Review and Analysis of Asia University's 2016 Freshman English Placement Test: The Need for Major or Minor Change?

Kathryn Mabe, Asia University

Abstract

Asia University's principal means of placing first-year students in year-long Freshman English classes is the Freshman English Placement Test (FEPT). This article reviews and analyzes the performance of the most recently administered FEPT. The assessments committee at the university's Center for English Language Education (CELE) is currently reviewing the FEPT as part of its ongoing endeavors to improve various performance aspects of the test. The results of the FEPT were analyzed using the Statistical Package for the Social Sciences (SPSS). Measurements focused upon mean, standard deviation, reliability, item difficulty and item discrimination. The analysis reveals that an alarming percentage of the test does not effectively discriminate between the high and low proficiency candidates. The paper concludes that improvements could be made by tweaking poor performing test items or by implementing more fundamental changes in the kind of test administered.

Background and Introduction to the Study

Incoming students in the Business, Business Hospitality, Economics, Law and Urban Innovation faculties at Asia University are placed in the year-long compulsory Freshman English courses run by the Center for English Language Education (CELE) using the Freshman English Placement Test (FEPT). The current FEPT test (version 2.7) consists of 74 questions comprised of 39 listening questions, 17 vocabulary items, 12 grammar questions and six reading questions. Freshman English classes are divided by faculty and limited to 20 students where possible. In 2016, there were 14 levels for the Faculty of Business, 22 for Law students, 16 for Economics and eight each for Business Hospitality and Urban Innovation, giving a total of 68 classes.

The assessments committee is responsible for the organization of the administration of the FEPT, reviewing its performance and making necessary changes. Much discussion has taken place previously regarding the performance of this test. Key concerns revolve around points widely regarded as critical to testing: reliability and validity (Bachman and Palmer, 1996). Poor scores on both item difficulty and item discrimination have raised questions regarding the reliability of the test (Hull, Brennan & Wells, 2015, Carpenter, 2016). Reliability is defined by Harris (1969) as "the stability of test scores" (p.14). If a test is reliable, the test scores will be consistent between different groups of students of the same level who take the test at different times. Concerns regarding the validity of the FEPT have also been expressed. A test can be said to have validity if the content is made up of the language skills, vocabulary and grammar for example, that the course covers. (Hughes, 2009). As Visiting Faculty Members (VFMs) from the CELE department only work at Asia University for a maximum of five years, one problem is that the original test makers are no longer in employment at the university. Thus, discerning the logic behind the making of the

original FEPT is problematic as we have no records as to what language skills, vocabulary or grammar the test makers intended to assess (Carpenter, 2016).

A further source of concern is the limitation on the length of the FEPT. The reason for this limitation, of 45 minutes, is due to the fact that as of until 2015, it had to be administered again to all the faculties at the end of the year by individual instructors within the time restraints of a 45-minute Freshman English class. The length of a test impacts upon its reliability and a test which is too short may not provide accurate information on the ability of the test takers (Hughes, 2009)

In recent years, attempts have been made to address these points of concern and the FEPT has been updated twice from version 2.5 to the current version 2.7. In addition, an entirely new pilot test was also written. However, as discussed in the 2015 results, neither item difficulty not item discrimination showed any improvement in either version. This paper will focus on the results of the 2016 FEPT and make recommendations based on in-depth analysis of them.

Introduction to the 2016 FEPT Analysis

The FEPT was administered in early April 2016 shortly before the start of the semester and the data was used to place1445 students into their Freshman English classes with CELE instructors.

The FEPT results were analyzed using the Statistical Package for the Social Sciences (SPSS). The results for mean and standard deviation tell us the distance of each candidate's test score from the mean and the larger the standard deviation, the more widely spread the test scores are (Carpenter, 2016). The main analysis focuses on analyzing the reliability of the test in three areas; namely to gauge to what degree the test items are measuring similar characteristics, item difficulty and item discrimination.

The statistical procedure of Cronbach's Alpha was used to measure the first point. Cronbach's Alpha measures how strong the relationship between the test items is and thus, if they are testing the same thing. This measurement gives a value between 0 and 1, with a measurement nearer to 1 indicating that there is a strong connection between the items (Carpenter, 2015).

