily conform to a normal distribution.

A norm-referenced test, on the other hand, measures general language abilities. Each student's score is interpreted relative to the scores of all the other students who took the test and the scores generally fall along a normal distribution curve. The TOEFL and TOEIC are norm-referenced tests. The purpose of the TOEFL is to evaluate the English proficiency of foreign nonnative speakers of English, primarily those who intend to study at colleges and universities in the United States or

Canada. Thus, the content of the TOEFL focuses on English for academic purposes. The TOEIC (Test of English for International Communication) is an English language proficiency test which measures how well non-native speakers of English can communicate in English with others in business, commerce, and industry.

Like TOEFL and TOEIC, the *Eiken* tests are English proficiency tests for non-native speakers of English. However, the *Eiken* tests are different in at least two ways. First, the *Eiken* tests are not just one test, but seven different tests. These divisions allow the test-makers to clearly define the material covered, a characteristic of a CRT. *Eikyo* described the contents of its pre-second level test, as follows:

> Successful examinees are able to understand and use general English needed in daily conversation. (High school level; appropriate for a wide range of ages, from high school students to adults in Japan.)
>
> The successful examinee is:
>
> (1) Able to converse about common daily topics. (Able to conduct simple business by telephone; to make easy explanations, leave messages, do simple interpretation, etc.)
>
> (2) Able to read material about common everyday topics. (Able to read news articles, letters, simple pamphlets, etc.)
>
> (3) Able to write about common everyday topics. (Able to write simple letters, notes, memos, etc.) (*Nihon Eigo Kentei Kyokai*, 1994a, p. 8)

Dividing the tests into seven levels is a practical way of handling the wide population, from junior high school to post-graduate levels, that *Eikyo* tests. Further, if students take and pass a test at a level appropriate to their ability, that success is seen to be motivational for their continued study.

The second point in which *Eiken* tests differ from TOEFL and TOEIC is its method of reporting scores. *Eiken* uses a pass/fail reporting system while TOEFL and TOEIC use a converted scale-score reporting system, which makes these two tests "'user-defined' in that scores can be considered in a variety of ways depending upon the requirements of a particular individual or client" (Wilson, 1993, p. 2). Although the reporting styles differ, all these tests follow NRT procedures by using some form of statistical analysis to translate raw scores into standard scores. In other words, none of the tests report their scores as absolute scores.

The question, "What kind of tests are the *Eiken* tests?" remains unanswered. According to Bostwick (1995), "the *Eiken* STEP claims to be a criterion-referenced test in that it specifies proficiency standards and attempts to identify whether the student can pass the pre-established standard" (p. 58). The fact that the *Eiken* tests are divided into seven levels, the

purpose of each level clearly defined by a set of specific tasks, with the language skills required to pass each level specifically defined, gives it qualities of a CRT. However, the way the test is scored (i.e., by converting raw scores to standard scores) is characteristic of an NRT. Therefore, it makes sense to call the *Eiken* tests hybrid CRT/NRT. Knowing that the tests are scored as NRTs is helpful in choosing appropriate reliability measures.

The Study

Method

Materials: The pre-second level test was originally developed for 2nd and 3rd-year high school students (16 and 17-year-olds) (*Nihon Eigo Kentei Kyokai*, 1994c). The most recent statistics show that this group forms the majority of test-takers. For the June, 1996 test, 75% of those who took the pre-second level (227,666 out of a total of 303,955) were senior high school students, 4% (12,471) were junior college students, and 3% (8,549) were university students (*Nihon Eigo Kentei Kyokai,* September, 1996).

There are 75 multiple-choice items on the pre-second level written test. Part 1, which tests vocabulary, idioms, grammar, usage, and reading composition, has 55 items, and Part 2, the listening section, has 20 items. Each item is worth one point for a total of 75 points.

Subjects: The subjects for this study were 182 first-year students (ages 18-20) in five classes at a junior college. Although the reports by *Eikyo* indicate that the pre-second level test is ideally suited to high school students, this higher age level was selected as it best matched the general ability and experience of the group, as outlined below.

A survey, in Japanese, accompanying the *Eiken* pre-second level test to determine the students' experience in taking it, showed that 17% of the students had tried the pre-second level test at least once before but failed, while 40% had never taken an *Eiken* test before. However, because the format of the *Eiken* pre-second level test is similar to high school English tests, it was concluded that lack of *Eiken* experience would not adversely affect the data. Fully 43% had previously passed the third level, confirming that the pre-second level was the most appropriate for this group.

