How Reliable is the Asia University Freshman English Placement Test? A Classical Internal Reliability Study

Hugh P. L. Molloy, Asia University

INTRODUCTION

The Freshman English Placement Test (FEPT) is given to about 1600 incoming first-year students each year (Koelbleitner, Gustavsen, & Alberding, 2003) at Asia University. It is used in conjunction with an oral proficiency interview (OPI) to place students in one of 22 levels of required first-year English classes (Koelbleitner, 2003).

The FEPT is what is called a "norm-referenced" test (Brown, 1996), a test that is used to answer the question "how does this test taker compare with other people?" In this case, the FEPT asks the question "how does this student compare with others with regard to English proficiency?" (The other kind of test is called a "criterion-referenced" test, which asks the question "can this student perform this task?" Achievement tests, such as in-class final exams, are criterion-referenced tests.)

The FEPT, then, can be considered a tool for rough ranking of students with regard to English proficiency. Because it is used with a second test, the OPI, the FEPT does not have to place students perfectly into groups: fine discrimination can be done with the OPI.

The general question addressed in this short report is this: how well does the FEPT work? As with all tests, the FEPT is a tool that is designed to measure something that cannot conveniently be measured quickly or directly. A blood-pressure test, for example, is designed to measure health (or, more properly, future health), a complex idea that cannot assessed directly except negatively—sick people are clearly not healthy—or after it not longer matters for the patient. The FEPT as well attempts to measure an idea (or "construct") that cannot be measured directly: English proficiency. Like all paper-based academic tests, the FEPT depends on the assumption that the ability to answer the questions on the test depends on how much proficiency the student possesses. If a student has a lot of English proficiency, he or she will be able to answer more questions correctly. If he or she has little English proficiency, he or she will be able to answer proportionately fewer questions correctly.

With tests, the question of how well a test "works" is addressed from two viewpoints: validity and reliability. The first viewpoint is the question of validity. "Validity" in testing refers to whether the measurement tools one is using are appropriate for measuring what you are interested in. For example, if we are interested in measuring people's heights, we might use a ruler or tape measure. Most people would agree that the ruler or tape measure is an appropriate measuring device for height. However, we might also try measuring people's height by asking their friends how tall they are. We might try it as well by recording the sizes of their shirts. Asking people how tall their friends are, however, we might argue is not so good a measure: affective factors might come into play, if, for example, people tend to see people they like as taller or shorter. Checking the size of shirts, we might argue, might be a bad idea because some people might wear larger or smaller shirts for fashion's sake, not simply because they are of a certain height.

Height is a relatively straightforward concept, but language proficiency is not. It is very difficult to agree on what constitutes proficiency in any language. (See, for example, Jacoby & McNamara, 1999.) Even when test makers agree on what proficiency is, the question remains of how it can be measured. Is an interview like the OPI a good approach? Is a paper test like the FEPT or the TOEIC a good approach? Do specific questions really measure what they purport to measure? 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.

Questions of test validity are usually decided at two times: when the test is being written, and when the test is being used for measurement. Test writers decide if particular questions actually measure what they are supposed to measure, and test consumers (that is, the people who use the results of the test) decide whether the test as a whole is a correct tool for the use to which the test is put. The two questions in the case of the FEPT, then, are: "Do the test questions in the FEPT actually measure skill in English use?" and "Does the FEPT allow us to rank students in English proficiency (rather than in something else)?" Neither question is a focus in this short paper, however.

This paper focuses on “reliability.” The question of reliability is this: How consistently does the test measure? That is, if we give the same test to the same student at two different times, will the student get the same score? If a test is reliable, the answer will be “yes.” Going back to the height example, if we measure a person’s height with a ruler on Monday and again on Wednesday, we will probably get roughly the same answer. Our ruler “test” will be fairly reliable. Note, however, that the ruler test will not be perfectly reliable: if we measure, for example, to the millimeter, it is likely that there will be some small discrepancies in the height we measure. Part of the discrepancy will come from the instrument we use: if the ruler is calibrated at the millimeter level, our average measurement will be about 0.5 millimeters off the “true” measurement. Part of the discrepancy will come from tiny actual changes in the person’s height. If, for example, the person drank a lot of water the day before the second test, the thickness of his or her skin might be slightly different.

