Review and Analysis of Asia University's 2012 Freshman English Placement Test, Transition from Version 2.3 to Version 2.4

Jeff Hull, Asia University

Abstract

The purpose of this article is to review and analyze the changes made in Asia University's Freshman English Placement Test, in particular the changes made in Version 2.3 to create Version 2.4 of the test for the April 2012 administration of the test. The revision of the test was undertaken by the Assessments Committee under the direction of the university's Center for English Language Education in order to produce a test that could be administered within a 45-minute class period. Standard measurements of test analysis were carried out in order to compare the two versions of the test, including measurements of the distribution of scores, means, standard error of measurement, reliability, item discrimination, and test difficulty. The analysis of the two versions of the test indicates that Version 2.4 will perform as well as a placement instrument as Version 2.3 despite being a shortened version of the earlier test.

Introduction

Since 1997, one or another version of the Freshman English Placement Test (FEPT) has been used at Asia University to test first year students at the beginning of the year for placement in Freshman English classes and then again at the end of the first year for placement in English classes after the first year. Over the years, the test has been modified a

number of times and has varied in length from a 75-item, 45-minute test to a 100-item, 60-minute test.

As a result of problems created by inconsistent administration of the test at the beginning and ending of the year combined with poor attendance rates by first year students in their Freshman English classes (Hull, 2012b, pp. 34-35), in 2011 the Assessments Committee at the Center for English Language Education (CELE) was assigned the responsibility of condensing the version of the test being used at that time from a 54-minute test to a 40-minute test. The goal was, as much as possible, to reduce the length of the test without compromising the acceptable degree of reliability its developers had achieved. More consistent and complete scores obtained for students at the end of their freshman year would enable the Academic Office to make placement of students in post first year English classes with greater confidence. In the end, a 72-item, 40-minute test was proposed, and approved, that would hopefully achieve the goal of being able to be administered in a 45-minute class without sacrificing a significant degree of test reliability (Hull, 2012a, pp. 9-10).

This paper reviews the changes that were made in Version 2.3 of the test in order to produce Version 2.4 and, using standard test analysis methods, evaluates whether the newer, condensed version of the test will function as well as a placement instrument as the former version of the test. In addition, the paper considers a direction forward in continuing the development of the test.

I. From Version 2.3 to Version 2.4 FEPT

As the Assessments Committee went through the editing process of reducing the 98-item, 54-minute version 2.3 of the FEPT to produce the originally proposed 72-item, 40-minute version 2.4 of the test during the

2011 academic year, it discovered that it would be possible to add three items to version 2.4 and still achieve the goal of a 40-minute test. Taking into consideration that historically the listening section of the test had been given greater emphasis because of the focus on oral communication skills in the Freshman English classes, two test items were added back into the listening section and one item to the vocabulary, grammar and reading section of the test. The result was a test that had 40 items in the listening section, and 35 items in the vocabulary, grammar and reading section.

The other major revision of the test was to re-record the Japanese parts of the audio with a native speaker of Japanese. For version 2.3, a non-native speaker delivered all of the Japanese instructions. The committee's position was that this presented at least an unnecessary distraction and potentially a source of confusion to new university students taking the test, since only a limited number of recent high school graduates have familiarity with non-native Japanese language speakers.

Other revisions made to the test included rewriting some of the Japanese instructions in the test to make them clearer, improving some of the pictures in part two of the listening section that were identified as potentially difficult for test-takers to understand, and spacing some of the test items farther out over the pages to make the test easier to read overall.

Otherwise, the test was kept the same as Version 2.3, so that the comparison of how the two versions performed would be based on the selective deletions of test items from the former version rather than a rewriting of test items that exist in both tests. It is important to note here that the position of the committee and the administration in regard to version 2.3 of the test was that, basically, it was functioning effectively. It was differentiating between students with a reasonable level of reliability for

placement purposes. But there was a consensus it could be improved, particularly if it could be administered in a more consistent way at the beginning and end of the academic year and yield more complete test scores at the end of the year.