Procedure: In May, 1996, the pre-second level test of June, 1994 was administered to all 182 Ss, along with the survey of *Eiken* experience. After a review of the results of the survey, the results for 14 Ss who had

previously passed pre-second level were eliminated from consideration. The remaining papers (N=168) were scored by hand. The test had 75 items, each worth one point.

Analysis

Analyses to evaluate test reliability were done on the data collected in three categories: Descriptive statistics, item statistics, and consistency estimates. Analyses to evaluate validity were done by comparing the contents of the test with the aims set out by *Eikyo* and the course of study for high school English education. The construction of the individual test items was also examined.

Reliability

Descriptive Statistics: Minimum score, maximum score, midpoint, mean, and standard deviation were calculated. The midpoint is the score which is halfway between the highest and the lowest score. The midpoint, together with the mean, are two statistics which help locate the middle or typical score (Brown, 1996, p. 109).

Item Statistics: Investigation into the reliability of the *Eiken* test began with a look at two types of item statistics: item facility and item discrimination. Item facility (IF) is "a statistical index used to examine the precentage of students who correctly answer a given item" (Brown, 1996, p. 64). The following formula for item facility was used to evaluate individual test items:

$$\text{IF} = \frac{N_1 \text{ (number of examinees who answered correctly)}}{N_2 \text{ (number of examinees who took the test)}}$$

Item discrimination (ID) "indicates the degree to which an item separates the students who performed well from those who performed badly" (Brown, 1996, p. 66). In order to calculate the ID index, it is necessary to differentiate the high scorers from the low scorers. In this study, the upper and lower thirds (33%, or 56 students each) were taken to represent the high scorers and the low scorers respectively. The ID was calculated as follows:

ID = IF (item facility of high scorers) - IF (item facility of low scorers)

That IF and ID are closely connected is apparent from the above formula: If the item is easy (i.e. has a high IF), there should be little discrimination (low ID), and if the item is rather difficult, the discrimination should be high. If the item is too difficult, there should be no

discrimination. According to Brown (1996, p. 69), ideal items in an NRT project have an average IF of .50 and the highest available ID. However, in reality, "items rarely have an IF of exactly .50, so that those that fall in a range between .30 and .70 are usually considered acceptable" (Brown, 1996, pp. 69-70). The items outside this range should either be set aside for revision or discarded. Examination of the ID of the remaining items further evaluates their suitability. The following guidelines were used to evaluate item discrimination:

.40 and up -	very good item
.30 - .39 -	reasonably good, but possibly subject to improvement
.20 - .29 -	marginal item, usually needing and being subject to improvement
below .19 -	poor item, to be rejected or improved by revision (Ebel, 1979, p. 267, cited in Brown, 1996, p. 70)

Consistency Estimates

In this study, the Kuder-Richardson Formula 20 (KR-20) reliability estimate and the standard error of measurement (SEM) were used (Brown, 1996, p. 199, p. 207) to estimate the *Eiken* pre-second level test's reliability. KR-20 was chosen over two other methods of calculating reliability, KR-21 and Cronbach's alpha, for two reasons: 1) it is reported to be the most accurate of the three (Brown, 1996, p. 199), and 2) the results could be compared with the reliability statistics of the TOEFL and TOEIC, which also use KR-20. Further, the *Eiken* pre-second level test follows criteria required by KR-20, in that each item is worth one point and is scored as correct/incorrect.

Test reliability is important because it measures the consistency of the test instrument. If a student takes a test on one day, and then takes the same test again a week or two later, it should produce nearly identical results, that is, if the test is a reliable instrument and little or no learning has taken place between the first and second testing. A test with a reliability coefficient of 1.0 would give precisely the same results for a particular group of test-takers regardless of when it was administered: it would be 100% reliable. Therefore, it is the goal of test-makers to attain the highest possible reliability coefficient.

The standard error of measurement (SEM) is another useful statistic for estimating reliability of NRTs. According to Brown (1996, p. 206), the SEM is "used to determine 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." A multitude of factors affect test perfor-

mance, only a few of which are connected with the test instrument itself, including items not matched to the purpose of the test, formats unfamiliar to the test-taker, and poorly constructed test items. The majority of factors affecting test performance are related to physical setting (i.e. the test room), and the mental and physical condition of the test taker. With all of these potential distracters, it is sensible to factor in the SEM when determining the cutoff scores.

