Improving the Freshman English Placement Test (FEPT): Some Thoughts on Validity and Information Management

Chris Koelbleitner, Asia University

The Freshman English Placement Test (FEPT) has recently been the object of attention at Asia University for a variety of reasons. Many teachers at the Center for English Learning (CELE) felt it was time to make cosmetic changes to the test. Meanwhile, AU administrators have been considering replacing the FEPT with a test that enjoys wide public recognition, such as TOEIC, TOEIC-bridge or G-TEC. Given this recent interest in the value of the FEPT, it may be useful to review the FEPT's history, reflect on how it has been used (and misused) since its creation, and consider some of the suggestions that have been made for its improvement.

BACKGROUND

The FEPT was created in 1997 by Douglas E. Forster and Michael Kearney, both visiting faculty members (Forster and Kearney, 1997). In 2001, oral proficiency interviews (OPIs) were incorporated in the placement process, which continues to serve as a way of checking and correcting the FEPT results. In 2004, there were improvements made to the appearance and audio track of the FEPT. These changes did not substantially change the content of the test. Please see Phil Barkman's "Improving the Asia University Freshman English Placement Test: Incorporating New Technology" (Barkman, 2005) for details.

THE USE AND MISUSE OF THE FEPT

The FEPT is currently being used as a placement test for freshman students, as well as an exit test at the end of the academic year. This much is in keeping with Forster and Kearney's intentions. However, there are a number of ways in which the use of the FEPT has deviated over the years from its makers' designs.

Forster and Kearney devised a special scoring system intended to maximize the FEPT's placement potential by first grouping students with similar listening scores and then subdividing these groups into classes according to their reading scores. Listening scores were to be scored first "because the medium of instruction in Freshman English is oral/aural" (Forster and Kearney, 1997, pp. 145) Kearney and Forster's article also advises that class levels would need to be renamed in accordance with their system (Forster and Kearney, 1997, pp.156).

At some unknown period of time, the test ceased to be scored in this manner. Instead, only the listening scores were used. The reading scores were not used at all. Not surprisingly, given that the scoring method was not followed, classes were not renamed to conform to Kearney and Forster's system. Unfortunately, no CELE Journal article or any paper trail whatsoever has been discovered to explain the deviation from Kearney and Forster's original scoring method. This change in scoring was discovered by accident at the end of the 2003-4 academic year.

Senior VFMs at that time justified this method of scoring by claiming that the listening scores were a better indicator of a students' general oral ability and thus better suited as an indicator of their FE level, since the FE program emphasizes speaking and listening over reading and writing. When asked why they continued administering the reading portion of the test, they

said that the reading section was viewed as a potential secondary system of placement, in the event that the students with identical listening scores needed to be placed in different classes to meet class size requirements. These senior VFMs seemed unaware that they were not scoring the test in accordance with the test's designers' intentions.

Kearney and Forster certainly cannot be blamed for this lapse in the information flow. Besides their CELE Journal article and the materials they placed in the committee files, they also issued a memo to several Japanese professors and administrators explaining their scoring system. A copy of the memo in the committee files is dated October 29, 1996 and is addressed to the following individuals: Mr. Kobayashi, Mr. Matsuta, Mr. Misawa, Ms Tanaka and Professor Miyama. From this list of addressees, only Professor Miyama remains at CELE, as its director. When questioned about this memo at the end of the 2003-4 academic year, Professor Miyama admitted he had neither any memory of it, nor did he know that the FEPT reading scores were not being used. Nonetheless, at some point in time, members of CELE instructed library staff to program the scoring software in such a way that only the listening scores of the FEPT would be used to place students.

What this sad chapter in the life of the FEPT demonstrates is the need for greater overall program continuity. In part, the mismanagement of the FEPT can be blamed on the vagaries of a 5 year limited term of contracted work. As VFMs come and go, some for only a year or two, it is difficult to maintain a high level of information management.

As evidenced by Professor Miyama's lack of knowledge about the FEPT, important changes at CELE can go entirely unnoticed by the permanent members of AU staff.