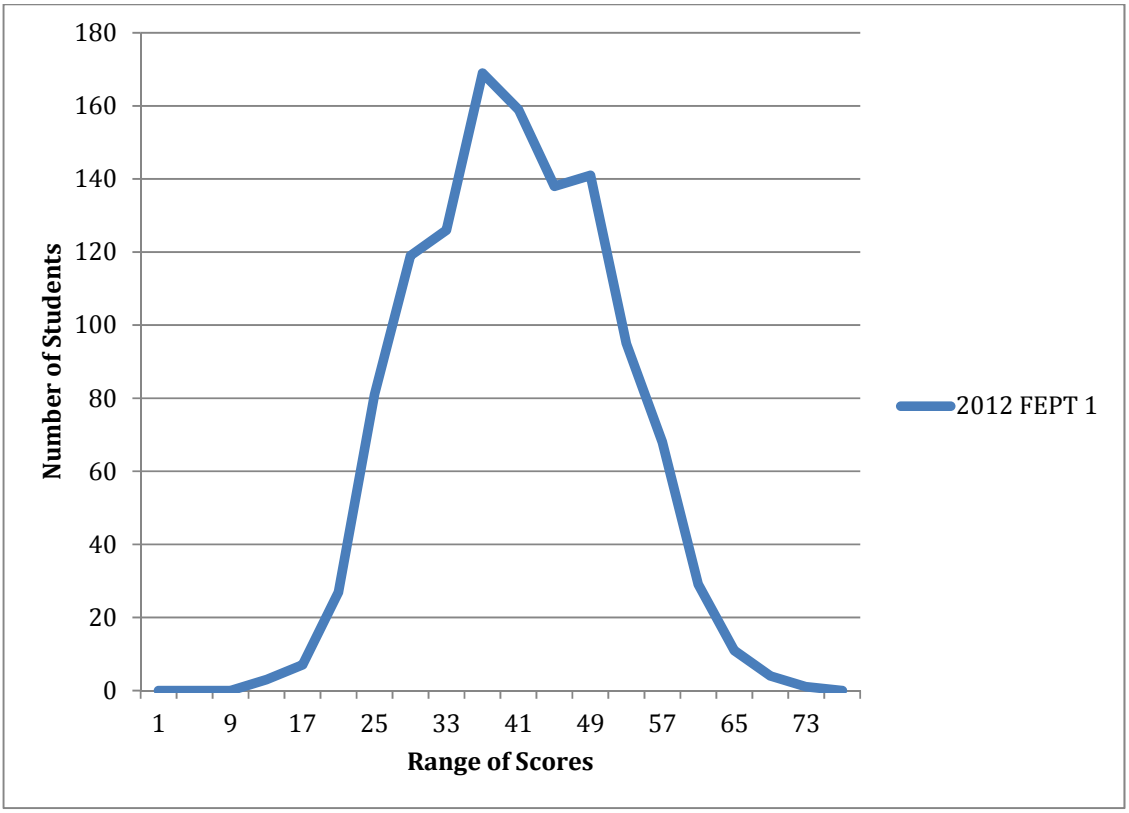
II. Analysis of the FEPT

A. Distribution of Scores

The distribution of scores for the April 2012 FEPT is comparable to those of April 2010 and 2011 (Hull, 2012a, p. 2). The range is smaller, to be sure, since there are 25 fewer items. However, like 2010 and 2011, most of the 2012 scores fall within the middle 60 percent of the distribution. Also, the shape of the graph is reasonably symmetrical, not noticeably skewed to the left or right, which is also similar to 2010 and 2011, indicating that the test was at approximately the appropriate level of difficulty for the population of students being tested.

Figure 1

Distribution of Scores, 2012 FEPT 1



The standard deviation of 10.5 for the 2012 administration of the test is very close to the measures for previous years despite the fact there is a significant reduction in the total number of items in the test. Generally speaking, in the case of placement tests, in which the goal is to separate out students into different class levels, a broader but symmetrical distribution of scores is helpful (Harris, 1969, pp. 125-126). A deviation of 10.5 for a test of 75 items separates out students' scores about as effectively as the slightly higher deviations of 11.7 or 11.9 seen in the 2010 and 2011 FEPT for the 98-item test. This dispersal of scores makes it much easier to place students in a range of classes than if there were a smaller level of deviation that resulted

in students' scores being bunched together around the median, a natural concern when reducing the overall number of items in a test.

Table 1:

Details FEPT Test Measurements, 2010-2012

FEPT Test	Number of Items	Number of Examinees	Mean	Std. Error of Measurement	Std. Deviation
April 2010	98	1259	48.6	4.45	11.9
April 2011	98	1106	48.1	4.51	11.7
April 2012	75	1178	39.2	3.9	10.5

The standard error of measurement also decreased with the 2012 test at a level that is appropriate in proportion to the number of items that were reduced from the test.