Validity

A test is a valid instrument if it measures accurately what it claims to measure. For example, an arithmetic test of addition should contain only test items which ask students to add numbers. Further, the test items should not contain ambiguities or misleading information. In the case of the addition test, we can tell whether it is a valid instrument or not just by looking at it. Assessing the validity of English proficiency tests like the *Eiken* is a bit more complicated.

Validity analyses are done on a regular basis by TOEIC and TOEFL. According to a TOEIC report:

> The first validity studies involved the administration of TOEIC to a representative population of Japanese managers, technicians, bankers, and other employees who require English in their work. Researchers compared the candidates' performance on TOEIC to their performance on direct measures of listening, speaking, reading, and writing and determined the correlations. (TOEIC, 1995, p. 3)

Therefore, TOEIC compares its test results to direct four-skills test results to determine test validity. TOEFL conducts a similar type of validity analysis: "TOEFL validation is based upon correlations between test performance of foreign students studying in U.S. colleges and universities and their performance in degree-granting educational programs" (TOEIC, 1995, p. 3). Both TOEFL and TOEIC test for what Woodford calls "concurrent" validity:

> If a language test is supposed to measure whether a person can read Japanese or not then the person who scores high on the test should be able to pick up the Japanese newspaper and tell us what the lead article says. The low scorer should not be able to do it. (1980, p. 4)

It is not known whether *Eikyo* is also doing such validity studies.

Other validation studies involve comparing the results of one test with the results of another. In 1993, TOEIC (Wilson, 1993) published a report of a research study which linked the TOEIC listening section

scores to the scores of the Language Proficiency Interview, a direct assessment of oral language proficiency developed by the Foreign Service Institute of the U.S. Department of State. The numerical correlations between LPI and TOEIC listening sections (.83) proved to be consistently high, suggesting that both tests are, as they claim, effective measures of the ability to understand and use spoken English.

Content validity compares the test specifications with the test contents. If the individual test items match the specifications, then the test can be said to have content validity. This is a subjective evaluation which should be done by a group of testing experts. For the present study, the resources necessary to do the type of validity research described above were not available. Instead, four general questions pertaining to test validity were posed: 1) Is the pre-second level test really appropriate for the group *Eikyo* aims to examine? 2) Do the contents of the test items reflect aspects of "daily life" in Japan, as *Eikyo* claims (*Nihon Eigo Kentei Kyokai*, 1994a)? 3) Do the test items really measure the abilities that they purport to? and 4) Are there any poorly constructed test items?

Results and Discussion

Reliability

Descriptive Statistics: Table 1 shows the descriptive statistics. The midpoint and mean indicate that the typical scores for those who took the test were just above 50%.

Table 1: Descriptive Statistics of the Pre-Second Level *Eiken* Test (N=168; k=75)

Min.	Max.	Midpoint	*M*	*SD*
17	64	40.5 (54%)	38.33 (51%)	9.21

The standard deviation of the test scores was 9.21. Figure 1 shows that the scores of 80 Ss (48%) fell below the mean (0-37 points), and 78 Ss (46%) fell above the mean (39-64 points), while the scores of 10 Ss (6%) were exactly on the mean. The results show that the test performed like a true NRT, conforming to a normal distribution.