Regarding item difficulty, once more, an index between 0 and 1 is used. Generally speaking, a range of between .25 and .75 is considered acceptable (Carpenter, 2016). If an item falls below that benchmark, it means a significant number of test takers answered it incorrectly and therefore indicates that it is too difficult. If the score is above .75, it shows that the item was too easy as many students chose the correct answer. In respect to item discrimination, again a scale between 0 to 1 is used, with 1 indicating the item perfectly discriminated between the high and low-performers and 0 meaning it did not discriminate at all (Miller, Linn, & Gronlund, 2009). Ideally, an item should score at least .3 to be said to be separating the best and worst performing test takers (Carpenter, 2016).

FEPT Results and Discussion of Results

Mean and Standard Deviation

The results of the 2016 FEPT mean and standard deviation were similar to those of the previous year (see Figure 1).

Figure 1

FEPT Mean and Standard Deviation

FEPT	Number of Items	Number of Examinees	Mean	Standard Deviation
April 2015	74	1243	39.2	9.4
April 2016	74	1445	39.3	9.7

As we can see from the results, many of the scores on the test are bunched reasonably close to each other. Reflecting previous discussion, it seems that the test population are relatively similar in level (Carpenter, 2016). It is worth noting, however, that the number of candidates increased as did the standard deviation. Asia University opened a new faculty, Urban Innovation, in 2016 and this may explain this aspect of the result.

Reliability

The 2016 FEPT's Cronbach Alpha's value was .84, very similar to last year's total of .83. As noted in previous years, this is considered an acceptable score for a "homemade test" (Harris, 1969, p17). As discussed last year, it can be said with some confidence that the FEPT test has a strong relationship between the items and is therefore testing the same thing. (Carpenter, 2009).

Item Difficulty and Item Discrimination

Turning to the remaining two areas requiring analysis, Figure 2 shows the performance of the FEPT in terms of item difficulty and item discrimination item by

item and Figure 3 shows the overall percentage of each section that did not meet the desired

scores of between .25 and .75 at for item difficulty and at least .3 for item discrimination.

Figure 2

FEPT Item Difficulty and Item discrimination

Item Difficulty						Item Discrimination					
Q1	.849	Q26	.382	Q51	.486	Q1	.272	Q26	.166	Q51	.230
Q2	.724	Q27	.427	Q52	.740	Q2	.328	Q27	.250	Q52	.422
Q3	.541	Q28	.516	Q53	.463	Q3	.307	Q28	.224	Q53	.167
Q4	.491	Q29	.367	Q54	.521	Q4	.340	Q29	.177	Q54	.291
Q5	.617	Q30	.310	Q55	.486	Q5	.194	Q30	.175	Q55	.347
Q6	.792	Q31	.426	Q56	.775	Q6	.122	Q31	.000	Q56	.390
Q7	.318	Q32	.533	Q57	.757	Q7	.114	Q32	.319	Q57	.275
Q8	.932	Q33	.827	Q58	.189	Q8	.221	Q33	.302	Q58	.175
Q9	.335	Q34	.438	Q59	.481	Q9	.195	Q34	.263	Q59	.186
Q10	.685	Q35	.447	Q60	.844	Q10	-.034	Q35	.404	Q60	.334
Q11	.807	Q36	.428	Q61	.537	Q11	.253	Q36	.150	Q61	.253
Q12	.462	Q37	.159	Q62	.322	Q12	.371	Q37	.099	Q62	.331
Q13	.461	Q38	.388	Q63	.328	Q13	.252	Q38	.095	Q63	.167
Q14	.614	Q39	.369	Q64	.665	Q14	.235	Q39	.130	Q64	.235
Q15	.285	Q40	.344	Q65	.396	Q15	.166	Q40	.202	Q65	.175
Q16	.556	Q41	.494	Q66	.538	Q16	.272	Q41	.230	Q66	.207
Q17	.603	Q42	.380	Q67	.538	Q17	.122	Q42	.213	Q67	.234
Q18	.799	Q43	.624	Q68	.803	Q18	.287	Q43	.356	Q68	.270
Q19	.698	Q44	.717	Q69	.688	Q19	.274	Q44	.313	Q69	.329
Q20	.264	Q45	.452	Q70	.658	Q20	.084	Q45	.305	Q70	.318
Q21	.419	Q46	.615	Q71	.425	Q21	.165	Q46	.331	Q71	.411
Q22	.576	Q47	.490	Q72	.372	Q22	.114	Q47	.211	Q72	.312
Q23	.473	Q48	.791	Q73	.316	Q23	.191	Q48	.395	Q73	.223
Q24	.295	Q49	.776	Q74	.463	Q24	.182	Q49	.217	Q74	.364
Q25	.769	Q50	.424			Q25	.102	Q50	.205		

