An Analysis and Review of the 2017 Freshman English Placement Test at Asia University

Daniel Bates, Asia University

Abstract

This article reviews the results of the Freshman English Placement Test (FEPT) administered in April 2017 to place incoming first-year students into Freshman English Classes (FE) at Asia University. Using the Statistical Package for Social Sciences (SPSS) to analyze a variety of measurements including the mean score and standard deviation, reliability, item difficulty and item discrimination, this analysis shows that while the reliability of the overall test was sound, there were significant deficiencies in the difficulty level of numerous individual items, as well as a large proportion of the test which did not discriminate between the best and worst performing test takers. This paper concludes by offering a number of possible options open to the Assessments Committee at Asia University's Center for English Language Education (CELE) in order to improve the overall effectiveness of the FEPT, ranging from replacing individual items to speculating on potential replacements for the current FEPT.

Introduction to the 2017 FEPT Analysis

The Freshman English Placement Test (FEPT) is administered at the beginning of each academic year at Asia University to place first-year students in compulsory, year-long Freshman English classes. This test has been in place for a number of years and has been regularly revised by the Assessments Committee in attempts to make the test more reliable and valid. As there were no changes made to the FEPT between 2016 and 2017, this paper will follow a similar structure to the 2016 paper, focusing on a detailed analysis of the reasons behind the contrast between particular items that performed well or poorly on the test, as well as making comparisons with the 2016 results where appropriate.

The FEPT was administered before the start of the 2017 academic year to place 1415 students into Freshman English classes based on their English language abilities. The FEPT results were analyzed using the Statistical Package for the Social Sciences (SPSS). The main purpose of this analysis is in testing three measurements: reliability (whether the items are testing for the same thing); item difficulty (whether the questions are too easy or too difficult for the test takers); and item discrimination (whether the items are discriminating between the best and worst performing test takers). The results of each section will be analyzed individually with comparisons made to the results from the 2016 FEPT.

2017 FEPT Results, Comparisons and Analysis

Mean and Standard Deviation

The results for the standard deviation show us how far a candidate's score was from the overall mean score and the larger the standard deviation, the more widely spread the test scores are (Carpenter 2016).

Figure 1

FEPT Mean and Standard Deviation

FEPT	Number of Items	Number of Examinees	Mean	Standard Deviation
2017	**74**	**1415**	**40.4**	**10.1**
2016	**74**	**1445**	**39.3**	**9.7**

The results show that there has been a small increase in both mean and standard deviation, both of which also rose slightly from 2015 to 2016 (Mabe, 2017, p. 5). The increase in the mean score indicates that there may be more higher-ability students taking this test compared to previous years. This corresponds with the rise in standard deviation which also suggests a wider range in the abilities of the test takers compared to the results from the past two years. While these increases are relatively small and would have little impact on the composition of the FE classes themselves, an increase in the ability of the test takers may negatively affect the item difficulty statistics, especially with a number of items already performing poorly in the 2016 FEPT as they were deemed too easy by the 2016 analysis (Mabe, 2017).

Reliability

A test that is reliable will show test scores to be consistent between separate groups of test takers at the same level, who sit the test at different times (Mabe, 2017, p. 2). The reliability of a test is determined by using Cronbach's Alpha to determine to what degree the test items are measuring similar characteristics. A value is given between 0 and 1, with higher values indicating a stronger relationship between the items and, as noted in previous years' analysis, a score of above .80 indicates an acceptable level of reliability (Carpenter, 2016 p. 61). The Cronbach's Alpha value for the 2017 FEPT was .85 which shows a slight increase on 2016 (.84) and 2015 (.83) (Mabe, 2017 p. 5). In line with results from recent FEPT tests, we can again surmise that there is a strong connection between the items on the FEPT and that the items are testing for the same thing.

Item Difficulty

Next, the analysis will focus on the performance of the 2017 FEPT in item difficulty and compare the results from each of the seven parts of the test with the results from 2016. Item difficulty is measured by SPSS and given a score between 0 and 1. In this case, a value between 0.25 and 0.75 is considered acceptable (Carpenter, 2016, p. 62). An item scoring above 0.75 indicates the question is too easy as a significant number of test takers answered correctly, while a score of under 0.25 suggests the item is too difficult as too many answers were incorrect. Figure 2 shows the overall percentage of each section on the 2017 FEPT that did not fall into the acceptable range of between .25 and .75 in contrast to the 2016 results.