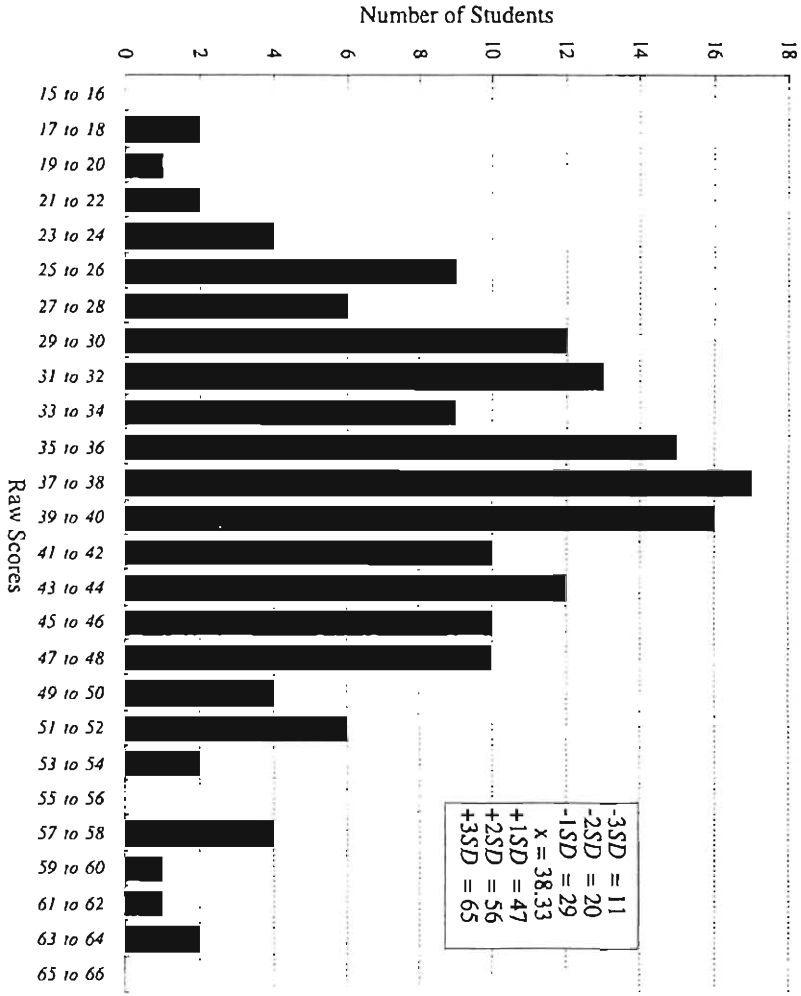

Figure 1: Distribution of Scores on Eiken, Pre-second Level (N=168, x=38.33, *SD* =9.21)

Item Statistics: You will recall that IF shows the percentage of students who answer a given item correctly. For example, 37 Ss answered the first item correctly. Therefore, the IF of item #1 (37/168) is .22 (see Table 2). This means that it was a difficult item because only 22% of the students got it right. Item #3, on the other hand, with an IF of .86, was an easy item for this group.

Table 2: Item Facility (IF) and Item Discrimination (ID) for *Eiken* Pre-Second Grade (N=168)

Item No.	*IF*	*ID*	*Eikyo's IF†*	Item No.	*IF*	*ID*	*Eikyo's IF†*
1	.22	.35	C	41	.81	.27	A
2 •	.36	.41	C	42	.31	.17	C
3	.86	.17	A	43	.72	.44	A
4	.38	.14	C	44•	.43	.30	B
5	.37	.22	D	45	.82	.27	A
6 •	.51	.31	C	46•	.51	.40	B
7 •	.42	.50	C	47•	.60	.37	B
8	.26	.08	D	48•	.55	.44	B
9 •	.50	.42	C	49•	.48	.38	B
10	.23	.14	D	50	.28	.21	C
11	.31	.23	C	51•	.42	.30	B
12•	.57	.37	B	52	.38	.22	C
13	.30	.27	D	53	.39	.16	B
14•	.48	.44	B	54	.18	.13	C
15•	.58	.42	C	55	.28	.16	C
16•	.48	.30	C	56	.60	.17	B
17•	.35	.30	C	57•	.66	.31	B
18	.60	.06	B	58	.89	.21	A
19	.12	.17	D	59	.77	.18	A
20•	.70	.42	B	60	.86	.09	A
21•	.37	.40	C	61•	.57	.37	B
22	.52	.26	B	62	.39	.02	C
23•	.48	.31	B	63•	.49	.40	B
24•	.41	.48	B	64	.65	.16	B
25	.26	.13	D	65	.54	.09	C
26	.80	.07	A	66	.79	.17	A
27	.35	.23	C	67•	.43	.30	B
28	.80	.25	A	68	.63	.37	A
29	.24	.14	D	69	.23	.07	D
30	.52	.21	B	70	.81	.21	A
31•	.55	.51	B	71	.51	.30	C
32	.40	.20	C	72	.94	.03	A
33•	.59	.33	B	73	.71	.19	B
34•	.48	.42	B	74	.26	.21	C
35	.86	.25	A	75•	.54	.45	B
36•	.68	.30	B				
37	.72	.23	A				
38•	.55	.35	B				
39	.48	.28	C				
40	.25	.25	D				

•= Good test items

† Note: IF reported in *Nihon Eigo Kentei Kyokai*, July, 1996, p. 30.