Figure 3

Overall percentage of unsatisfactory performance of each section of FEPT

Section of test	Item difficulty	Item discrimination
Listening: Part 1	37.5%	62.5%
Part 2	14.3%	87.5%
Part 3	20%	100%
Part 4	7%	78.5%
Vocabulary: Part 5	17.6%	53%
Grammar: Part 6 (Fill in the blank)	42.9%	71.4%
Grammar: Part 6 (Find the Mistakes)	20%	100%
Reading: Part 7	0%	16.6%

Item Difficulty

As shown in Figure 3, the worst performing part of the test is *Part 6 Grammar Fill in the Blank*, where students are asked to fill in the blanks from a choice of four possibilities. 28.6 % of the section was too easy for students and 14.3% too difficult, giving the overall total of 42.9%, although it should be noted that this part only consists of seven questions. Looking in detail at the items which were too easy for the test takers, question 60 had an item difficulty value of .84, well over the desired .75 value. This item presents test takers with the following:

"Mariko got a bad grade on her test _____she didn't study."

The options are A) because B) so C) yet and D) nor. In my experience, students at lower levels do often make errors using 'because' and 'so'. However, it may be the case that the other distractors were ineffective and therefore, the test takers settled upon the correct answer with relative ease. It is well-documented that writing multiple-

choice questions is notoriously challenging as the distractors must be both plausible and grammatically possible (Brown & Hudson, 2002, Hughes, 2009,).

Turning now to the item 58 which received an item difficulty value of .19, indicating it was too difficult for the test population. This item comprises of:

"Mr. Kim lived in New York _____two years ago."

The options are A) since B) for C) until D) by. As the use of the key 'until', is often problematic for even high level students in Japan, this could be the reason why it performed badly.

As we can see from Figure 3, *Listening Part 1 Word Discrimination* also performed extremely poorly with 37.5 % of the test too easy for the students. In this section, test takers are asked to identify from a choice of five words the one which they hear at the end of the sentence. As previously mentioned, one problem for the current assessments committee is the lack of information as to the theory behind the making of the test. Although we can speculate from the Cronbach's Alpha rating that the items on the FEPT are closely related, a clear outline of the methodology underpinning the choice of these items is unavailable (Carpenter, 2016). However, it is much documented in EFL literature that identification of phonemes has been said to play an important part in listening skills (Field, 1998). Hughes describes the value of testing lower level listening skills, such as discriminating between vowel phonemes and consonant phonemes, in placement tests (2009). Therefore, the inclusion of item discrimination seems reasonable in my opinion. As to why the test takers found it too easy, the answer may lie in the distractors given. For example, in item eight, which received the worst score within this section of 0.93, students are asked to identify the word they hear at the end of the following sentence:

"What did you do then?",

with the alternative choices being ‘they’, ‘bay’, ‘den’ and ‘men’. Although distinguishing between the phonemes /ð/ and /d /can be problematic for Japanese learners, this relatively familiar question pattern of using a wh- question word followed by the past tense auxiliary may have guided a large number of test takers to the correct answer. In addition, the phonemes are all at the end of the utterance. This eases the burden on the test takers as they only have to listen for the last word. These results suggest that asking students to only do this may be making this item too easy for the test population.

Turning to the better performing parts of the test, *Part 7 Reading: Sentence Comprehension*, was notably the strongest, with all of the items receiving a standard score of between .25 and .75. As an example, item 69 reads:

“Since I don’t have enough money, I may not go to Hokkaido”.

Students are then asked to choose the option below which has the same meaning;

A) I will go to Hokkaido, B) I must go to Hokkaido, C) I probably won’t go to Hokkaido and D) I will go to Hokkaido in the future.

From a pedagogical point of view, the validity of this as a test of reading skills could be called into question if we consider how we read in an authentic situation, for instance reading with a purpose and a given context; neither of which are present here. Moreover, a knowledge of modals expressing future possibility is needed to successfully answer this question; therefore, it is also testing grammatical prowess. It should be noted though that the pilot test appears to have attempted to redress this issue by extending the reading tasks and giving clear contexts but in fact the results for item difficulty and discrimination on the pilot test were even poorer than the ones for the current FEPT (Carpenter, 2016). Therefore, as this is the best performing section of the test, it seems misguided to criticize it too deeply. As Hughes comments, grammar and

vocabulary “are both tested indirectly in every reading test” (2009, p.138). Therefore, there seems to be some pedagogical basis for this part of the test.

