Review and Analysis of Asia University's 2013 Freshman English Placement Test, Transition from Version 2.4 to Version 2.5

Jeff Hull and Jay Brennan, Asia University

Abstract

In this article, we review and analyze the changes made in Asia University's Freshman English Placement Test, in particular the changes made in the word discrimination part of Version 2.4 to create Version 2.5 for the April 2013 test administration. The Assessments Committee in the university's Center for English Language Education (CELE) undertook the test revision in order to improve its overall performance and placement accuracy. We carried out standard measurements of test analysis, including measurements of the distribution of scores, means, standard error of measurement, reliability, item discrimination, and test difficulty in order to compare the two versions of the test. The analysis of the two versions of the test indicates that the changes made in the word discrimination part resulted in key test measurements declining, indicating that we need to reconsider how to improve the test.

Introduction

During the 2011 academic year, the Assessments Committee revised the Freshman English Placement Test (FEPT), reducing it from a 98-item, 54-minute test (Version 2.3) to a 75-item, 40-minute test (Version 2.4). The primary purposes of making that change were to make the administration of the test more consistent at the beginning and end of the academic year and to

obtain more complete scores for students at the end of the year. That would enable the Academic Office to make better placement decisions for students in English classes after their first year. Every effort was made not to compromise the reliability of the test. An analysis of how the condensed Version 2.4 of the test performed after its first year of use indicated that the Assessments Committee achieved its goal of developing a shorter test that places students in Freshman English classes as accurately as the longer version. A fuller account of the history of the test and the changes that were made for the 2012 administration of the test, along with an analysis of how the reduced Version 2.4 performed, can be found in two consecutive articles published in the *CELE Journal* in 2012 and 2013 (Hull, 2012a, pp. 1-11; Hull, 2013, pp. 1-17).

Hull's analysis in the more recent of the two articles also identified a direction forward for the Assessments Committee to make additional improvement to the test (Hull, 2013, pp. 13-15). The most important area of improvement to focus on was the test's first part, word discrimination. Only this part of the new version of the test did not show improvement. Consequently, changing that part became the Assessments Committee's primary focus for revising Version 2.4 to produce Version 2.5. Table 1 summarizes the changes that have been made in the test in recent years to help clarify the work the committee carried out in 2012.

Table 1: Summary of FEPT Development in Recent Years

Test/Year	Number of Items	Time	Sections/Parts
Version 2.3 2007	98	54:00	Listening Section: Parts 1-5 Vocabulary, Grammar and Reading Section: Parts 6-8
Version 2.4 2011	75	39:30	Listening Section: Parts 1-4 Vocabulary, Grammar and Reading Section: Parts 5-7
Pilot Test 2012	11 55	---	Listening Section: Alternative 1 Part 1 Word Discrimination Listening Section: Alternative 2 Part 1 Word Discrimination
Version 2.5 2013	75	39:30	Listening Section: Parts 1-4 (with 11 items from Alternative 2 Part 1 to make the new Part 1 of the test) Vocabulary, Grammar and Reading Section: Parts 5-7

Our purposes in this paper, then, are (a) to review and analyze how the condensed version of the test performed in its second year of use to determine whether it performs consistently over multiple administrations; (b) to assess how well the new word discrimination part that the Assessments Committee developed for the test performed; (c) to compare the number of complete scores obtained for students at the end of the 2012 to 2013 academic year with previous years in order to assess the overall effectiveness of the condensed version of the test in providing the Academic Office the scores it needs; and (d) to consider additional means of improving the test.

I. From Version 2.4 to Version 2.5

The Assessments Committee's approach to improving the word discrimination part of the test was to develop two alternative versions of that part, do a trial test of those two versions with International Relations students (because they no longer take the FEPT), evaluate whether one of the two versions discriminated among student levels more effectively than

the original, and then replace the original with the more effective of the two alternatives for the 2013-2014 test administration.

The Assessments Committee created one of the alternative versions by starting with the word discrimination format that already existed in Version 2.4. Several of the eleven items in Version 2.4 were edited, rewritten, and/or re-recorded. One of the main differences the committee focused on for the second alternative version was removing the contextual clues test takers could refer to (in the first alternative version) to help them identify a target word among possible answers.

Other minor revisions made to the test included editing a few items in other parts that the Assessments Committee felt would result in clearer answer choices for students and correcting a few language errors in a couple of test items. Otherwise, the test was kept the same as Version 2.4 so that the comparison of how the two versions performed would be focused on the revision of the word discrimination part.

In order to keep an accurate record of the Assessments Committee's approach to developing the FEPT and to potentially serve as a guide to future CELE Assessment Committees, we will provide a brief explanation here of the two alternatives developed for the word discrimination part of the test along with an analysis of the outcome of test piloting the two alternatives to arrive at Version 2.5.

A. Alternative 1: Word discrimination part

The format of a given question on the previous test, Version 2.4, was a stimulus sentence presented aurally with a target word under primary

phrase stress[1] (See Example 1). The target word was always the last word in the stimulus sentence. The test taker's task was to identify the target word from the five possible choices, four of which were distractors. The previous stimulus sentences with the target words and distractors from Version 2.4 were used for items 1-4 and 9. However, due to poor audio quality and poor performance, items 5-8 and 10-11 were redesigned and re-recorded. The Version 2.4 format was employed in Alternative 1. However, we tightly focused stimuli (questions), targets, and distractors (incorrect responses) on vowel or consonant discrimination, holding all other segments of a word constant whenever possible, as illustrated below in Example 1.