Item #2 (Table 2), has an IF of .36, and an ID of .41, and therefore meets the above IF and ID criteria. It is fairly well centered and discriminates well between high- and low-scoring students. Other items in the test which could be called good test items are indicated by an asterisk. However, there are many items which discriminate poorly. In fact, more than half of the items (59%) do not meet the ID and IF requirements of good test items, falling outside the acceptable range of .30 to .70. Based on these results, further refinements and improvements of many of the test items are needed.

Eikyo published a table approximating IF by assigning letter values to five ranges (Table 2): A=.8–1.0; B=.6–.79; C=.4–.59; D=.2–.39; E=0–.19 (*Nihon Eigo Kentei Kyokai*, July, 1996, p. 30). Because the ranges are so wide (0.19 points each), however, it is difficult to draw any conclusions about IF from these statistics. Besides, IF alone only tells what fraction of the group got the item correct. In order to draw any concrete conclusions about whether the item is functioning well, ID statistics are needed.

Consistency Estimates: The KR-20 result for this study was .82. It is important to keep in mind that this figure of 82% reliability is based on the in-house scores of this sample of 168 Ss. The actual test population, a much greater number with a broader age range, would have had a different reliability index. Unfortunately, this information has not been made available by *Eikyo*. Only a hint as to the reliability of the *Eiken* test was made by an *Eikyo* representative of the test development section, who said that in the years up to 1992, the reliability of *Eiken* tests was between .80 and .90 (*Eikyo* representative, name withheld, personal communication, July 25, 1995). However, he did not divulge the type of analysis done. In any case, this information is not directly relevant to the present study, which deals with the pre-second level test, first introduced in 1994. It is worth noting that in 1989 and 1990, test reliability for TOEIC using the K-R20 formula was .96 (Woodford, 1992).

The SEM for this test administration was 3.9, meaning that the band around which a student's score should be considered is ±4. Therefore, if a subject who scored 37 on the test were to take the test repeatedly, the scores could vary between 33 (-4) and 41 (+4). The passing level for the June 1996 pre-second level test set by *Eiken* was 38 and above, meaning a subject with 37 would have failed. With no apparent margin for error, all it takes is to be one point short to fail.

Validity

Four general questions pertaining to the test validity were posed. First, is the pre-second level test appropriate for the group *Eikyo* claims?

To determine if the test items were suitable for senior high school level students and above, the items of the June, 1994 pre-second level *Eiken* were compared with the nationally approved senior high school course of study (Wada, 1992) and one Ministry of Education approved textbook, *The Crown English Reading* (Hirano, et al., 1996). This revealed that most of the words and idioms on the pre-second level *Eiken* test are taught at some point during the three years of senior high school. [Exceptions are noted as follows: Section 1, item 10 - *vacant*; item 19 - Bill *came all the way from* Florida; item 20 - *take it easy*; Section 2(B), item 9 - *French, Italian, or Thousand Island salad dressing*; Section 3, item 2 - *Are you having some problems there?*; and Section 4(B), item 9 - *farther inland* (*Nihon Eigo Kentei Kyokai*, 1994b).]

The second question regarding test validity asked whether the contents of the test items reflect aspects of "daily life" in Japan, as *Eikyo* claims (*Nihon Eigo Kentei Kyokai*, 1994a)? At least two do not. The following observations are those of the author, not the results of rigorous evaluations by a team of testing experts.

Section 1, item 10:

When the sign on the door of a rest room says "( )," it means someone is using it.

1 OCCUPIED 2 VACANT 3 LIMITED 4 EMERGENCY

(*Nihon Eigo Kentei Kyokai*, 1994c, p. 9)

This item is problematic because restroom doors in Japan seldom have signs indicating whether the stall is vacant or occupied. (The exception is on airplanes, where restroom doors are equipped with such signs). The chance that students are familiar with the context of this item, in English or in Japanese, is remote.

Section 2(B), item 10:

A "Do you have the receipt?"

B "Well, it was a present, but it's too small."

C "What's the problem?"

D "I'd like to exchange this skirt, please."