While developing the FEPT, Forster and Kearney envisioned an "FEPT Coordinator" who would "be responsible for seeing that the test is scored and students are properly placed" as well as ensuring that "the content of the test questions…be analyzed and revised (or changed) as necessary" (Forster and Kearney, 1997, pp. 156). In order to ensure that greater care is taken in managing the FEPT, it may be prudent to have one individual whose yearly project it would be to carry out such analysis and revision of the FEPT.

It should be noted that Forster and Kearney's system of scoring the FEPT will be given greater scrutiny during the April 2005 intake period, at which time it will be compared with other methods of processing FEPT results.

REVISING THE FEPT

When approaching the possibility of revising the FEPT, it is important that we are sure we know what areas are in need of improvement. The FEPT, perhaps due its lowly status as an in-house test, has been the object of a number of groundless criticisms. I would like to consider some of these, beginning with a couple of points raised by H.L.P. Molloy's article ""How reliable is the Asia University Freshman English Placement Test? A Classical Internal Reliability Study" (Molloy, 2004). It needs to be said that Molloy's highly informative article has ultimately raised the status of the FEPT in the eyes of administrators and faculty and paves the way for a much more informed approach to its management as a placement tool. Nevertheless, there are a few minor issues raised by this article that need to be addressed.

Molloy's article argues that a test must be evaluated according to two registers: validity and reliability. For Molloy, validity "refers to whether the measurement tools one is using are appropriate for measuring what you are interested in" (Molloy, 2004, pp. 64). Meanwhile, the reliability of a placement test refers to its ability to yield the same results if it were given to the

same student at different times (Molloy, 2004, pp. 66). Molloy makes the focus of his paper the reliability, not validity, of the FEPT, but not before criticizing the validity of one of its questions. Here is his objection to question 59 in its entirety:

> Consider question 59 of the 1999 version of the FEPT.
> 59. Keiko is a very pretty girl; _____, she is extremely intelligent.
>
> A. therefore
> B. as a result
> C. moreover
> D. on the contrary
>
> This is a question from the "grammar" section of the FEPT. The purpose of the grammar section of the test is to test participants' (students') knowledge of grammar, yet each answer fits makes a grammatically correct sentence. Given that it is well known that physical attractiveness is often associated with greater or lesser estimates of intelligence in different circumstances (Etcoff, 1999, pp. 46, 52), we might consider this question one that is more valid for measuring social attitudes than for measuring the ability to manipulate the linguistic code in English. (Molloy, 2004, pp. 65)

Most of us are familiar with psychology studies in which individuals are found to over-attribute various qualities like intelligence or kindness to attractive people. These studies may use photographs or video clips of people to measure to what degree test subjects viewing these images associate physical attrraction with other personal qualities. However, such tests do not measure subjects' logical thought processes, nor their conscious belief structures. These tests measure sub-rational impressions and the emotional responses that underpin them. Sub-rational impressions of an attractive person's abilities are quite different from an individual's logical understanding of the relationship between two qualities like prettiness and intelligence. A pretty girl may seem more intelligent because she is attractive, but there are very few people whose rational beliefs include the notion that prettiness is a reliable predictor of intelligence, as suggested by options A. and B. For a student with sufficient English ability, question 59 asks only whether it is logically sound to say that Keiko's intelligence is a natural product of her prettiness (options A and B), or merely another one of her positive qualities (option C). Option D does not yield any meaning, logical or illogical, that can be paraphrased, at least for this reader. Thus Question 59 does not measure the emotional and sometimes irrational processes that constitute "social attitudes" as Molloy suggests; it measures a test-takers ability to use conjunctions to create a logically sound sentence. Measuring the correct usage of a conjunction would surely fall within the parementers of the grammar section of a placement test.

Molloy raises another point, this time concerning the FEPT's reliability, that needs to considered more carefully. He introduces the issue of reliability by listing a number of ways that this value can be measured. Among the methods he mentions is one in which students' results on one test are compared to their results on another.