Example 1:

Modified Version 2.4, FEPT Pilot 2012
SECTION I: LISTENING
Stimulus: Why did they cut your share?
(A) share (B) fair (C) bear (D) hair (E) chair

B. Alternative 2: Word discrimination part

Because the format of questions on the previous test and Alternative 1 was stimulus sentences with the target words under primary phrase stress (See Example 1), we chose to eliminate phrases and use only a single word stimulus for each item in Alternative 2. The new stimulus in Alternative 2 word discrimination is illustrated in Example 2 below.

[1] Primary Phrase Stress is the syllable of a phrase that stands out because of its longer duration, louder sound, and its contrasting pitch, or some combination of these three acoustic features.

Example 2:

Version 2.4

SECTION I: LISTENING

Stimulus: "bug... bug…"

(A) bag (B) big (C) bug (D) beg

For the purposes of this section of our review, a word discrimination item will be considered valid only if the target and all distractors vary either the vowel phoneme or consonant phoneme and hold all other segments constant.

We revised the word discrimination part of Version 2.4 primarily because of concerns about item validity, as defined above. However, the Assessments Committee understands that Version 2.5 has validity issues; in particular, item seven violates our own definition of validity. Yet, there is considerable improvement in this area from Version 2.4 to Version 2.5. For example, item 7 is the only item in Version 2.5 that varies both vowel and consonant phonemes. However, a large number of items violate this principal in Version 2.4.

For example, consider Example 3 from FEPT Version 2.4 below.

Example 3:

Version 2.4

SECTION I: LISTENING

Stimulus: "May I borrow your **pen**?"

(A) pan (B) pen (C) pin (D) ban (E) bin

A student who chooses *ban* may do so because of a vowel discrimination problem or a consonant discrimination problem. We believe

that test items should be more closely focused. That is, each item should vary only a vowel in one position, holding all other segments constant whenever possible, if the test is trying to determine whether there is a vowel discrimination problem: alternatively, each item should vary only the consonant in one position, holding all other segments constant, if the test is trying to determine whether there is a consonant problem. If this section of FEPT 2.4 is carefully examined, one will find the majority of targets and distractors are not closely focused on either vowel or consonant discrimination and instead are ambiguous.

Additionally, FEPT 2.4, Part 1 has validity concerns for other reasons. Consider the question in Example 3, *May I borrow your pen?* This item uses a variable target. That is, some educated speakers of English pronounce *pen* as /*pɛn*/ and other educated speakers pronounce *pen* as /*pɪn*/. Therefore, it could be argued that both (B) and (C) are correct responses. Thus, all variable targets and distractors in vowel or consonant phoneme questions should be avoided.

Next, Version 2.4 items, such as Example 4 below, are not testing vowel or consonant phoneme identification. They are testing linking, and in this instance consonant-to-vowel linking. Although Messerklinger (2007, p. 16) cites Judy Gilbert's *Clear Speech* (1993) as the source of the idea for the past word discrimination portion of the test, upon reviewing this text we could not find one example of Gilbert mixing linking assessment with consonant and vowel discrimination assessment. These are two clearly separate areas to be assessed.

Example 4:

Version 2.4

SECTION I: LISTENING

Stimulus: "The teacher said that's **it**."

(A) at (B) it (C) sat (D) sit (E) set

If the testing of linking and trimming is the object of this part of the test, then instead of word identification on the basis of phoneme distinction, it would be appropriate to have one type of target and avoid mixing them as occurs in Version 2.4. For example, we could use varieties of trimming and varieties of linking to fill the eleven items of this part of the FEPT. This may be something for the committee to consider for future development of part one of the test.

Method

To address the problems described above, the Assessments Committee chose to focus targets and distractors on vowel or consonant discrimination, holding all other segments constant whenever possible, as illustrated below in Example 5.

Example 5:

Version 2.5

SECTION I: LISTENING

Stimulus: "bought… bought…"

(A) boat (B) bought (C) bout (D) boot

Also, because the format of items on the original test was a stimulus sentence with the target under primary phrase stress (see Example 1), we chose to eliminate the phrase and use only a single word stimulus. We felt that since students are notified in the instructions that the target will occur last, the phrase wasn't necessary given our intention was to test word discrimination as we have outlined. The new stimulus is illustrated in Example 5 above. In this example, the test taker would hear two iterations of the stimulus *bought*.

Item Selection Criterion

For selecting targets we used a three step process:

First, we consulted *Learner English* (Swan, M. and Smith, B. 2001, pp. 297-299) to determine which particular conflations pose the greatest difficulty for Japanese speakers. The most noticeable problems for Japanese speakers are /oʊ/ and /ɔ/ which are both pronounced as a long pure /oʊ/, causing confusion in minimal pairs like caught and coat, bought and boat. We used *Learner English* to create a list of the most common problems for Japanese learners to include in Alternative 2.

Second, we referred to Brown's *Teaching English Pronunciation* (1991, pp. 211-224) to determine the functional load or relative importance of each conflation. By relative importance, we mean relative frequency of occurrence of words that are set apart by only one unique feature. For example, the aforementioned /oʊ/ vs. /ɔ/ has a functional load of 10, which signifies the highest functional load score, or high relative frequency of occurrence in the English language. That is, there are many words in English that differ only by the vowel sounds /oʊ/ vs. /ɔ/.

Third, we designed test items giving priority to conflations of greatest importance (see pilot items with targets bolded in Appendix A). The committee created 55 items for the official pilot. These items were numbered 12 through 66.