1 C-B-A-D 2 A-D-C-B 3 D-C-B-A 4 B-C-D-A

(*Nihon Eigo Kentei Kyokai*, 1994c, p. 12)

This item, which asks students to put the four sentences in sequential order (3 is the correct choice), is problematic for three reasons: First, it is culturally inappropriate as Japanese do not customarily exchange items

of clothing that they have purchased, let alone received as a gift. Second, it is illogical. If the person received the skirt as a gift, it is unlikely that she would have the receipt. The third point is not concerned with item validity as much as consistency. In Section 2(B) there are five items, the first four of which follow a question-answer-question-answer sequence. This item, however, has a statement-question-answer-question structure. The fact that it is different from the other items in this section may be a source of confusion for the test takers. If *Eikyo* intended this confusion, it would be of interest to know the reason.

The third question regarding test validity to be asked is if the items on the test really measure the abilities that they claim to. *Eikyo*'s statement about what the successful pre-second level examinee is able to do, converse, read, and write about daily topics, implies that the *Eiken* tests test all four skills. Only listening and reading are actually tested. The TOEIC is similar to the *Eiken* tests in design, as it too tests only listening and reading. The TOEIC differs from the *Eiken* tests in that it measures listening and reading directly, and speaking and writing indirectly. Validity studies have been done to confirm a high correlation between TOEIC results and speaking and writing skills (Woodford, 1980). As *Eikyo* has not published studies to show correlations between its listening and reading tests to speaking and writing abilities, its claims that the successful examinee is "able to converse about daily topics" and is "able to write about common everyday topics" (*Nihon Eigo Kentei Kyokai*, 1994a) cannot be confirmed.

The final question regarding test validity asked if there were any poorly constructed test items. One example is from Section 3 (item 2):

Helen: What did you want to talk to me about? You sounded so mysterious on the telephone.

Sherri: Sorry, but I wanted to tell you this news face to face. I've decided to move.

Helen: But I thought you liked your neighborhood. (2)?

Sherri: No, everything is fine. I just need a change.

2 i) Are you having some problems there? ii) What's the problem?
iii) Isn't everything fine? iv) Why are you moving?

(*Nihon Eigo Kentei Kyokai*, 1994c, p. 13)

The difficulty with this item is in the first of the four possible responses, which also happens to be the correct answer, "Are you having some problems there?" According to Swan (1995), "*some* is most common in affirmative clauses, while *any* is common in questions and negatives" (p. 548). Further, "we use *some* in questions if we expect people to

answer "Yes," or want to encourage them to say "Yes" such as in offers and requests" (Swan, 1995, p. 548) (i.e., *Would you like some more coffee?*). Because choice '(i)' is a question, not an offer or request, it could be argued that it is inappropriate to use *some*.

Two examples of poorly constructed test items occur in the reading passages in Section 4. In the questions following the reading of "Volunteer Guides at Museums" (see Appendix), Item 5 required too much inferencing to make it a viable item:

(5) Professionals at some museums

1 think volunteers should not be paid.
2 feel they know less about museums than volunteers.
3 dislike volunteers because they know more than the professionals.
4 think volunteers cannot do the work of professionals.

(*Nihon Eigo Kentei Kyokai*, 1994c, p. 14)

The correct answer, number 4 (*Nihon Eigo Kentei Kyokai*, 1994c, p. 22), is primarily based on inference, the only clue in the reading passage being, They [professional scholars] feel that amateurs should not do the work of professionals, and that some volunteers act as if they knew everything. The problem is the interpretation of *cannot* in the answer and *should not* in the reading passage.

The second reading passage in Section 4, entitled, "Rainfall in Australia," is problematic in that the text does not correspond to one of the test items (number 7). The pertinent paragraphs and the test item in question are excerpted below:

> Most parts of Australia do not receive enough rainfall. In some places there are long periods when it doesn't rain at all. This lack of rainfall is one of the major reasons why such a large country as Australia has such a small population.
>
> Only one-sixth of the continent—a belt of land along the north, east, and south coasts—receives more than 40 inches of rain a year. The rest receives less than 40 inches, and farther inland are somewhat drier areas that receive between 10 and 20 inches.

(7) Where in Australia do they get more than forty inches of rain?

1 In the center of the south coast.
2 In narrow areas along the coasts.
3 In the areas which have monsoon climates.
4 In wide inland areas.

