The Freshman English Placement Test at Asia University 2018: Still Viable?

Kim Pollard, Asia University

Abstract

This article evaluates the viability of the Freshman English Placement Test (FEPT) taken by freshman students at Asia University in Tokyo. It examines the reliability of the test, item difficulty and discrimination using statistical analysis. It finds that the test is still a useful tool for separating students into classes and highlights factors that should be considered in future rewrites. Smaller short-term test question modifications are also proposed.

Introduction to the 2018 FEPT

The Freshman English Placement Test (FEPT) was used to place 1311 Freshman English (FE) students from five faculties into compulsory English classes at Asia University in Tokyo. This paper will examine the viability of the test in 2018 by evaluating reliability, test item difficulty and discrimination. These results will then be compared to those from 2016 to 2017 to determine if the test is still a useful tool for placing students or requires significant alteration. Factors to be considered in future rewrites and current test modifications will be proposed.

Administered at the beginning and end of the FE academic year, the test aims to place students into classes based on their English language abilities. The test has 74 multiple choice questions and comprises of Listening, Vocabulary, Grammar and Reading sections. Even though the FE course is communicative and based on speaking skills, there is no speaking component to the test due to a lack of resources. The test is administered by teachers from the Centre for English Language Education (CELE) at Asia University. It is computer marked using a scantron format and the results are analysed using SPSS.

Comparisons and Analysis of FEPT Results 2016-2018

Mean and Standard Deviation

By examining the standard deviation for each year, the variation in students' scores can be measured. A high standard deviation indicates that students' scores are spread out from the mean, a low standard deviation that data points are close to the mean score. A placement test where all the scores are close to the mean has failed to differentiate between students effectively (Alderson, Clapham & Wall, 1995).

The standard deviation and mean scores have increased slightly year on year (Bates 2018; Mabe, 2017; Carpenter, 2016). As noted by Bates (2018) and Mabe (2017), this has implications for analysing test difficulty and assessing whether parts of the test are performing as well as they should. This standard deviation of 10.4 in 2018 shows that scores are spread further out from the mean year on year but that the test population are still similar in ability. Anecdotal evidence suggests that most students in FE are Beginner, Elementary and Pre-intermediate on the Common European Framework of Reference (CEFR) scale.

Table 1

FEPT Mean and Standard Deviation

FEPT Year	Number of Items	No. Examinees	Mean	Std. Deviation
2018	74	1311	41.3	10.4
2017	74	1415	40.4	10.1
2016	74	1445	39.3	9.7

Test Reliability

Cronbach's alpha was applied to the FEPT results to determine whether test items are consistently testing for the same thing. This index establishes the reliability of the test and whether the test performs consistently from year to year with different students. The higher the coefficient, the likelier it is that the test items are testing the same concept. If the theoretical value of alpha varies from 0-1. A score of 0.7 or higher, it is deemed acceptable for determining reliability (George & Mallery, 2003). A reliability analysis was carried out on the 74 test items. Cronbach's alpha showed internal consistency of test items to be good, α = 0.86, a slight improvement on previous years. The test can therefore be deemed to be reliable in testing for the same thing amongst different students. Although this is a good score for a homemade test, it should be noted that the alpha does not assess test validity (Hull 2012, p. 4).

Item Difficulty

The ability of the FEPT to discriminate between students' abilities from year to year can also be assessed by analysing test item difficulty scores. A mean score of below 0.3 indicates that a test item was too difficult for a significant number of test takers as most got the question wrong. A score above 0.7 shows that the test item was too easy as most students answered the question correctly, including low scoring students. Table 2 below compares the percentage of unsatisfactory performance in item difficulty for the last three years.

Table 2

	2018 (%)	2017 (%)	2016 (%)
Listening			
Part 1	50	50	37.5
Part 2	14.3	42.9	14.3
Part 3	40	50	20
Part 4	14.29	14.3	7
Vocabulary: Part 5	47.06	31.3	17.6
Grammar: Part 6 A (Gap fill)	42.9	28.6	42.9
Grammar: Part 6 B (Find the mistakes)	40	20	20
Reading: Part 7	33.3	0	0

Table 3 shows the percentage of questions deemed too easy and too difficult in each section of the 2018 test.