B. Reliability

Measuring a test's reliability, its ability to give consistent results with a particular test population from one administration of the test to another, is a critical step in analyzing how well it is functioning. Two standard measures of reliability, Cronbach's alpha and Kuder-Richardson 21, were calculated for the 2012 administration of the test in order to compare it with past results. Again, as we can see in Table 2, the measures for the 2012 test are very similar to reports of previous tests. There was no loss in reliability and actually a slight improvement.

Although there is no definite standard for what is considered an acceptable reliability value for a placement test, some suggest that a listening comprehension test should be in the range of .80 to .89 while a vocabulary, structure and reading test should be in the range of .90 to .99 (Hughes, 2009, p. 39). Harris (1969, p. 17) states that lower reliability measures in the .70s or .80s are more typical of what he refers to as

"homemade" tests, tests which are not produced by independent professionally recognized testing organizations. The FEPT could certainly be considered in this category of tests since it is produced by CELE teachers with limited resources, support and time at Asia University.

Normally, one way of increasing the reliability of a test is to lengthen it, as long as the additional items are of similar quality and difficulty in comparison to the original test. The point worth noting here, however, with the first administration of the new version of the test, is that the level of reliability has been maintained; although, the number of items in the test has been reduced by more than 20 percent.

Table 2

Measurements of Reliability for the FEPT, 2008-2012

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

Based on the analysis of the first administration of version 2.4 of the FEPT, it would appear that the test does as reasonable a job of placing students in Freshman English classes as version 2.3. As long as the test is not expected to make very detailed distinctions between student levels for placement purposes, the new version of the test should perform adequately. It can separate students into four or five broad levels of ability although with some degree of overlap across the levels just as the previous version had.

C. Item Discrimination

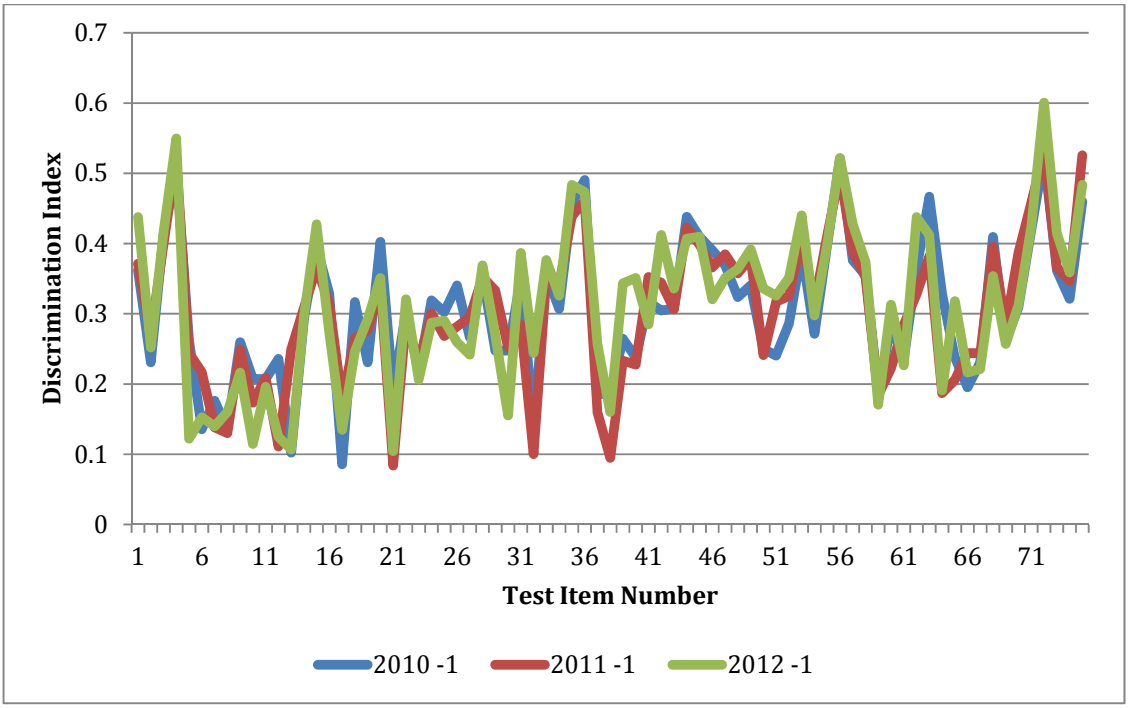
Item discrimination, analyzing how well or poorly individual test items divide students of greater and lesser proficiency, was used to decide

which items were best to eliminate from version 2.3 (Hull, 2012a, p. 6). Those items with the lowest discrimination indexes, particularly those with a value below .2, were immediate candidates for elimination although issues of balancing the number of items in the different sections of the test and the overall difficulty level of the test were taken into consideration.