(*Nihon Eigo Kentei Kyokai*, 1994c, pp. 15-16)

Eikyo states that the correct answer is number 2, which implies that all four coasts receive rainfall. However, the supporting statement in the text is . . . *a belt of land along the north, east, and south coasts—receives more than 40 inches of rain a year*, specifically states only three coasts.

To summarize, it is clear that while most of the vocabulary found on this form of the *Eiken* test is appropriate for the intended examinees, problems of context and item construction make the validity questionable. Without evidence from *Eikyo*, it is difficult to conclude that the *Eiken* pre-second level test is a valid instrument.

How are the Eiken *tests scored?* According to *Eikyo*, the passing score for the pre-second level is "approximately 65%" (*Nihon Eigo Kentei Kyokai*, 1994a, p. 7). However, test score statistics since its introduction in 1994 show that the passing scores are much lower. In 1994, the passing percentages and scores were 55% (41+); in 1995, 56% (42+); and in 1996 (June), 49% (37+). This information is somewhat misleading as it is not made clear that these are standard, not raw, scores.

Eikyo's score reporting system is made an even greater mystery by the fact that students never actually see their test scores. All they receive is a report which states either "pass" or one of three categories of "fail," A, B, C.[5] An *Eikyo* representative explained: "A-level failure encompasses scores up 10 points below the passing score; B is up to 15 points below A, and C covers the remaining scores down to zero" (name withheld, personal correspondence, July 16, 1996). These "guidelines" conflict with a report of the pre-second level test of June, 1994, in which *Eikyo* stated that the passing score was 41; A was 34-40 (7 points below the passing score); B, 27-33 (7 points below A); and C, 26 or lower (*Nihon Eigo Kentei Kyokai*, 1994d). These ranges were consistent for 1995 and 1996 tests as well. The discrepancy between the explanation and the published scoring brings into question the integrity of the reporting system.

The following information (in translation) about how the *Eiken* test is scored was received from an *Eikyo* representative in the Planning Division:

> The passing score for the pre-second level is set at approximately 65%. However, the difficulty of the test inevitably varies from time to time, which leads to adjustment of the passing scores each time a new test is given. In the past, the adjustment of the scores has been done by a thorough item by item analysis, looking at the difficulty of each item [IF], and by using point biserial coefficiency. Recently, *Eikyo* has begun experimenting with another method of analysis, Item Response Theory (IRT), as a replacement for the above-mentioned item analysis procedure. (name withheld, personal correspondence, July 17, 1996)

Point biserial correlation is a calculation which shows item discrimination by computing the correlation between individual item responses and total test scores. Like the ID analysis, an item with a low point biserial coefficient may be discarded from the scoring, resulting in changes in the passing score each time a different test is given. However, there is no indication of whether items have ever been discarded by *Eikyo*.

Item response theory true score equating, also used by TOEFL and TOEIC (TOEFL, 1995, p. 9), converts raw scores to equivalent scaled scores. Although *Eikyo* claims to evaluate its test results using IRT methods, there are no published reports to substantiate these claims.

One final concern regarding test scoring is whether *Eikyo* sets a cut-off for the number of people who can pass. Reports in *Eikyo*'s monthly newsletter, *STEP News*, between 1991 and 1996, and information in their brochure, *The STEP Test?* suggest that this is a possibility. The percentage of people who passed the second level test has been consistent at 18% (1991-1996) and the pre-second level at 30-33% (1994-1996) for a number of years.

Conclusion

The results of this study indicate that the reliability and validity evaluations of the pre-second level *Eiken* test are not favorable. First, the reliability in this study is only .82. Is a test that is 82% reliable good enough? For the uninformed consumer, maybe; for test-makers, definitely not. The validity checks in this study show that the content of the test matches the intended group of test takers, perhaps the test's greatest strength. However, there are problems of clarity and context in the items themselves which need to be corrected. Finally, the item facility (IF) and item discrimination (ID) results in this study indicate that more than half of the test items should be revised or removed as they discriminate only fairly or poorly.

Eikyo has been operating a successful testing business in Japan for more than 30 years. In all likelihood, this trend will continue. However, published reports of studies by *Eikyo* on item construction, reliability, and validity are urgently needed to help consumers become better informed about the test, and to encourage research that would improve the quality of the test so that someday the *Eiken* tests might approach reliability in the high .90s.