With tests of mental ability, as tests of language proficiency are often considered to be, checking reliability is a little more difficult than testing physical properties. We cannot assume that giving the exact same test two days in a row will work. The student is likely to remember the answers to some questions, which will give him or her more time to spend thinking about the answers for more difficult questions. He or she will have had time to think about the answer to questions. The student might have gone home and checked some words in a dictionary or studied a chapter in a textbook that might help in answering questions. All of these factors can lead to a higher score. On the other hand, the student might be bored or resentful at having to spend time doing the same test again. He or she might assume that having to take the same test again means he or she did badly on the first one and change previously correct answers for this reason. These and similar factors might lead to a lower score. The student might have a headache on one particular day, which might also lead to discrepant scores.

Checking the reliability of a language test, then, is usually not done by giving the student the same test twice, a reliability checking method known as the “retest” method. Instead, at least one of several other methods is used.

One method to check the reliability of a test is to correlate students’ test scores with their scores on another test. 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 scores.

Another method is known as the parallel-forms methods. This involves having two tests that measure the same thing and giving both tests to the students. For example, if there were two versions of the FEPT, we could give half of the students version A one week and version B the second week and reverse the order for the other half of the students, then correlate the scores on the two tests. (It would be necessary to give the tests in reverse order to half the students because taking the test A might allow students to learn some English that will improve their scores on test B. Counterbalancing the order allows the learning effects to cancel themselves statistically.)

A third method of checking reliability in language tests is called internal reliability checking. This is by far the most popular method of reliability measurement, and it is the one used in this paper.

There are currently two principal methods of internal reliability checking in the language testing field: "classical" internal reliability checking and a new method known variously as Rasch, latent-trait, or item-response-theory measurement (see, for example, Baker, 2001, for an introduction to this latter approach).

Classical reliability measurement has the advantage of being conceptually easier to understand and mathematically less burdensome to calculate.

Classical reliability measurements make two assumptions about the test: that all of the items ("items" here meaning test questions) measure the same thing (in this case, English proficiency), and that each item is independent.

Independence here means that the answer to one question is not connected with the answer to another question. Consider the following two questions:

A. Which city in Japan has the largest population? ______________.

B. What two colors are in the flag of Japan? ___________ and ___________.

My answer to question A does not have anything to do with my answer to question B. These two items are independent. On the other hand, consider these two questions:

A. Which city in Japan has the largest population? __________________.

B. What is the population of that city? ___________________.

If I make a mistake with question A and write "Osaka," my answer to question B will be "about 2.5 million." My answer to question B depends on my answer to question A—notice that we cannot reverse the order of the questions—and so these questions are not independent.

Classical test reliability measures, then, assume that each question is a different, independent way to measure the same thing. The more times and the more different ways we measure something, the assumption holds, the more accurate our final answer will be. Hence, if we measure someone’s height ten times using ten different rules and average the answer, we assume that the answer is closer to the “true” height than if we measure the height only once. With the FEPT, the assumption is that each of the questions measures English proficiency. Students who can answer a given question correctly are considered more proficient than students who cannot. With the FEPT, we are taking 75 such measurements of students’ English proficiency.

Note, however, that the FEPT is divided into three sections: listening, grammar, and reading. Each section has a different number of questions. The listening section has the greatest number of questions, followed by grammar, then reading. Classical test reliability theory leads us to assume that, all other things being equal, a greater number of test items (questions) will lead to greater accuracy. Hence, before checking any statistics, we can guess that the listening section will be the most reliable section of the FEPT, that the reading section will be the least reliable, and that the FEPT as a whole will be more accurate than any of its parts.

However, the division of the FEPT into three sections can actually lead us to the opposite prediction: that the reliability of each subtest (listening, grammar, and reading) should be higher, not lower, than the reliability of the test as a whole. If we assume that each of the subtests is designed to test a skill or quality that is wholly different from the others, it makes sense that a group of questions designed to test one and only one skill (grammar, for example) will be more consistent than a group of questions designed to several skills.