Figure 2 shows how the 75 items of Version 2.4 performed in 2012 compared to how those same 75 items in Version 2.3 performed in two previous occasions the test was administered. The graph exhibits a great deal of consistency in how the 75 items functioned despite the fact that version 2.3 had 23 more items. Worth noting here, then, is that the ability of the 75 items that were retained in the test to discriminate among students was not adversely affected by the deletion of the 23 other items.

Figure 2

Item Discrimination for the FEPT, 2010-2012



As shown in Table 3, the average discrimination index for Version 2.4 of the test is .31, a clear improvement over that reported for previous years (Hull, 2012a, p. 6; Messerklinger, 2009, p. 53). This provides additional support to the position that Version 2.4 places students as effectively as previous versions. On the other hand, although there are no clear guidelines for what D.I.s are acceptable, item writers are often satisfied with an item D.I. of +.4 (Alderson, Clapham & Wall, 1995, p. 82) indicating that there is still considerable room for improvement in the FEPT in this area.

Table 3

Average Discrimination Index for the FEPT, 2007-2012

FEPT Test	Version	Number of Items	Average Discrimination Index
April, 2007	2.2	98	.25
April, 2008	2.2	98	.26
April, 2010	2.3	98	.27
April, 2011	2.3	98	.26
April, 2012	2.4	75	.31

Another perspective from which to view the test's ability to discriminate among students is to examine how each part of the test functions individually. Table 4 shows the average discrimination index by part. Comparing the 2012 values with those of previous years, we see increases in the values and therefore improvement in the ability of the test to separate out students in all but part one. Another observation that can be made here is the appropriateness of removing part five of Version 2.3 of the test because of that part's very poor performance in discriminating among students. The removal of part five alone may have done more to help the new version retain the previous version's reliability and ability to discriminate among students than any other change that has been made.

Table 4

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

D. Test Difficulty

Version 2.4 of the test has become a little easier than Version 2.3. Table 5 shows the average score by section of the test and overall for the last three years. Generally, an average score of around 50% indicates an appropriate level of difficulty for a test population (Brown & Hudson, 2002, p. 33). Whereas previous administrations of the test reported here were slightly under or right at 50%, the 2012 administration was 52 %, or two percentage points over what would be absolutely ideal for this test population.

This is a natural outcome of the higher number of more difficult test items being removed because of their low discrimination values. However, this is an area the Assessments Committee may want to devote some attention to in the future. In particular, the table shows that the vocabulary, grammar and reading section has become comparatively easier. Test items should be revised or replaced in this section with more difficult items to help distinguish students who are at the top third of the performance scale.

Table 5

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%

Table 6 provides more detail about the difficulty of the test by breaking it down by each part. Comparing the difficulty levels of the different parts makes clear how mismatched part five of Version 2.3 was with the rest of the test. Not only did that section have the poorest discrimination value by far, but it was also disproportionately more difficult

than the other sections of the test. As originally designed, the test was to proceed from easier to more difficult items. However, the jump of approximately 9% in difficulty level from Part 4 to Part 5, combined with the low discrimination value of the section, resulted in a series of items that did not effectively separate out students and must have been frustrating for most students since so few of them were able to respond correctly. The increase in difficulty of the subjects and vocabulary in Part 5 compared to Part 4 was most likely responsible for this.

Another observation that can be made about the average scores by part is in regard to Part 8. As can be seen from the 10% increase in correct responses from 2011 to 2012, Part 8 has become significantly easier. This is not necessarily a problem since the percentages of correct responses for Parts 6, 7 and 8 show a progression from easier to more difficult, an order that is generally considered appropriate for tests (Forster & Kerney, 1997, p. 145). The key point to consider is how effectively Part 8 discriminates among student levels. The fact that Version 2.4, part 8 has a significantly better DI compared to previous versions of the test, as can be seen in Table 4, indicates that Part 8 has actually been improved despite the fact that it has become easier.

Table 6

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%

III. Modifications of the FEPT

Although the findings here must be considered preliminary since version 2.4 has been administered only one time, the findings are clearly favorable. The early indications are that the new, reduced form of the FEPT will perform as well as the previous, longer versions of the test in terms of placing students in Freshman English classes at the beginning of the year and will have the added advantage of providing more complete scores at the end of the academic year.