Acknowledgments

Special thanks are given to Mikiya Koarai for his assistance with this paper and translations of the communications with Eikyo, *Yuji Ushiro for help deciphering Japanese statistical terms, and two anonymous* JALT Journal *reviewers for their advice.*

Laura MacGregor is an Associate professor at Sapporo International University (formerly Seishu University). She also teaches at Sapporo International University Attached Kindergarten.

Notes

1. The Ministry of Education endorses a total of 15 proficiency tests. In addition to the *Eiken* tests, three others are *Kohitsu Shosha Kentei* (penmanship), *Mohitsu Shosha Kentei* (calligraphy), and *Katei Ryori Gino Kentei* (cooking).
2. The *Eiken* tests are ranked from highest, *i-kyu* (first level), to lowest, *go-kyu* (fifth level). *Kyu* is translated here as level, rather than grade, as more appropriate for the *Eiken* ranking system.
3. The total number of people who took the *Eiken* test from 1990-1994 were as follows: (1990) 2,624,106; (1991) 2,761,771; (1992) 2,830,496; (1993) 2,895,912; and (1994) 3,374,140.
4. There is no second stage interview test for the fourth or fifth levels. For these levels, however the written tests are offered three times a year, in January, June, and October.
5. This style of score reporting is not unique to the *Eiken* tests. It is also used in the tests of secretarial skills (*Hisho Kentei*), and *kanji* proficiency (*Kanji Noryoku Shiken*).

References

Bostwick, M.R. (1995). Evaluating young EFL learners: Problems and solutions. In J.D. Brown & S.O. Yamashita (Eds.), *Language testing in Japan* (pp. 57-65). Tokyo: The Japan Association for Language Teaching.

Brown, J.D. (1995). Differences between norm-referenced and criterion-referenced tests. In J.D. Brown & S.O. Yamashita (Eds.), *Language testing in Japan* (pp. 12-19). Tokyo: The Japan Association for Language Teaching.

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

Brown, J.D. & Yamashita, S.O. (Eds.). (1995). *Language testing in Japan*. Tokyo: The Japan Association for Language Teaching.

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

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

Hirano, K., Homma, N., Tanabe, Y., Ikegami, Y., Matsuka, H., Yamamoto, S., Sugano, D., Schneider, D., & Wilson, B. (1996). *The crown English reading:*

Teacher's manual. Tokyo: Sanseido.

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

MacGregor, L. (1995). More on the *Eiken. The Language Teacher, 19*(11), 51, 86.

Nihon Eigo Kentei Kyokai. (1994a). *The STEP test?* Tokyo: Author.

Nihon Eigo Kentei Kyokai. (1994b). *Heisei 6-nen dai 1-kai kentei 1-ji shiken jun 2-kyu (1994, Heisei 6, pre-second level test, 1st stage, 1st administration).* Tokyo: Author.

Nihon Eigo Kentei Kyokai. (1994c). *Jun 2-kyu gaido (Guide to pre-second level).* Tokyo: Author.

Nihon Eigo Kentei Kyokai. (1994d). *Eiken jun kaijo reporuto (Eiken pre-release report).* Tokyo: Author.

Nihon Eigo Kentei Kyokai. (1996, July). *STEP News, 367.* Tokyo: Author.

Nihon Eigo Kentei Kyokai. (1996, September). *STEP News, 369.* Tokyo: Author.

Swan, M. (1995). *Practical English usage* (2nd ed.). Oxford: Oxford University Press.

TOEFL. (1995). *TOEFL test & score manual, 1995-96 edition.* Princeton, NJ: Educational Testing Service.

TOEIC. (1995). *Guide for TOEIC users.* Princeton, NJ: Educational Testing Service.

Wada, M. (1992). *Kotogakko gakushu shido yoryo (The course of study for senior high school).* Tokyo: Kairyudo.

Wilson, K. (1993). Relating TOEIC scores to oral proficiency interview ratings. *TOEIC Research Summaries, 1.* Princeton, NJ: Educational Testing Service.

Woodford, P. (1980, November). *The test for English for international communication.* Paper presented at the English-Speaking Union Conference, London.

Woodford, P. (1992). A historical overview of TOEIC and its mission. *The 35th TOEIC seminar* (pp. 10-15). Tokyo: The Institute for International Business Communication.

(Received Sept. 27, 1996; revised Dec. 17, 1996)