Which is the correct answer? Does the FEPT actually seem to be testing three differentiable skills (listening, grammar, or reading), or does it seem to be testing only of general skill of English proficiency? The answer to this question depends much on the test writers’ or test consumers’ judgments about the validity of the test, but we can gain much evidential information that can help us to answer the question mathematically by seeing if different students are answering the different types of questions differently. If we find that one group of students does well on the grammar section and a different group does well on the reading section, this can be considered evidence that grammar and reading (as characterized by the questions in the FEPT) can be considered different skills.

For this short paper, I will consider three research questions:

1. How reliable is the FEPT?
2. How reliable is each subtest of the FEPT?
3. How accurate is the score for each student?

4. Are there really three subtests in the FEPT?

The first two questions I will approach by using classical test reliability calculations. Further information on these calculations can be got from any basic work on testing written for the social sciences (e.g., Hatch & Lazaraton, 1991; Brown, 1996). I will also use some calculations on item functioning. Item functioning is a general term that refers to several algorithms for characterizing how well individual questions on a test work. The most common item function calculations can be found in Brown (1996), especially in chapter 4.

The third question I will approach by calculating the standard error of measurement, as a calculation that shows us how the test score relates to an estimated "true" score of the student.

The last question I will approach with a statistical technique called factor analysis. Factor analysis is a computationally tedious procedure that works, roughly, by looking for groups of questions answered in the same way by groups of test takers. If the listening, grammar, and reading subtests of the FEPT actually do test different skills, then the questions in those three sections should show different answer patterns. A detailed explanation of factor analysis can be found in Tabachnik and Fidell (2001) and a more user-friendly explanation in Hatch & Lazaraton (1991).

METHOD

Participants

The participants were 1535 entering Asia University students who took the FEPT in April 2003. Judging from the names of the students, the majority seem to be Japanese.

Sixty participants in the data set analyzed had scores of zero, indicating some sort of problem with the test. These participants were eliminated from the analysis, giving a total of 1475 participants.

Materials

The FEPT is a 75-question test in multiple-choice format (Forster, Kearney, & Ridge, 1999) comprising nine parts. Parts 1 through 5 (questions 1-54) are designed to test listening; parts 6-8 (questions 55-71) are designed to test knowledge of grammar; and part 9 (questions 72-75) is designed to test reading. Each multiple-choice question has 4 answer choices (one correct answer and three distracters), which the exception of Part 1, which has five answer choices.

The analyzed data were recorded in a Microsoft Excel (Microsoft Excel, 1999) spreadsheet and comprised the participants' answer choices. For the classical reliability studies, the data were recoded to binary data: that is, participants' answers were recorded as simply correct or incorrect.

The binary data were transferred to SPSS format (SPSS, 1999) for further analysis.

Analysis

Classical reliability analysis calculations were done step by step in Microsoft Excel (Microsoft Excel, 1999), using formulas presented in Brown (1996) and checked with formulas provided in Hatch & Lazaraton (1991) and with the formulas used in SPSS versions 10.0.1 and 9.01 (SPSS, 1999; SPSS, 1998).

Four different reliability calculations were used:

Cronbach alpha. This is the most common reliability measure used in second-language research. It functions by comparing one half of the test to the other half. Usually the test is split in half by treating the odd-numbered questions as comprising one test and the even-numbered questions another test; the Cronbach alpha formula then compares the two halves of the test. The formula involves calculating the standard deviation for each half of the test. Standard deviation is the most common way to show how widely dispersed test scores are: it can be considered the "average differentness" from the average score on the test. For example, for the April 2003 administration of the FEPT, the mean (arithmetic average) of the test scores was 39.51. The standard deviation was 8.92. What this means specifically is that about 68% of the students scored from 30.59 to 48.43 on the FEPT. If the standard deviation had been 1, it would have meant that about 68% of the students scored between 38.51 and 40.51 (and that the scores are piled up in the middle of the range); if the standard deviation had been 15, it would mean about 68% scored from 24.51 and 54.51 (and that the scores are spread out all over the range of possible scores). The standard deviation of 8.92 is not problematic for this test.

Adjusted split-half reliability. This calculation is easier than Cronbach alpha, but usually a little less accurate. It works by (again) splitting the test into two parts and comparing the scores from one test to the scores from the other. It is slightly less accurate than Cronbach alpha because it does not account for variation in scores. (That is, it does not involve the standard deviation.)