C. Results of Test Piloting the Two Alternatives

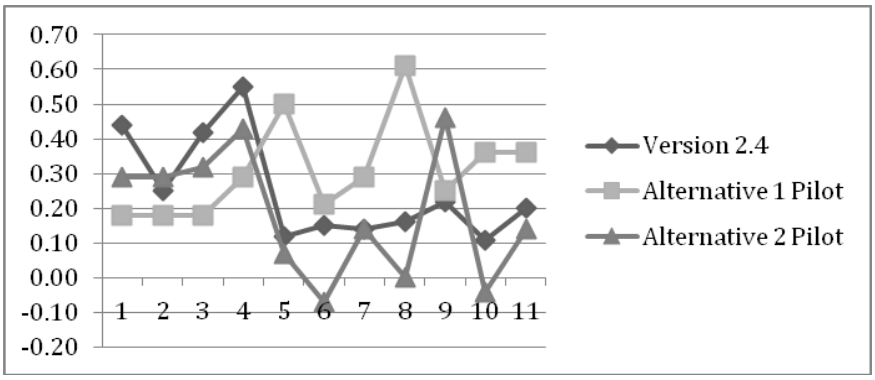
A comparison of item discrimination values of Version 2.4, from the 2012 administration of the test given to the entering class of freshmen, and Alternatives 1 and 2, from the pilot test (Table 2) given to International Relations students in November of 2012, suggests that Alternative 2 discriminates among student levels more accurately than both Alternative 1 and the original Version 2.4. Figure 1 shows a visual comparison of how each performed and supports the decision to use Alternative 2 in the 2013 administration of the FEPT Version 2.5.

However, it is interesting to note that Alternative 1 did not perform as well as the original Version 2.4. This is better illustrated in Figure 1, which shows that the modified items of Version 2.4 in Alternative 1 resulted in much lower discrimination values than the original Version 2.4.

Table 2: Average Discrimination Index Comparison of 2012 FEPT Ver. 2.4 and Alternative Pilot

FEPT Word Discrimination Version 2.4 & Pilot Alternatives 1 & 2	Item Disc Average for Items 1-11
Version 2.4	0.25
Alternative 1	0.16
Alternative 2 (Version 2.5)	0.31

Figure 1: Item Discrimination: Version 2.4 and Piloted Alternatives 1 & 2



If we consider Table 3, we can see that not one of the modified items resulted in an increased discrimination value.

Table 3: Comparison of Version 2.4 and Alternative Discrimination Values

Item Number	Version 2.4 Discrimination Value	Alternative 1 Discrimination Value
5	0.12	0.07
6	0.15	-0.07
7	0.14	0.14
8	0.16	0.00
10	.11	-0.04
11	.20	0.14

After examining the results of the pilot test, the Assessments Committee decided to implement Alternative 2. We examined Alternative 2 pilot data and used eleven out of the 54 items that resulted in the best discrimination values.

However, we omitted items that assessed redundant word discrimination contrasts. For example, the items in Table 4 were omitted even though they had discrimination values higher than other items that were

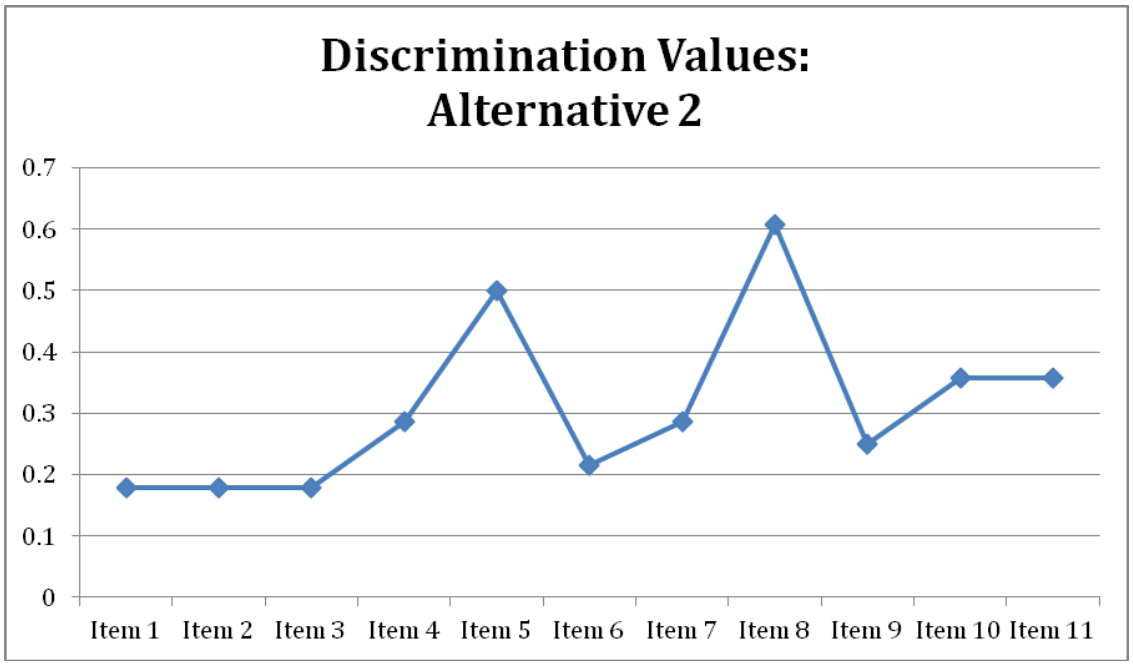
included in Version 2.5. This was because they all assessed sound discrimination between /æ/ and /ʌ/ (e.g., *rag* vs. *rug*).

Table 4: Sample of Alternative 2 Omitted Items

Omitted Item Number	Alternative 2 Discrimination Value	Corresponding Item Included in Version 2.5
21	0.36	8
33	0.25	8
43	0.25	8
62	0.25	8

The discrimination values of the items we selected for Version 2.5 word discrimination can be examined visually in Figure 2 below. For the purpose of clarity, the items are listed as they appear in Version 2.5 of the FEPT. However, the item number can be determined by referencing Appendix A. For instance, Item 12 of the Alternative 2 Pilot became item 9 in Version 2.5 of the FEPT.