Table 3	2018 too easy %	2018 too difficult %
Listening		
Part 1	50	0
Part 2	14.3	0
Part 3	30	10
Part 4	7.1	7.1
Vocabulary: Part 5	29.4	20.6
Grammar: Part 6 A (Gap fill)	28.6	14.29
Grammar: Part 6 B (Find the mistakes)	40.0	0
Reading: Part 7	33.3	0

The number of questions that are too easy for most test takers seems substantial in most sections. This may, in future, cause problems in differentiating between the abilities of students.

Part 1 Listening is deemed particularly problematic each year as most of the test takers got the questions in half of this section correct. Easy questions are useful in that they can provide a good lead in for students and help put them at ease (Heaton 1989, p 179). However, questions should get progressively more difficult in order to discriminate between students more effectively (Carpenter, 2016). Just one or two easy questions should be sufficient to ease students in.

As in the previous two years, Grammar parts A and B are also consistently too easy for students (Bates, 2018; Mabe, 2017; Carpenter, 2016). In broad terms, this could be because students may come from a predominantly Grammar Translation pedagogy at Junior (JHS) and Senior High School (SHS). In general, greater emphasis is placed on grammatical forms and accuracy over fluency in JHS and SHS examinations and tuition (McNamara & Rover, 2006; Sato, 2002). Also, unlike previous years, the reading questions were found to be too easy by most test takers with 33% of questions being unsatisfactory in this regard. As this has only been the case for this year, it would be prudent to see if this trend continues before making changes to this section. Although there were some difficult items which fail to discriminate effectively amongst most students, they are still useful for distinguishing good and very good students (Heaton 1989, p. 179).

Item Discrimination

Item discrimination is another index that helps to determine whether a test is functioning effectively. This index shows "the extent to which the item discriminates between the testees, separating the more able from the less able" (Heaton, 1989, p. 179). A discrimination index of above 0.3 confirms that the test item is at the correct level of difficulty and discriminates between higher level and lower level students well. Conversely, an index below 0.3 means that the item discriminates poorly. A minus score indicates that the lower level students answered the question correctly, but higher level students did not. The table below shows the percentage of scores under 0.3 for each section of the FEPT over the last three years.

Unfortunately, only 51.35% of the test discriminates between students effectively. However, this is an improvement on the 74% achieved in 2015 (Carpenter, 2016, p. 58). It seems then that the test is discriminating between students more effectively than in previous years. The vocabulary and reading sections of the test continue to perform relatively well in terms of item discrimination and item difficulty combined.

Table 4

Item discrimination % of questions that performed poorly	2018	2017	2016
Listening			
Part 1	50	75	62.5
Part 2	42.9	100	87.5
Part 3	70	80	100
Part 4	57.1	71.4	78.5
Vocabulary: Part 5	35.3	56.3	53
Grammar: Part 6 A (Gap fill)	57.1	57.1	71.4
Grammar: Part 6 B (Find the mistakes)	80	80	100
Reading: Part 7	33.3	16.6	16.6

Table 5

Sections that performed poorly in terms of difficulty and discrimination (%).

	Difficulty	Discrimination
Listening		
Part 1	50	50
Part 2	14.3	42.9
Part 3	40	70
Part 4	14.29	57.1
Vocabulary: Part 5	47.06	35.3
Grammar: Part 6 A (Gap fill)	42.9	57.1
Grammar: Part 6 B (Find the mistakes)	40	80
Reading: Part 7	33.3	33.3

Problematic Questions

To assess the performance of test items, it is important to examine their level of difficulty and ability to discriminate between higher and lower level students (Heaton, 1989). In terms of improving the test as it stands, questions that fail in terms of difficulty and differentiation year on year should be modified or removed. Doing so may also improve Cronbach Alpha. Based on these criteria, the committee could look at replacing or removing questions 6, 8, 20, 37, 53, 54, 55 and 58 as they have performed poorly in both regards each year. If replacing these questions, it is important to ask why the students failed to answer the question correctly so that a suitable alternative can be found (Heaton 1989, p. 182).

As Mabe (2017) has stated, the questions in the Listening section, are too easy because students only have to listen for the final word and can guess the word from the context of the sentence. An example is Question 6: Q6. What did you do _______? a) they b) bay c) den d) then e) men. It may also be that some of the students are guessing the correct answer to some questions because the distractors are unfamiliar to them (Mabe, 2017). This test item does not, therefore, test for phonemes as it was designed to. Mabe (2017) suggests that more difficult words such as “won't” and “want” in the middle of the phrase would challenge and differentiate students more.