> We could correlate students' FEPT test scores with TOEIC scores, for example, and see if students who score high on the FEPT score high on the TOEIC. A perfect correlation would be if a one-point change in students' FEPT scores was

> always paralleled with, say, a five-point change in their TOEIC score. (Molloy, 2004, p. 66)

Molloy's article mentions this method only in passing; however, there is a dangerous assumption in this passage that a students' TOEIC scores would be an accurate reflection of their English ability, and therefore a touchstone for the reliability of the FEPT. The question of whether TOEIC scores can be used as a way of measuring the FEPT's reliability is an important one at AU, since in the past TOEIC scores have been used to question the usefulness of FEPT on a number of occasions. Even after in-house research and direct communication from TOEIC representatives indicated that TOEIC was not a reliable measure of AU's freshman students, teaching staff and administrators continued to consider TOEIC the more reliable test, compared to the FEPT. For more information about the use of TOEIC at AU, please see "An Examination of the Proposed Use of TOEIC at Asia University" (Koelbleitner, Gustavsen, Alberding, 2003).

There are a number of other objections that have been made concerning the FEPT that need to be addressed. These are criticisms that are based on informal impressions of the test and tend to lack evidence to support them. Nevertheless, these informal impressions have been allowed to influence administrative decisions concerning changes to the FEPT. These objections are that the speed of delivery of some of the questions is too fast, and that some of the questions are too difficult to be useful in placing students.

The problem with such objections is that they tend to focus only on the apparent difficulty of a listening passage, without considering what exactly students are being asked to listen for. On a multiple choice test like the FEPT, the difficulty of a test item will depend a great deal on the obviousness of the correct answer and/or the complexity of the distracters. The only way to truly evaluate the value of questions on the FEPT is to determine how well, statistically speaking, they perform as a placement mechanism. In some cases, revising a question to meet subjective criteria without recourse to statistical data may actually lower its value as a placement device. To illustrate this point, let us consider questions 43-45. In the 2000 version of the FEPT, the rate of delivery of the passage connected to these questions is quite rapid. Among VFMs familiar with this version of the FEPT, it was a commonly held belief that this question was too difficult to be of any use as a placement tool. The "self-evident" difficulty associated with the rapid rate of delivery of this passage was among the chief reasons the audio track of the FEPT was revised in 2004. However, if we take a look at the overall results of the students who wrote the FEPT in 2003, we find that 61% of students answered question 43 correctly. The number of students who answered question 44 (37%) and 45 (22%) correctly were much lower but this only highlights the fact that the quality of multiple choice answers will determine the usefulness of a listening question as much as factors associated with its spoken delivery . By slowing the delivery, we may in fact find that some questions of the 2000 FEPT may become too easy to be useful for placement purposes. The passage associated with questions 37-39 in the 2000 edition has also been unfairly maligned by numerous VFMs. This passage concerns James Joyce, and is delivered by a speaker whose Irish accent is quite evident. Both the complexity of the vocabulary, and this individual's accent was viewed by many VFMs (who, at this time, are primarily North American) as sources of excessive difficulty for students. If we consult the results, we once again find that a high proportion of students were able to answer two of these "difficult" questions correctly. 42% of students answered question 37 correctly while 41% answered 39 correctly. Question 38 was the only one that seemed difficult for students – only 17% answered it correctly. It is possible that in this case North American VFMs allowed their

unfamiliarity with the inflections of Irish English to influence their judgement of the usefulness of this question. Of course, having an Irish person deliver a passage about James Joyce bestows a measure of cultural diversity upon the FEPT that would be unfortunate to lose.

It should be noted that counting the number of students who answered a question correctly provides only limited information about the overall reliability of a placement test. There are far more sophisticated measures that would account for lucky guesses. My observations are only intended to suggest that closer scrutiny of statistical measures of the FEPT should be used to evaluate its usefulness in the future. For a thorough account of the ways in which the reliability of a placement test can be measured, see Molloy (add biblio). However, it also needs to be said that despite the availability of any amount of statistical data demonstrating its validity, the FEPT will probably only survive as CELE's primary placement tool if it can also satisfy the subjective and aesthetic whims of that portion of administrators and faculty members who do not trouble themselves with objective measures of the FEPT's value.

References

Barkman, P. (2005) "Improving the FEPT: Incorporating New Technology". *CELE Journal* No. 13, pp. 93-95.

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

Molloy, H.P.L. (2004) "How reliable is the Asia University Freshman English Placement Test? A Classical Internal Reliability Study" *CELE Journa*l No. 12, pp. 64-86.

Koelbleitner, C. Gustavsen, E., Alberding, M. (2003) "An Examination of the Proposed Use of TOEIC at Asia University." *CELE Journa*l No. 11, pp. 115-124.