Figure 2

Overall percentage of unsatisfactory performance in Item Difficulty for each section of FEPT

	2017	**2016**
Listening		
Part 1	**50%**	**37.5%**
Part 2	**42.9%**	**14.3%**
Part 3	**50%**	**20%**
Part 4	**14.3%**	**7%**
Vocabulary: Part 5	**31.3%**	**17.6%**
Grammar: Part 6 A (Fill in the blank)	**28.6%**	**42.9%**
Grammar: Part 6 B (Find the Mistakes)	**20%**	**20%**
Reading: Part 7	**0%**	**0%**

As shown in Figure 2, the worst performing sections in 2017 were Listening Part One: Word Discrimination and Listening Part Three: Question and Answer. Both sections performed exceedingly poorly with half of the total questions failing to satisfactorily challenge the students. Five of the seven sections saw a notable increase in unsatisfactory performance in item difficulty from 2016, with the majority of those unsatisfactory scores (85%) being above 0.75 (see Appendix 1), and therefore being too easy for the test takers. As already noted, this links to the increase in mean and standard deviation which indicate an increase in the number of higher-ability test takers than in previous years.

In section one, 50% of the eight questions were too easy for the test takers. In this section, test takers are asked to identify the correct word from a choice of five to complete a sentence. As with 2016, items 1 and 8 performed particularly poorly with respective scores of .86 and .94. In item one, students are asked to identify the word they hear at the end of the following sentence: "The team has everyone's support" with the alternative choices being 'spirit', 'port', 'sprout', and 'sport'. This question may be testing for syllable types and consonant clusters, an area of some difficulty for Japanese learners of English (Ohata, 2004, p. 35) yet the context of the sentence renders the alternative answers immediately obsolete. With little else being tested in this item, most students were able to correctly identify 'support' as the correct answer. In comparison, question 4, which received a score of .54, asked test takers to identify the following word in bold; 'She is a secretary, isn't **she**?' with the alternative choices being 'agency', 'sea', 'C' and 'see'. Question 4 performed well on item discrimination, perhaps because Japanese

learners of English have difficulties distinguishing between the / ʃ / and /s/ phonemes (Ohata, 2004, p. 13). In addition, question tags can provide some difficulties for lower level learners and listing homophones requires the students to have an awareness of spelling. The questions in this section all follow the same pattern—a simple sentence is spoken and the test takers have to identify the final word in the utterance. Evidently, the year-on-year results indicate that this structure is too easy for the test takers. By having a range of words placed throughout the sentence, suprasegmental features such as linking words, elision and assimilation could be tested alongside segmentals, thereby increasing the difficulty level for the test takers. Part three also had 50% of the 10 questions being unsuitable for the test, with three questions proving too easy, and two being too difficult. In this part, test takers hear a question followed by three possible answers from which they choose the most appropriate response. Item 18 scored a value of .817, making it too easy for those taking this test. It began with the prompt, "Where are you going?" to which the students are given the options of: a) tomorrow b) to class c) Tuesday. Evidently, given that this simple interrogative 'where' question is commonly taught at low levels and the fact that the wrong answers both give a time, if the student correctly identifies the 'Where' at the beginning of the question, they are likely to get the correct answer. On the other hand, item 24 has a value of .286 and was at the low end of acceptability, with it being too difficult for the majority of test takers. Students first heard "Is that our English teacher?" with significant emphasis on 'that' and with a surprised intonation. This was followed by the options: a) Yes, in an hour b) I have English next hour c) It certainly looks like him. The question requires students to understand sentence stress and intonation as well as grammatical meaning (the use of 'that' to describe a person). Stress for emphasis and showing emotion through intonation is unlikely to be taught to lower-level learners and a question like this may prove beyond the capability of most incoming Freshman English students at Asia University. This question also scored unsatisfactorily on item discrimination, meaning it did not discriminate between high and low-level test takers.