Another problem lies in the vocabulary section of the test where low frequency, high level words are used in some questions. Most students are low level and so the test should be trying to distinguish between students of that level. Questions 53 and 55 (see below), for example, asks students to find the opposite of a word from a choice of four other low frequency, high level words. Prefixes are typically problematic for low level students. Similarly question 54 uses phrasal verbs that lower level students probably will not have encountered yet (Mabe, 2017).

Problematic Question Examples:

Q. 53. Find the opposite of practical a) accidental b) impractical c) intentional d) imperial.

Q 54. Find the opposite of continues a) break off b) carry on c) start up d) start off.

Test Validity

As stated before by Carpenter (2016, p. 67), Mabe (2017) and Bates (2018) there are several key problems with the validity of the FEPT. One such problem is the lack of diversity in the language structures and lexis tested. There are many questions that cover the same language point, for example, prepositions of place and directions. Consequently, the full language ability of the students may not be tested effectively. Also, where possible, both productive and receptive skills should be tested (Mabe 2017, p. 13). In the FEPT, only the latter is being tested.

Secondly, the FE course is communicative and predominantly focuses on speaking skills. However, most test items do not test for language used in a communicative way (Mabe, 2016). Given that students may come from JHS and SHS classrooms that are non-communicative in nature, this is understandable. However, this may not be so in the future with changes being implemented by the Ministry of Education (e.g. LEEP and the introduction of the Japan CERF) (Nagata, 1995). In the future students will have to complete an oral test as part of their entrance examinations. Ideally test questions should give authentic situations where students are listening or reading for a given purpose and in a specific real word context (Mabe, 2017). Test items could be constructed using the multiple choice automatic scantron format and still fulfil this criterion. This would, however, require another major rewrite of the test.

Finally, the test does not appear to take fully into consideration the FE curriculum and language taught at each textbook level (Brown, 2002). This means that improvement in students' ability cannot be measured even though the test is taken again at the end of the year. Unfortunately, there is no indication of the reasoning behind the test design of each component as the test creators have left CELE (Bates, 2018; Mabe, 2017; Carpenter, 2016). Given CELE's current resources it is not possible to do a speaking level check for over 1300 students. It should be possible, however, to collaborate with the curriculum development committee to fit the test better to the curriculum (Carpenter, 2016).

Conclusion and Recommendations

The test as it stands still does a reasonable job of differentiating between students but has some room for improvement. In the short term, questions that have repeatedly scored poorly in terms of difficulty and discrimination each year could be replaced or removed and this would require little rewriting (Bates, 2018; Mabe, 2017; Carpenter, 2016). In the long term, the test should also be overhauled to fit in with changes occurring in JHS and SHS and government English education policy and the FE curriculum. A more communicative, standardised approach that tests both receptive and productive skills would enable students to track their progress more effectively.

References

Alderson, J., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge Language Teaching Library. Cambridge: Cambridge University Press.

Bates, D. (2018). An Analysis and Review of the 2017 Freshman English Placement Test at Asia University. *CELE Journal*, 26, 1-9.

Brown, J. (2002). English Language Entrance Examinations: A Progress Report, Curriculum Innovation, Testing and Evaluation. *JALT 1st Annual PAN-SIG Conference*, 95-112.

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

George, D., & Mallery, P. (2003). *SPSS for Windows Step by Step: A Simple Guide and Reference. 11.0 update (4th ed.).* Boston: Allyn & Bacon.

Heaton, J.B. (1989). *Writing English Language Tests.* Longman Handbooks for Language Teachers. London: Longman.

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

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.

McNamara, T & Rover, C. (2006). *Language Testing: The Social Dimension.* Oxford: Blackwell.

Nagata, H. (1995). Testing Oral Ability, ILR and ACTFL Oral Proficiency Interviews, Language Testing in Japan, *JALT Applied Materials*, 12, 108-118.

Sato, K. (2002). Practical Understanding of Communicative Language Teaching and Teacher Development (pp. 41-81). In S.J. Savignon (Ed.), *Interpreting Communicative Language Teaching.* New Haven, CT: Yale University Press.