Item Discrimination

The best performing section of the test was once more *Part 7 Reading* with 83.4% of the items discriminating between the high and low proficiency students. Thus, as discussed previously, this part of the test can be deemed effective and reliable as a placement tool on the basis of the data analysis in this report.

However, as Figure 2 shows, much of the test does not effectively separate the higher and lower proficiency students. Calculated as an overall percentage, 71.6% of the test failed on this point. This is deeply concerning as the placement test’s function, by very definition, is to separate students accurately into multiple levels. Looking at the sections in detail, the two worst performing were *Listening Part 3 Question and Answer* and *Grammar Part 6 Find the Mistakes*, with none of the questions achieving the desired .3 value in either section. *Listening Part 3*, item 20, which obtained the lowest value in this section of only .08, begins with the prompt:

“How can I get in touch with you?”

The students then hear A) Yes, you can B) before 6pm and C) Call me. Despite having to correctly identify the question word ‘how’, as well as understand the phrasal verb ‘get in touch with’, this item achieved an acceptable value on item difficulty, albeit a lower end one of .26 (see Figure 2). As to why it did not effectively differentiate between high and low proficiency test takers, this is beyond the scope of this analysis. It may be of relevance that this is the section of the listening test where test takers have no visual clues to help them. In parts 1 and 4, the students are able to see the possible answers and in Part 2, there is a photo for each item. It could be that the absence of any visual hints meant that students were not confident of the answers and guessed. This could explain why a number of students in low level classes answered correctly and

those in high levels did not. As Hughes comments, "Guessing may have a considerable but unknowable effect on test scores…The trouble is we can never know what part of any particular individual's score has come about through guessing" (2009, p.76).

Turning to *Grammar Part 6 Find the Mistakes*, test takers are asked to choose which of four underlined words in a sentence is grammatically incorrect. The worst performing item was 65, with a score of .18:

"Midori rarely reads newspapers and ever hardly listens to the news."

It is difficult to speculate as to why this item failed to successfully divide the test takers. It is possible that including two nouns as alternatives (newspapers, the news) was too easy for the students, leaving them just two reasonable options to choose from, or as mentioned earlier, guess from.

The lowest individual score was for item 10, in *Listening Part 2 Picture Identification,* with a value of minus .03. The picture is of a train next to a platform in the station with the rubric:

"A) The doors are open B) They're staying on the platform C) The plane is waiting D) Please use this stationary."

In the 2015 FEPT administration, this question also received a very low discrimination value, therefore consistently being a problematic item, despite revision attempts (Brennan, 2015, p.28).

Other extremely poor performing sections were *Listening Part 2 Picture Identification* and *Listening Part 4 Dialogs*. In the former, 87.5% did not effectively separate the high and low proficiency students and in the latter, 78.5% did not.

Not only does a large part of this test fail to discriminate between the high and low proficiency students, a further alarming feature is that multiple items score poorly both on item difficulty and item discrimination. For example, item 8, despite being too easy for the test takers as described in the previous section, did not discriminate

between the high and low-performers, receiving a score of .2. This is particularly perplexing as it seems a reasonable assumption that if a question were too easy, the higher level students would be able answer the question correctly. Nor was this an isolated instance. In fact, a total of eleven questions (Items 1,6, 8,11, 18, 25, 37, 49, 57, 58 and 68) fell short of a satisfactory score in both item difficulty and item difference. Two of the questions were too difficult (Items 37 and 58) yet again counter-intuitively it appears that this does not mean the lowest level students were unable to answer them. It should be noted as this point that at least anecdotally, according to CELE instructors, there is a notable difference between the students placed in high and low level classes. However, there are also many classes in the middle of the level ability, therefore the question remains as to how effectively they are being separated (Carpenter, 2016). In addition, the evidence of the above data is compelling as it is not only one isolated case. This leads to the inevitable question as to why this phenomenon is occurring.