As with 2016, Part 7 Reading: Sentence Comprehension was again the strongest with all the items falling between the acceptable score of .25 to .75. Item 74 fell closest to the middle with a value of .476. It reads: "The professor had already given the homework assignment due Monday when he remembered that it was a holiday." Students are then asked to choose the option which refers to 'it,' with the options being: a) The professor b) The homework c) Monday

d) The assignment.

Unfortunately, with the original test makers no longer working at Asia University and with no definitive written records, we can only make educated guesses as to what the test makers intended to assess with these sections (Carpenter, 2016, p. 60). However, it is clear that item difficulty is continuing to perform poorly, particularly the first three listening parts which are performing worse year on year. If some listening sections are to be replaced, Part 1: Word Discrimination and Part 3: Question and Answer, would be the logical places to begin as half of the questions in these sections are either too easy or too difficult for this placement test. If individual items are to be replaced, those items scoring above .75 should be replaced first.

Item Discrimination

Item discrimination shows the separation between the best and worst performing test takers. Here, a scale of 0 to 1 is again used with a higher value indicating that the item has discriminated between high- and low-performing test takers. An item with a score above .300 is said to be separating the high and low performers on the test. Figure 3 shows the overall percentage of unsatisfactory performance (namely, a score of below .300) for item discrimination for each part of the FEPT.

Figure 3
Overall percentage of unsatisfactory performance of each part of the FEPT

	2017	**2016**
Listening		
Part 1	**75%**	**62.5%**
Part 2	**100%**	**87.5%**
Part 3	**80%**	**100%**
Part 4	**71.4%**	**78.5%**
Vocabulary: Part 5	**56.3%**	**53%**
Grammar: Part 6 A (Fill in the blank)	**57.1%**	**71.4%**
Grammar: Part 6 B (Find the Mistakes)	**80%**	**100%**
Reading: Part 7	**16.6%**	**16.6%**

The best performing part of the test was Reading with 83.4% of the items discriminating between the high and low proficiency students. As with 2016, this part of the test remains the most effective and reliable in terms of placing students into classes based on language

proficiency. As Figure 3 shows, 66.2% of the 2017 FEPT was unsatisfactory in separating the students based on language proficiency. While this is slightly less than the 2016 results (71.6%) (Mabe, 2017, p. 11), over half of each of the remaining six parts did not discriminate effectively between high- and low-level test takers. In 2017, Listening Parts One and Two saw the most alarming increase in unsatisfactory performance in item discrimination with the entire section of Part Two performing unsatisfactorily.

The following items scored poorly in both item difficulty and item discrimination (see Appendix 1 for data). Questions 1, 2, 6, 8, 11, 49 and 68 scored above .75 on item difficulty and so can be considered too easy for this placement test, while questions 20, 24, 37 and 58 scored around or below .25 and can be deemed too difficult. While a number of other items also scored poorly on item difficulty, the above-mentioned items also had an unacceptable score on item discrimination and are arguably not serving any significant purpose on this test. If some minor alterations are planned for the 2018 FEPT, replacing these items would be an ideal place to start.

Suggestions for Further Study

Anecdotally, many Visiting Faculty Members have expressed concerns that the range in the abilities of students in their FE classes is too wide, especially when attempting activities such as presentations or communicative activities that go beyond the scope of the textbook. A study focusing on both VFMs' perception of their students' varying abilities and a study looking at FE students' perceptions of their own abilities in comparison to their peers may give some context to this statistical analysis and the conclusions drawn from it.

Conclusions and Recommendations

With no changes made to the FEPT between the 2016 and 2017 tests, and with broadly similar results for reliability in both years, the same problems remain. At present, there are arguably three options available to the Assessments Committee going forward. The first option is to replace the test entirely. With increasingly affordable online-based standardized placement testing (such as VERSANT), it could be feasible to replace the FEPT. These online tests have many advantages over the current FEPT in the sense that they address the concerns about reliability and offer a more valid test with a speaking element. The main hurdle to implementing

a test of this kind is likely the costs rather than an educational concern, but further discussion is beyond the scope of this paper.