Figure 2: Discrimination values of Alternative 2 as they appear in Version 2.5



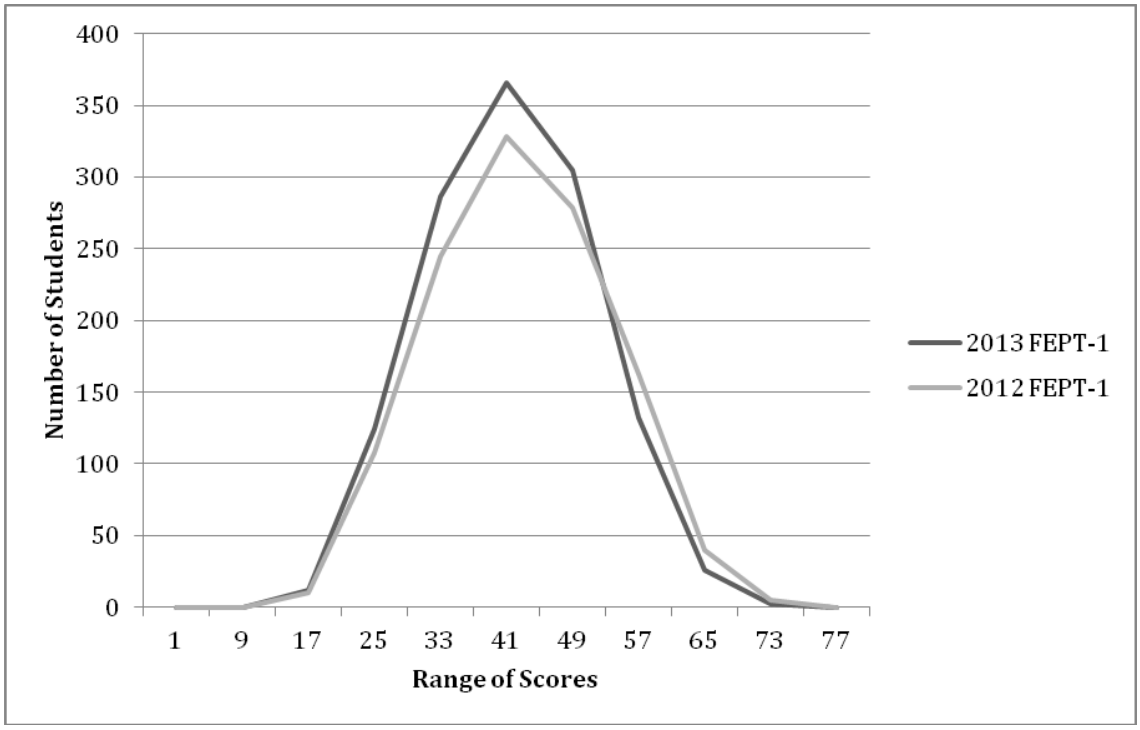
Based on the performance of these items, we concluded that Alternative 2 in the 2013 FEPT was clearly worth attempting. Therefore, based on the results of the piloting, the committee decided to proceed with Alternative 2 for the Version 2.5, April 2013 administration of the test.

II. Analysis of Version 2.5 FEPT

A. Distribution of Scores

The distribution of scores for the April 2013 FEPT as shown in Figure 3 is very similar to that of April 2012 (Hull, 2013a, p. 4-5). Like the 2012 results, most of the 2013 scores are within the middle 60 percent of the distribution. In addition, the shape of the graph is symmetrical, not noticeably slanted to the left or right. This, too, is similar to 2012 and indicates that the test was at an acceptably appropriate level of difficulty.

Figure 3: Distribution of Scores, April 2012 and 2013 FEPT



The standard deviation of 9.7 for the 2013 FEPT, as shown in Table 5, is somewhat lower than that of 2012. This indicates that the newer version of the test results in scores that are distributed less evenly over a narrower range than the 2012 version. Although one of the primary goals of placements tests is to separate students into different class levels (Harris, 1969, pp. 125-126), the 2013 distribution of scores resulted in more students being bunched together in the middle of the distribution. The standard error of measurement is the same as last year, again at a level appropriate to the scale of the test. The mean dropped slightly, which indicates that the newer test is slightly more difficult. We will address this point in section D (dealing with test difficulty).

Table 5: Details FEPT Test Measurements, 2012-2013

FEPT Test	Number of Items	Number of Examinees	Mean	Std. Error of Measurement	Std. Deviation
April 2012	75	1178	39.2	3.9	10.5
April 2013	75	1254	38.1	3.9	9.7

B. Reliability

Measuring a test's reliability (its ability to yield consistent results with a particular test group) is an important step in analyzing how well it functions. We calculated two of the most common measures of reliability, known as Cronbach's alpha and Kuder-Richardson 21, for the 2013 administration of the test so that we could compare it with the 2012 test. Table 6 shows that the reliability figures for the 2013 test were somewhat lower than the 2012 test. Although not a statistically significant loss in reliability, it does represent a slight step backward to the level of reliability that had been obtained in 2008 with a previous version of the test. Since that year, the level of reliability had shown an increase until this year.