Aside from the aforementioned possibility of multiple examinee guessing skewing the results, one possible cause could be that the FEPT relies only upon multiple choice items. This raises the issue of whether a test is designed to assess candidates' productive or receptive skills. Multiple choice questions only test recognition knowledge. As Hughes comments: "A multiple choice grammar test score, for example, may be a poor indicator of someone's ability to use grammatical structures" (2009, p.76). It is possible that including only receptive skill testing items is affecting the results in the above puzzling way. Including question types such as gap fill or short answer would mean students would be required to actively produce language and might well improve results on item discrimination. However, as the FEPT results are scored through an automatic scantron format, the correct answers must be limited to what are able to be read and checked automatically. Although in the past, there was also an oral component to the FEPT, this was removed in 2009. It is the view of the assessments

committee that it would not be beneficial to return to this element of the FEPT as the practical considerations of training CELE instructors to be able to effectively interview hundreds of new intake students on the same day using an agreed criterion is unfeasible.

Recommendations and conclusions

In terms of short term recommendations, revision of *Grammar Part 6 Fill in the blanks* could well be worthwhile as this received the worst score for item difficulty. This is feasible time-wise as this would not require any re-recording of the audio. Items such as the aforementioned 58 and 60 could be rewritten with relative ease to adjust the level of difficulty. In addition, *Grammar Part 6 Find the Mistakes*, as previously discussed in the item discrimination section, had no items which successfully separated the high and low proficiency students. Rewriting some of these items may help increase item discrimination scores. Again, no re-recording of audio would be necessary.

Turning to mid-term possible revisions and the poor performance of *Listening Part One*, it would be interesting to see whether inclusion of mid-sentence word identification improved the item difficulty rating of this section. Research has shown this to be an area of difficulty for English learners, for example differentiating between items such as ‘want’ and ‘won’t’ in mid-sentence speech (Field, 2003, p.325). Other sections of the FEPT listening also received extremely poor item discrimination scores and thus need further scrutiny. Alterations would require re-recording of the audio however, therefore it is not feasible in the short-term as the assessments committee is limited in time available for such a project. Additionally, as many of these items have already been changed in the past, it may be the solution lies to look at other options rather than rewriting.

Longer term changes may be possible due to changes agreed in 2016 by the university administration. To date, the assessments committee has also been responsible for coordinating the administration of the End of Year Placement Test (EYPT), which is the identical test. This was because it was believed that faculties used the results to place students into Sophomore English classes. It has now been agreed with university administration that this is unnecessary for all but one faculty. This opens up the possibility of making the test longer than 45 minutes and thus possibly improving the reliability. A longer test could, for instance, include more items in the Reading section, which received the best scores out of all the sections for both item difficulty and item discrimination. I would also recommend changing the name of this section to something like Multi-Skills Section to reflect that it is not merely testing reading skills in isolation but also aspects of language such as grammar.

As previously discussed, it seems plausible that some of the more inexplicable poor performance of the FEPT could be caused by the students guessing. One solution could be actively involve guessing as part of the test. The ability to guess meaning of unfamiliar vocabulary from context is widely regarded as a beneficial tool in language learning (Griffiths, 2015). Test takers could be presented with vocabulary items specifically designed to be unfamiliar and asked to guess their meaning from the context of the rest of the sentence.

In addition, the usage of only one kind of test question (i.e. multiple choice) is also possibly a factor in the overall poor performance of the FEPT. The solution to this would require a completely different test format and scoring system and thus, discussion of this is beyond the scope of this paper but nevertheless is a long-term consideration.

References

Bachman, L.F. & Palmer, A.S. (1996). *Designing and Developing Useful Language Tests.* New York: Oxford University Press.

Brown, D.B., & Hudson, T. (2002). *Criterion-referenced Language Testing*. Cambridge University Press.

Carpenter, J. (2016). Past, Present and Future Placement testing Practices at CELE: A view from 2015. *CELE Journal, 24*, 52-77.

Field, J. (1998). Skills and strategies: towards a new methodology for Listening. *ELT Journal, 52/2,*110-118. Oxford University Press.

Field, J. (2003). Promoting perception: lexical segmentation in L2 listening. *ELT Journal 57/4*, 325-333. Oxford University Press.

Griffiths, C. (2015). What have we learnt from 'good language learners'? *ELT Journal 69/4*, 425-433.

Harris, D.P. (1969). *Testing English as a second language.* MacGraw-Hill, Inc.

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

Hull, J., & Wells, L. (2015). Review and Analysis of Asia University's freshman English placement test: Transition from version 2.5 to version 2.6. *CELE Journal, 23,* 21-49.

Miller, M., Linn, R., & Gronlund, N. (2009). *Measurement and assessment in teaching.* (10th ed.). Upper Saddle River, NJ: Pearson Education, Inc.