Kuder-Richardson formula 21. This commonly reported measure is calculated from the number of items on a test, the standard deviation of the test, and the mean of the test. It usually gives an estimate of reliability that is more conservative (that is, lower) than the other calculations.

Kuder-Richardson formula 20. This is the most accurate formula, but the most difficult to calculate, as it involves calculations derived from the answers given to each question on the test. (The other three reliability measures depend only on total scores or subtotal scores.)

Standard error of measurement. This formula is used to calculate how accurate the score for the typical student is. What it shows is the range of scores the student would be likely to get if he or she were to take the test again. For example, if a student gets a score of 38 on the FEPT and the standard error of measurement (SEM) is 3.74, we can assume that the student's "real" or "true" score is probably between 34.26 and 41.74. What does this mean in terms of the FEPT test? It means that with an SEM of 3.74, we cannot say that the abilities of a student with a score of, say, 33 and one with a score of 38 are different. Again, though the students' scores are different, we cannot conclude that their abilities are different, because their scores are not different enough.

Item functioning. Item functioning calculations can help in determining which questions on a test are the most useful ones for the purpose of the test. Regardless of how good a question is, if it does not fit in with the rest of the test, it should not be used in the test. A mathematics question expressed in figures only could be put in the FEPT and work perfectly reliably, but it would probably not tell us much about the test-takers' English proficiency.

The several item functioning calculations I made were as follows:

1. Item facility (IF). This is a measure of how easy a question is. For the FEPT, for example, the IF for question 1 was about 0.90, which means that 90% of the students got the question right and it is a relatively easy question. The IF for question 20, on the other hand, was about 0.40, which means it is a much more difficult question.
2. Item discrimination (ID). This is a measure of how well the question differentiates between students with high total scores and students with low total scores. An item with a high ID will be answered correctly by most of the top test-takers and incorrectly by most of the least able test-takers. An item with a low ID will be answered correctly by an equal proportion of top and bottom test-takers, which would mean that the test item is not useful for discriminating between "good" and "bad" students. An item with an ID of 0.00 would not be discriminating at all between "good" and "bad" students and probably should be dropped from the test. Item discrimination calculations depend on defining a group to represent the highest scoring students and another to represent the lowest scoring students. In this paper, I have simply chosen the 500 students with the highest scores to

represent the high group and the 500 with the lowest for the low. We could just as easily choose the top 25% and the bottom 25%.

3. Point-biserial correlation (r_{pbi}). This is another calculation that shows how well items discriminate between high- and low-level students. It is a little more accurate than ID, but more difficult to calculate. It is slightly better than simple item discrimination calculations in some cases, as it includes all of the students who took the test, not only the top and bottom groups.
4. Factor analysis. Factor analysis is, as mentioned earlier, a complicated procedure. It is also one accompanied with myriad disputes about approaches to calculation: for a beginning summary, the reader is referred to the relevant chapter of Tabachnick and Fidell (2001). For the calculations presented here, I used all of the most popular settings and set SPSS, the statistics program I used (SPSS, 1999), to look for three divisions in the data. This was because there are three subtests in the FEPT: if the test actually does measure three different skills, then the factor analysis program should give a reasonably economical summary of the three tests.

RESULTS

In this section, I simply present the results of the several calculations I did and explain how to read the tables. Interpretation I leave to the following section.

Descriptive Statistics

Descriptive statistics is a term that refers to several calculations designed to summarize the information in a group of numbers, in this case, scores on the FEPT and its subtests. Results are presented in table 1 and were calculated in Microsoft Excel (Microsoft Excel, 1999) and checked with SPSS version 10.0.1 (SPSS, 1999).

Table 1.