No consensus exists for what is considered an acceptable level of reliability for placement tests. However, Hughes suggests that a range of .80 to .89 is desirable for a listening comprehension test and a range of .90 to .99 for a vocabulary, structure, and reading test (2009, p. 39). Harris (1969, p. 17), on the other hand, points out that tests which are not produced by independent professional testing organizations more typically have lower reliability measures in the .70s or .80s. He refers to these tests as "homemade tests." The FEPT, which has been produced and developed by CELE teachers over a number of years with limited resources and time, is certainly the kind of test Harris is referring to. Even as such, however, the FEPT has achieved an acceptable level of reliability, and the 2012 and 2013

administrations of the test continue to exhibit this. Despite the slight drop in the reliability measure, Version 2.5 still does an acceptably reliable job of placing students in Freshman English classes. It still fulfills the goal of separating students into four or five broad levels of ability, although there is some degree of overlap across the levels as there has always been.

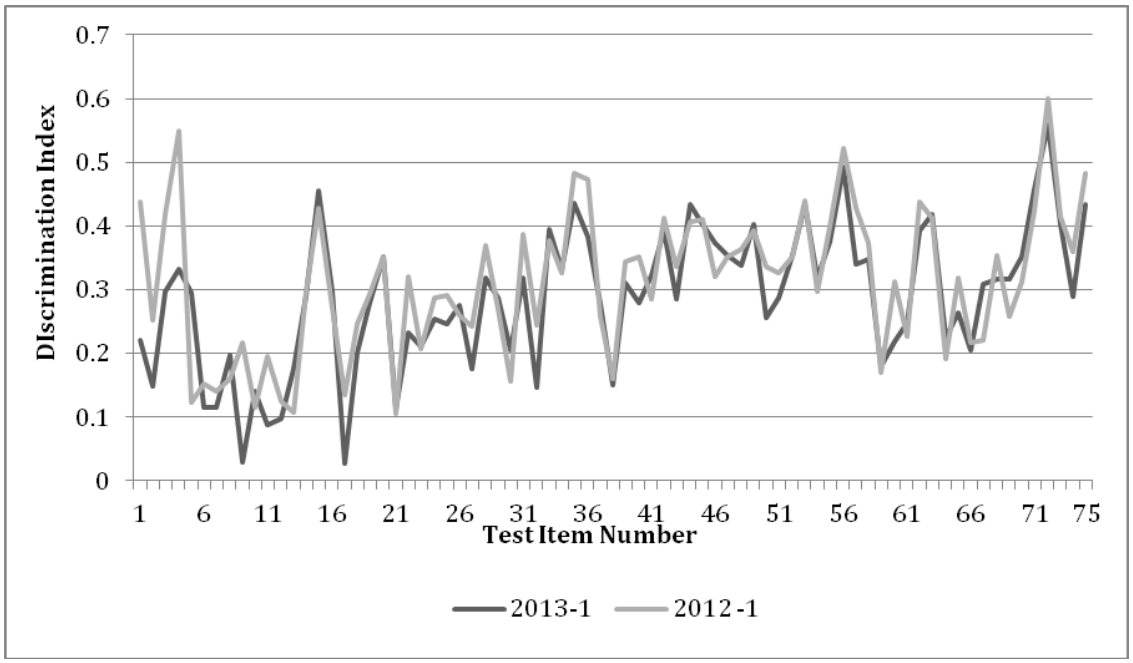
Table 6: Measurements of Reliability for the FEPT, 2008-2013

FEPT Test	Version of FEPT	Number of Items	Cronbach's alpa	KR21
April 2008	2.2	98	.84	.81
April 2010	2.3	98	.86	.84
April 2011	2.3	98	.85	.83
April 2012	2.4	75	.86	.84
April 2013	2.5	75	.84	.81

C. Item Discrimination

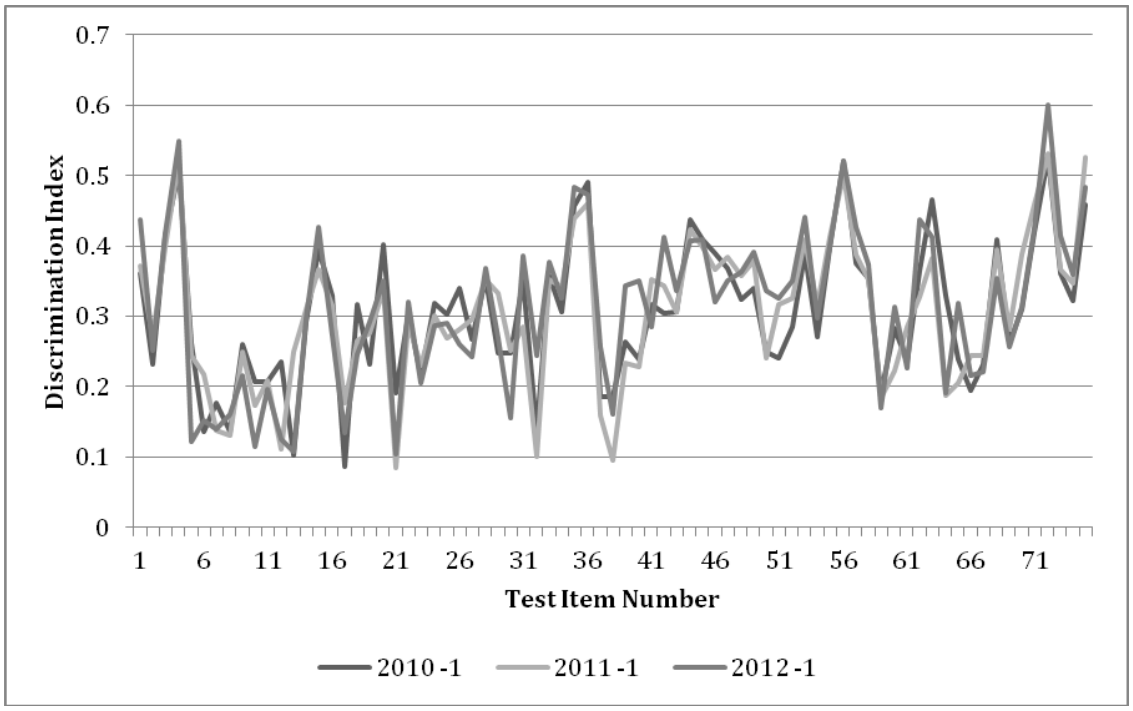
Measuring the item discrimination values of each test item reveals how well or how poorly it divides students of greater and lesser ability. Figure 4 shows a visual comparison of how versions 2.4 and 2.5 performed. With the exception of the early items in the test, the graph shows a considerable amount of consistency across the 75 items. The early items, which make up the word discrimination part, indicate a bigger gap between discrimination values in the two versions than at other sections along the graph.