The second option would be a large-scale revision of the current FEPT. However, a new FEPT format was piloted in 2015 and the respective analysis of it revealed that it had similar strengths and weaknesses to the current FEPT format (Carpenter, 2016, p. 71). This would be a large project to undertake with past experience showing that there may be little to no improvement in a new, bespoke placement test.

Finally, and most realistically, minor alterations to individual test items that have continued to perform poorly could be revised and replaced. It should be remembered that the Cronbach's Alpha value has remained consistently high, thus the test as a whole is achieving a suitable level of reliability. However, as seen in Figures 2 and 3, individual items, and in some cases whole sections, are not performing well in other areas. My recommendation is that the Assessments Committee focuses on replacing the items, which were identified in the item discrimination section of this paper, and have consistently performed unsatisfactorily in both item difficulty and item discrimination. By replacing these items before the 2018 FEPT, next year's analysis can focus more closely on how these changes affected the test and whether further alterations of a similar nature could be made or if one of the other two more drastic options should be put in to place.

References

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

Mabe, K. (2017). Review and Analysis of Asia University's 2016 Freshman English Placement Test: The Need for Major or Minor Change? *CELE Journal, 25,*1-16.

Ohata, K. (2004). Phonological Differences between Japanese and English: Several Potentially Problematic. *Language Learning*, *22*, 29-41.

Appendix 1: Item difficulty results by item

Item number	Item Difficulty								
Q1	.861	Q16	.567	Q31	.459	Q46	.637	Q61	.589
Q2	.769	Q17	.613	Q32	.574	Q47	.495	Q62	.360
Q3	.567	Q18	.817	Q33	.832	Q48	.788	Q63	.360
Q4	.538	Q19	.719	Q34	.454	Q49	.791	Q64	.674
Q5	.564	Q20	.253	Q35	.455	Q50	.427	Q65	.407
Q6	.793	Q21	.423	Q36	.448	Q51	.472	Q66	.539
Q7	.355	Q22	.552	Q37	.165	Q52	.749	Q67	.564
Q8	.940	Q23	.517	Q38	.379	Q53	.510	Q68	.805
Q9	.315	Q24	.286	Q39	.374	Q54	.560	Q69	.680
Q10	.715	Q25	.741	Q40	.389	Q55	.514	Q70	.674
Q11	.790	Q26	.361	Q41	.467	Q56	.782	Q71	.448
Q12	.709	Q27	.436	Q42	.408	Q57	.754	Q72	.406
Q13	.466	Q28	.495	Q43	.629	Q58	.201	Q73	.341
Q14	.632	Q29	.400	Q44	.726	Q59	.493	Q74	.476
Q15	.320	Q30	.307	Q45	.435	Q60	.856		

Appendix 2: Item discrimination results by item

Item number	Item Difficulty								
Q1	.286	Q16	.303	Q31	.068	Q46	.291	Q61	.244
Q2	.299	Q17	.133	Q32	.320	Q47	.219	Q62	.351
Q3	.314	Q18	.311	Q33	.301	Q48	.434	Q63	.095
Q4	.413	Q19	.268	Q34	.308	Q49	.288	Q64	.203
Q5	.123	Q20	.119	Q35	.350	Q50	.209	Q65	.131
Q6	.129	Q21	.164	Q36	.147	Q51	.299	Q66	.204
Q7	.090	Q22	.181	Q37	.073	Q52	.484	Q67	.329
Q8	.223	Q23	.179	Q38	.226	Q53	.236	Q68	.246
Q9	.142	Q24	.167	Q39	.156	Q54	.382	Q69	.337
Q10	.042	Q25	.109	Q40	.249	Q55	.395	Q70	.370
Q11	.275	Q26	.079	Q41	.191	Q56	.351	Q71	.441
Q12	.227	Q27	.290	Q42	.211	Q57	.355	Q72	.337
Q13	.298	Q28	.212	Q43	.374	Q58	.191	Q73	.237
Q14	.268	Q29	.211	Q44	.367	Q59	.260	Q74	.376
Q15	.210	Q30	.118	Q45	.369	Q60	.323		