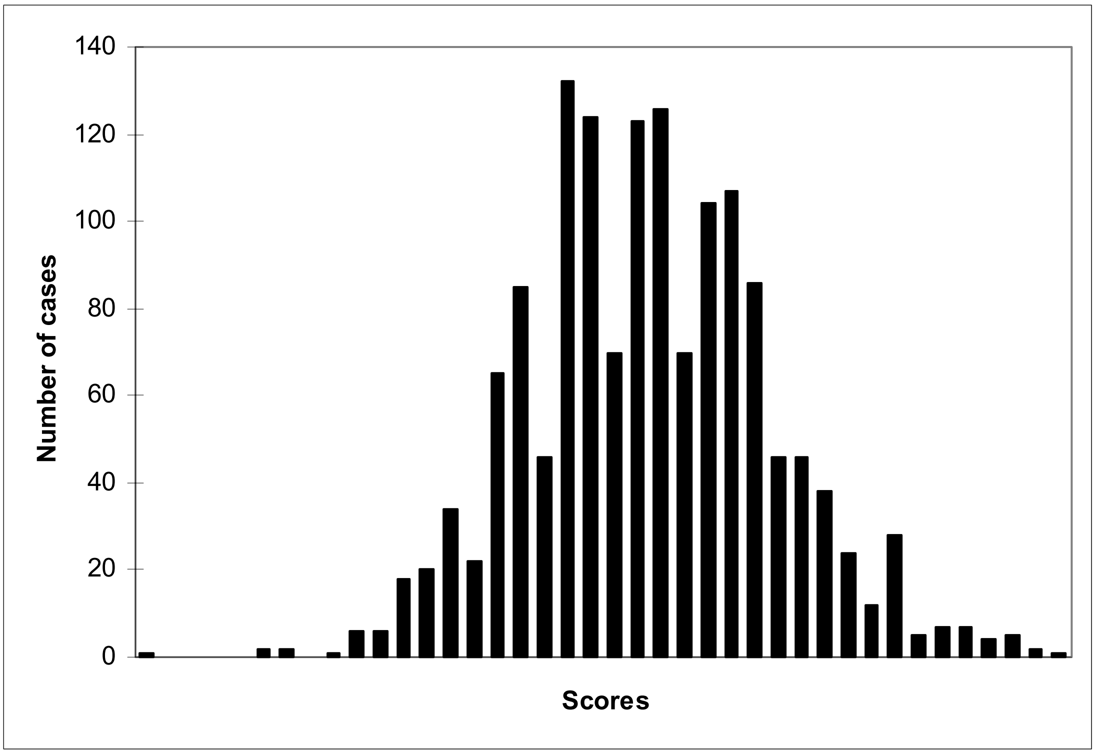
Descriptive statistics for April 2003 administration of FEPT

	k	x	C.I.	s	Kurtosis	Skew	Range	Min.	Max.
Total	75	39.52	0.45	8.90	0.27	0.16	66	4	70
Listening	54	30.52	0.33	6.49	0.49	-0.05	49	4	53
Grammar	17	7.06	0.14	2.82	-0.27	0.32	16	0	16
Reading	4	1.94	0.06	1.16	-0.85	0.18	4	0	4

The table can be read as follows.

The first column, k, represents the number of items (questions). The second, x, is the (arithmetic) mean (average) for the test. The third column, C.I., represents the confidence interval. This is the amount by which the mean might vary, so that for the total test, the C.I. of 0.45 means that there is a chance of about 68% that the true mean score for this particular group of 1475 students will be between 39.07 and 39.97. The next column, s, is the standard deviation which, as explained earlier, in this test means that about 68% of the students got scores between 30.62 and 48.42. The next two columns, kurtosis and skew, represent how far the scores deviate from a perfect bell curve. Positive kurtosis means that the curve made by these scores is a little higher and narrower than the perfect bell curve, meaning that there are comparatively many scores in the middle of the range and only a few at the top and bottom ends. Negative kurtosis means the curve is flattened and that there are relatively many scores at the top and bottom ends. Skew is a rough measure of whether the curve is pushed to the right (negative skew) or the left (positive skew) which would mean that there are some very low scores (negative skew) or very high scores (positive skew). As can be seen in the histogram showed in Figure 1, the scores for the total FEPT are very close to the normal curve.

Figure 1. Histogram of total scores for the April 2003 administration of the FEPT. High total test scores are on the right and low on the left. “Number of cases” represents the number of students who scored at that level.



As can be seen from Figure 1, the distribution of scores looks more or less like a normal bell curve. It may look a little too tall, but the deviation is not big enough to be worrisome. The shape of the distribution is important: if the shape is close enough to a normal bell curve, many standard statistical tests are possible: most of these tests depend on comparing the numbers one wants to check with the known mathematic properties of the normal bell curve. If your data don't look like a normal bell curve, using many statistical procedures is inappropriate, a matter of comparing apples and oranges.