Figure 4: Item Discrimination for the FEPT, 2012 and 2013



This is easier to see when comparing this graph with Figure 5, taken from Hull (2013a, p. 9), which compares discrimination values for the years 2010 through 2012. Figure 5 indicates a greater level of consistency at a slightly higher discrimination index in the early part of the test than is indicated in Figure 4.

Figure 5: Item Discrimination for the FEPT, 2010, 2011, and 2012



Isolating the discrimination values for the word discrimination part of the test and then comparing it with the values for the rest of the test reveals this difference in performance even more clearly. Table 7 shows that Items 1 through 11, which make up the word discrimination part, have a significantly lower discrimination value than the rest of the test. Isolating the performance of these same items on the previous versions of the test for the years 2010 to 2012 also reveals a noticeable decline in test performance in 2013. On the other hand, the rest of the items of the test exhibited the same level of performance across the same years. Table 7 also shows that this drop in the performance of the first part impacted the overall ability of the test to discriminate among student levels. Also worth noting here is that the discrimination index value for the word discrimination part dropped below

the 2.0 level that Hull (2012a, p. 6) identified as the level at which items should be reviewed for possible elimination or replacement from the test.

Table 7: Average Discrimination Index for the FEPT, 2010-2013

FEPT Test	Number of Items	Item Disc Ave for Items 1-11	Item Disc Ave for Items 12-75	Item Disc Ave for All Items
April 2010	75 (adjusted)	0.26	0.31	0.31
April 2011	75 (adjusted)	0.27	0.31	0.30
April 2012	75	0.25	0.32	0.31
April 2013	75	**0.18**	0.31	0.29

To complete the analysis of how the new version of the test performed in terms of item discrimination compared to last year and previous years, we can measure its individual parts. With the exception of Part 1, Table 8 shows that measures for Version 2.5 are reasonably consistent with that of Version 2.4. Although a condensed form of the test has been administered for only two years and the performance of the test will continue to need to be reviewed in future years, our results provide early evidence that the test is performing consistently over multiple administrations.

Table 8: Discrimination Index by Part

	Listening					Vocabulary, Grammar and Reading		
TEST	**Part 1**	**Part 2**	**Part 3**	**Part 4**	**Part 5**	**Part 6**	**Part 7**	**Part 8**
April 2010	.25	.22	.24	.29	.16	.32	.28	.28
April 2011	.25	.21	.22	.28	.14	.33	.26	.31
April 2012	.25	.25	.25	.32	Removed	.38	.29	.43
April 2013	**.18**	.26	.22	.29	Removed	.36	.29	.42

D. Test Difficulty

Table 9, which lists the average score by section of the test for the last four years, reveals that Version 2.5 has become slightly more difficult than Version 2.4. Like others in the field, Brown and Hudson (2002, p. 33)

consider an average score of around 50% to be the ideal level of difficulty for a test population. So the movement closer to a 50% level in 2013, although not statistically significant, is actually a welcome development.

Similar to last year, however, the table shows that the Vocabulary, Grammar and Reading section is comparatively easier than the listening section and easier than it was in previous years. This resulted from the committee's work in 2011 to condense the test to create Version 2.4. While that effort clearly improved discrimination values for parts 6-8 (as seen in Table 8), this is an area to which the Assessments Committee may want to devote some time in order to balance the difficulty level of the two sections of the test and move in the direction of an overall average score near the 50 percentile level. Easier items in the Vocabulary, Grammar and Reading section should be revised or replaced with more difficult items to identify students who are at the top third of the test group, while some of the more difficult and less discriminating items in the listening section could be replaced with less difficult items that differentiate students from the bottom to middle ability levels.

Table 9: Average Scores by Section and Overall (reported as percent correct)

Test	Listening	Vocabulary, Grammar, and Reading	Overall Average Score
April 2010	48.3%	50.9%	50%
April 2011	47.1%	51.3%	49%
April 2012	48.3%	56.8%	52%
April 2013	**45.8%**	56.3%	50.7%

As with the analysis of the item discrimination values, we carried out a more detailed comparison of the difficulty level of the word discrimination

part relative to the rest of the test (Table 10). The comparison reveals that the word discrimination part has become noticeably more difficult than in past years. On the other hand, the performance of the rest of the test is very similar to that of past tests. Although this increase in the difficulty level of the word discrimination part brings the overall difficulty level of the test closer to the ideal 50% level, it also contributes to the imbalance in the difficulty level of its two major sections.