Nevertheless, there is still much room for improvement. First of all, the test needs to be reviewed and analyzed each year to see how it is performing. It will be particularly important to compare the number of complete scores obtained for students at the end of the 2012-13 academic year with previous years in order to assess the overall effectiveness of reducing the test to 40 minutes. It will also be important to review the performance of the test after April of 2013 to see whether the results reported in this paper are consistent over multiple administrations.

Second, the analysis of item discrimination and test difficulty reveal an immediate direction for revision of the test. The analysis above of item discrimination reveals that Part 1 of the listening section was the only part in which the index did not improve. In fact, looking more closely at the first part reveals an unbalanced set of discrimination values across the items that make up that part. As Table 7 indicates, although a few of the eleven items have relatively strong discrimination indexes of .4 or above, five have very weak indexes below .2 and three others are just above .2, still considerably below the discrimination index of .31 for the test overall.

Table 7

Discrimination Index Values for Part One

Discrimination Index Values											
Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
April 2012	.44	.25	.42	.55	.12	.15	.14	.16	.22	.11	.2

In addition to the weak discrimination values, this part of the test violates a basic testing principle that the original makers tried to incorporate into the test: within each of the two major sections of the test, the items would proceed from easiest to most difficult (Forster & Kerney, 1997, p. 145). Although the easy to difficult order may not apply in all testing contexts, as Bachman points out (1990, pp. 120-121), tests that are designed to measure level of ability are typically sequenced this way.

Reviewing the charts presented in the analysis of test difficulty above, one can see that, with the exception of the first part, the test does proceed in that order. The percentage of correct responses decreases with each successive section. For all of the years reported, part one had a higher level of difficulty than the two parts that followed it. However, one issue the committee will have to pay attention to, as noted in section D above, is the overall difficulty level of the test. If Part 1 is made easier, it will make the overall test easier, as well. To compensate for that, the committee may need to consider making items in later parts of the first section of the test more difficult to help identify students at a higher level of proficiency.

This year, the Assessments Committee is making another attempt to improve this first part of the test by working within the word discrimination format that currently exists there, as well as an alternative version of that first part which is very similar to it. Seven or eight of the eleven items in the original version of Part 1 will be revised to see if they can be made to

discriminate any better among student levels. The revised items, along with the alternative version of Part 1, will be trial tested with first year International Relations students since they no longer take the FEPT. The outcome of that trial will be analyzed, and then adjustments will be made to version 2.4 of the test for the spring 2013 administration. If the attempt to bolster Part 1 does not succeed this time, it may be time to seriously consider a different testing concept altogether for that part.

Additionally, the Assessments Committee could devote some attention to parts three and four since those parts also have lower discrimination ability than the other parts of the test. Beyond that, continued revisions and replacements can be made of test items that have low discrimination values. There are still a sufficient number of test items with weak discrimination values to keep the committee preoccupied with this direction for some time to come.

IV. Final Thoughts

Time and resources have always been limited for the development of placement tests at the Center for English Language Education. If that were not the case, more serious consideration could be given to commercial alternatives to an in-house test like the FEPT. The expertise and resources that have gone into the creation and development of the FEPT are simply not equal to those of a recognized, professional test making organization. That being said, a limited attempt to improve the existing FEPT can be made from year to year, such as the changes that were made in version 2.3 to produce version 2.4 and the changes that are being worked on for the coming year, without requiring an unreasonable amount of the Assessment Committee's time. Certainly, we can anticipate incremental improvement of the test.

Working within the confines of the budget and resources available to placement testing, it would also be possible to consider implementing, over a longer period of time, a significantly different test concept than currently exists in the FEPT. The current test has no clear connection to the curriculum and materials being used in the Freshman English classes and is, therefore, of limited value. The FEPT seems to be based more on the TOEIC than on the oral communication skills content that makes up the textbooks and materials used in Freshman English. A test that has the consistency and reliability of the current FEPT but which is more closely aligned to the curriculum of the Freshman English program could potentially result not only in making better student placements, but also come much closer to providing some measure of student achievement than the current FEPT.

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, 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.

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. (2012). Modifying Asia University's Freshman English Placement Test. *CELE JOURNAL, 20*, 1-11. ---. (2012). Results of the 2010-11 FEPT and TOEIC tests. *CELE Journal*, *20*, 34- 38.

Messerklinger, J. (2009). Results of the 2008 FEPT. *CELE Journal*, *17*, 49-59.