Reliability measures

For the total test, four types of reliability measures (and the nearly equivalent SEM) were calculated for the total test. For the subtests, I only calculated Cronbach alpha and split-half reliability. Results are presented in Table 2.

	Cronbach alpha	Adjusted split-half	K-R21	K-R20	SEM
Total test	0.84	0.83	0.78	0.83	3.74
Listening	0.77	0.67			
Grammar	0.58	0.57			
Reading	0.50	0.39			

Note that, as mentioned earlier, the K-R21 calculation is lower than the others. Again, as mentioned earlier, the K-R20 calculation is probably the most accurate.

Item functioning

Appendix A shows the item functioning calculations for each of the questions. As explained above, item facility (IF) shows how difficult a question is (with higher numbers being easier questions) and item discrimination (ID) and point biserial correlation (r_{pbi}) show how well the question differentiates between high scoring and low scoring students.

Factor analysis

A factor analysis was performed with principal components extraction, varimax rotation, and 3 components extracted. Three components were chosen because there are three subtests in the FEPT, implying that there should be three distinct patterns of answering, if the three subtests actually do test different skills. If the three subtests do not test separate skills, then no clear patterns should emerge. Appendix B presents a rotated components matrix.

Factor analysis tables are difficult to read without training. However, some common rules of thumb are that if an item has a number of higher than 0.30 in a particular area, it can be said to "belong" to that area. Items that have numbers higher than 0.30 for two areas are probably testing two things at once. Items that have no numbers higher than 0.30 are probably testing something different from any of the three areas.

I have highlighted all of the numbers higher than 0.30. If the three subtests actually test three different (and differentiatable) skills, then the majority of the items from the listening subtest, for example, should have numbers higher than 0.30 in the same column. An important thing to remember is that the factor analysis procedure in this case only "accounted for" 13.83% of the variation in test scores. This is a very low number and gives us a hint that the three main patterns detected in the scores probably cannot tell us much about the test.

DISCUSSION

I will present the discussion according to the research questions.

How reliable is the FEPT?

According to the statistics calculated in the analysis, the FEPT as a whole is a reasonably reliable test: a reliability measure of 0.80 is usually considered good enough for most purposes, and the reliability for this administration of the FEPT was 0.82, as calculated with the most accurate formula, the K-R20 formula.

For comparison, the reliability of the most well known standardized tests of English proficiency (the TOEFL, TOEIC, and Eiken tests) hovers around 0.90 or higher.

Judgments of how reliable a test has to be have to be made with reference to the purpose to which the test is put. The FEPT is used as a preliminary ranking device to be followed by the OPI. So long as the OPI is used to make final decisions about the placement of students, the FEPT is probably sufficiently reliable.

To make a test more reliable, there are two main strategies employed: making the test longer, and using better items. Making tests longer will always make a test give higher classical reliability measures (as will be apparent when one considers that most such measures include in their calculations the number of items), but there is a tradeoff with lengthening a test: long tests are more difficult to develop and score, and fatigue comes to play a role. A 500-question test might give more reliable results, but only if fatigue does not come into play.

The second strategy, improving the items, is a more reasonable option for development of the FEPT. The item functioning statistics presented in Appendix A show that some of the questions (such as questions 9, 15, 45, and 65) do not differentiate well between the most proficient and the least proficient students. Replacing these items with more discriminating items would make the test more reliable. Indeed, question 45 should probably be replaced. The item facility statistic (0.22) shows that it's a relatively difficult item, but the item discrimination statistic (0.01) and point-biserial correlation (0.01) show that least able and most able students have roughly equal chances of answering the question correctly. We might suspect that all of the correct answers came about by simple guessing and that there is some problem with the item. Simply deleting such items entirely would probably not impact reliability much.

How reliable is each subtest of the FEPT?