Table 10: Average Scores by Word discrimination part and Overall (reported as percent correct)

Test	Number of Items	Facility Values Average for Items 1-11	Facility Values Average for Items 12-75	Fac. Ave for All Items
April 2010	75 (adjusted)	0.50	.53	.50
April 2011	75 (adjusted)	0.49	.52	.49
April 2012	75	0.50	.53	.52
April 2013	75	**0.45**	.52	.51

If we look at the greater detail Table 11 presents about the difficulty level of each part of the test, we can see again that the newest version of the test, with the exception of the first part, has measures that are reasonably consistent with the previous 2012 Version 2.4. One important note to make here, however, is that the FEPT was originally designed for the examinees to proceed from easier to more difficult items as they moved through each of the two major sections (Forster & Kerney, 1997, p. 145). This follows a rather standard progression recommended for tests designed to measure level of ability (Bachman, 1990, pp. 120-121). Table 11 shows that Part 1 has historically not achieved this goal and that the revised version of that part for the 2013 Version 2.5 has moved even further in the wrong direction. This added decline in the discrimination index value for the first part indicates a

clear direction for the Assessments Committee to continue to focus its attention.

Table 11: Average Scores by Part (reported as percent correct)

	Listening					Vocabulary, Grammar, and Reading		
TEST	**Part 1**	**Part 2**	**Part 3**	**Part 4**	**Part 5**	**Part 6**	**Part 7**	**Part 8**
April 2010	48.8%	55.7%	50.3%	46.6%	37.4%	56.4%	52.1%	40.6%
April 2011	47.3%	53.9%	50%	45.2%	37%	56.8%	51.5%	41.9%
April 2012	49.5%	51.5%	50%	45.1%	Removed	60%	54.8%	51.9%
April 2013	**44.6%**	51.1%	47.9%	43.3%	Removed	58.9%	55.4%	50.9%

E. Complete versus Partial Scores for End-of-Year FEPT

One of the original reasons for condensing the FEPT from a 98-item, 54-minute test to a 75-item, 40-minute test was to reduce the number of partial scores that resulted due to the way the test had been administered at the end of the year (Hull, 2012, p.1). Unlike the beginning of the year, when students are assigned to large testing halls according to department, and sufficient time is scheduled for them to take the test in one sitting, Freshman English instructors administered the test in their 45-minute classes at the end of the year. With this time limitation, instructors often chose to administer the test in two halves, the listening section in one class and the Vocabulary, Grammar and Reading section in another. As a result of this division of the test and inconsistent student attendance at the end of the year, a significant number of students ended up with scores for only one of the two sections of the test. This was difficult for the Academic Office, which relies on the end-of-year scores when placing students in English classes after their freshman year. For students with partial scores, the Academic Office has to refer back to entrance scores. An additional disadvantage of administering the test over two class periods is that one instructional class is lost.

Table 12 shows a significant decline in the number of partial scores as a result of the complete test being administered in one sitting at the end of the year, from a level of around five to seven percent down to 1.5 percent. Attendance issues still play a role in this and perhaps will never be entirely eliminated. There will always be students who either do not attend the class on the date the test is scheduled or arrive too late to complete enough of the test to be scored. Although the condensed form of the test has been used for only one and a half full years and the results here should therefore be considered preliminary, it appears that the Assessments Committee may have achieved its goal of reducing the number of partial scores at the end of the year. Another avenue the committee can explore to increase the number of complete scores would be to implement a make-up exam system for students who miss the test at the end of the year. This, too, would help reduce the gap between the number of students who take the test at the beginning of the year and at the end of the year.

Table 12: Complete versus Partial Scores for End-of-Year FEPT, 2010-2012

	Number of Examinees	Number of Complete Scores	Number of Partial Scores	Percentage of Partial Scores
2010-11	1047	979	70	6.6%
2011-12	871	827	44	5%
2012-13	916	902	14	1.5%

III. Conclusion

The analysis here of the performance of Version 2.5 of the FEPT indicates that the revision of the word discrimination part of the test has not resulted in improved performance of the FEPT. On the contrary, the slight decline in the reliability and item discrimination values indicates a small step

backward in the test's ability to differentiate among levels of students, although the difference is minor. Despite the small step backward, the test continues to retain an acceptable level of reliability in identifying a range of student ability levels sufficient for placement purposes at the beginning of the academic year. Furthermore, the committee appears to have made progress in reducing the number of partial scores at the end of the year, which has been a long-term problem.

Still, we should expect improvement in the test's performance as the committee continues to develop it. Based on the analysis of the performance of the test this year, and in particular the new version of the word discrimination part, one option for the committee would be to continue to focus on improving that first part of the test for the coming year. It continues to discriminate poorly among student ability levels and to violate the principle of sequencing the test so that it proceeds from easier to more difficult items within each of the two sections.

However, in follow-up discussions, after analyzing the outcome of the 2013 administration of the test, the committee has decided to take a slightly different approach for the coming year. We have reached the conclusion that it may be in our best interest to remove this historically problematic first part of the FEPT altogether. Assessing word discrimination in a homogenous student body may not be an accurate method of discriminating between student abilities. We cannot recall seeing this type of assessment being done anywhere else, and we believe very few Japanese students have exposure to focused word discrimination training.

Our effort may be better invested in developing existing parts of the FEPT that seem to perform more effectively. Parts two, three, and four have historically performed better than the first part, and if items that discriminate

more effectively can be added at a lower level of difficulty to those parts, it would not only help improve the performance of the listening section but also bring it into greater balance with the difficulty level and discrimination performance of the second half of the test. It would also result in a test that moves from easier to more difficult in each section, rather than beginning with items that are among the most difficult.

Beyond this, the committee should continue to monitor the overall performance of the test each year to confirm its consistency of performance over multiple administrations and identify items that continue to discriminate poorly among student levels so that they can be replaced with more effective items.

At the same time, the committee should continue to explore the possibility of using a commercial alternative to the FEPT. Despite its best efforts, the Assessments Committee, with its limited time and resources, cannot compete with the resources of a professional test-making organization. The committee has recently learned that the publisher of one of the textbook series currently being used for Freshman English has a placement test that is specifically tailored to placing students in one or another of the levels in that textbook series. The publisher has indicated it will allow the Assessments Committee to modify the test to fit the constraints of the CELE program, specifically the limited amount of time we have available for administering the test. If the committee is able to modify the test successfully, it has the potential of not only providing a test of higher professional quality but also meeting more desirable standards of a placement test (Hull, 2013, p. 16). These standards posit that if the test was more closely connected to the curriculum of the Freshman English program

it would result in better student placements and come much closer to a true measure of student improvement from the beginning to the end of the year.

References

Alderson, C. J., 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, A. (1991). Functional load and the teaching of pronunciation. In Brown, A. (ed.). *Teaching English pronunciation*. London: Routledge.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Forster, D. E., & Kearney, M. (1997). Writing the Freshman English Test (FEPT). *ELERI Journal*, *5*, 144-157.

Gilbert, J. B. (1993). *Clear speech*, Cambridge: Cambridge University Press.

Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill, Inc.

Hughes, A. 2009. *Testing for language teachers*. Cambridge: Cambridge University Press.

Hull, J. (2013). Review and analysis of Asia University's 2012 Freshman English Placement Test, transition from version 2.3 to version 2.4. *CELE Journal*, *20*, 1-11.

Hull, J. (2012). Modifying Asia University's Freshman English Placement Test. *CELE Journal*, *20*, 1-11.

Hull, J. (2012). Results of the 2010-11 FEPT and TOEIC tests. *CELE JOURNAL*, *20*, 34-38.

Messerklinger, J. (2007). The new Freshman English Placement Test, *CELE Journal*, *15*, 11-22.

Messerklinger, J. (2009). Results of the 2008 FEPT. *CELE JOURNAL*, *17*, 49-59.
Swan, M. and Smith, B. (2001). *Learner English: A teacher's guide to interference and other problems, volume 1*. Cambridge: Cambridge University Press.

Appendix A

12 (9)	a	boat	**b**	**bought**	c	bout	d	boot
13	a	sin	b	shin	**c**	**thin**	d	tin
14	a	stock	**b**	**stack**	c	stuck	d	stoke
15(11)	**a**	**stir**	b	store	c	star	d	steer
16	a	share	b	fair	c	bear	**d**	**hair**
17	a	team	b	tear	c	tease	**d**	**teeth**
18	a	berry	b	ferry	**c**	**very**	d	Perry
19 (6)	a	sim	b	sing	**c**	**sin**	d	sit
20	a	binger	**b**	**bigger**	c	beamer	d	beener
21	a	mad	b	med	**c**	**mud**	d	mode
22	**a**	**bowl**	b	bill	c	ball	d	bell
23 (5)	a	fur	**b**	**far**	c	for	d	fair
24	**a**	**fair**	b	share	c	care	d	hair
25	**a**	**thigh**	b	shy	c	sigh	d	tie
26	a	they	b	say	c	shay	**d**	**Jay**
27	a	binger	**b**	**bigger**	c	beamer	d	beeler
28	**a**	**do**	b	Jew	c	two	d	chew
29	a	low	b	so	**c**	**no**	d	toe
30 (4)	a	faith	b	fate	c	fade	**d**	**phase**
31	a	bin	b	fin	c	vin	**d**	**pin**
32 (1)	a	verse	b	first	**c**	**thirst**	d	burst
33	a	rag	**b**	**rug**	c	rogue	d	rig
34	**a**	**parse**	b	purse	c	pierce	d	Paris
35	a	van	b	pan	c	fan	**d**	**ban**
36	a	rum	b	run	**c**	**rung**	d	rug
37	a	she	**b**	**see**	c	zee	d	gee
38	a	beep	b	jeep	c	cheap	**d**	**deep**
39	**a**	**caught**	b	coot	c	cell	d	coat
40	a	source	b	horse	**c**	**force**	d	course

41(10)	a	thank	**b**	**sank**	c	tank	d	shank
42	a	thaw	b	saw	c	law	**d**	**jaw**
43	a	swarm	b	swim	**c**	**swam**	d	swum
44	**a**	**foam**	b	comb	c	home	d	dome
45	**a**	**sick**	b	thick	c	tick	d	Shick
46	a	butter	b	cutter	**c**	**putter**	d	rutter
47	**a**	**flow**	b	fly	c	flee	d	flaw
48	a	they	**b**	**say**	c	Jay	d	shay
49	a	page	b	pave	c	pays	**d**	**paid**
50	a	barn	b	born	c	burr	**d**	**burn**
51 (3)	a	phone	**b**	**hone**	c	shown	d	lone
52	a	breathe	**b**	**breeze**	c	breed	d	bereave
53	a	goat	b	boat	**c**	**vote**	d	note
54	**a**	**win**	b	wing	c	whim	d	wig
55	a	ache	b	age	c	ate	**d**	**aid**
56 (8)	a	bag	b	big	**c**	**bug**	d	beg
57	**a**	**fur**	b	for	c	fair	d	far
58 (7)	a	sip	**b**	**seep**	c	sheep	d	sleep
59	a	rows	b	froze	c	those	**d**	**throws**
60	a	night	b	kite	**c**	**light**	d	fight
61 (2)	a	cooled	**b**	**cold**	c	called	d	killed
62	**a**	**luck**	b	Luke	c	lack	d	lock
63	a	pert	b	port	c	paired	**d**	**part**
64	a	bold	b	boiled	**c**	**bald**	d	billed
65	**a**	**turn**	b	torn	c	tarn	d	term
66	a	right	b	night	c	might	**d**	**light**