Each subtest of the FEPT is not particularly reliable. This lack of reliability probably has two sources. First, as explained earlier, most classical reliability measures depend on the number of items directly or indirectly and the grammar and reading subtests have especially few items. Second, as can be seen in the factor analysis results presented in Appendix B (and discussed further below), the questions in the three subtests are not being answered in ways we would expect if the subtests were really testing separable skills. Note that the questions from the listening subtest seem to "belong" to two different groups, which may be taken a an indication that two different things are being tested in this subtest, which contributes to the low reliability measurements.

Regardless of the interpretation of the particular results of this test, it is clear that it would be irresponsible to use scores on the three subtests as indications of skill in listening, grammar, or reading. The subtests are simply not reliable enough for such an interpretation of scores.

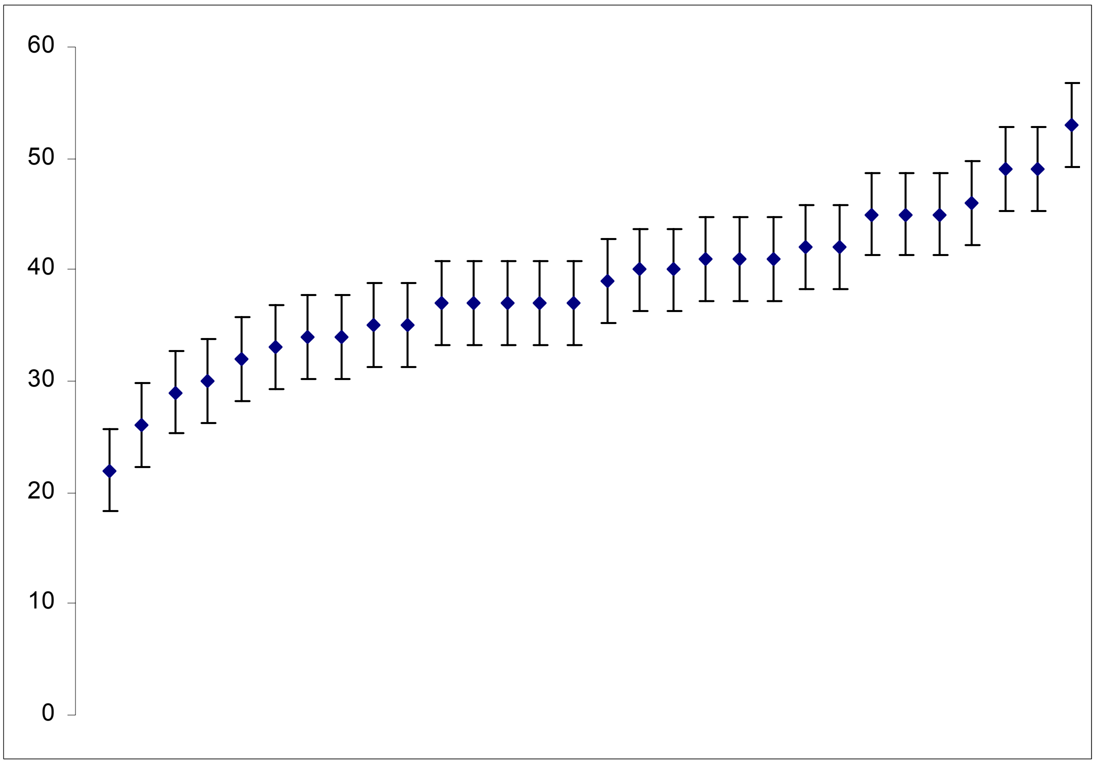
How accurate is the score for each student?

This is a very important point: Table 2 shows that the SEM for the entire FEPT test is 3.74, which means that each student's score should be considered to belong somewhere in a range from –3.74 to +3.74 the score he or she obtained on the FEPT.

The most important implication of this SEM is that students who score 3.74 points apart cannot and should not be distinguished from each other. If we have three students who score, respectively, 25, 28, and 31, we can say that the first and third students probably differ in proficiency, but not that student 2 differs from either student 1 or student 3.

Figure 2 gives an illustration of this phenomenon. To make this chart, I randomly selected 30 students from the 1475 who took the April 2003 FEPT. I chose 30 because this is a typical university class size. The chart shows the total FEPT scores (diamonds) and the associated error of ±3.74.

Figure 2. FEPT scores and associated error for 30 randomly selected students. Error bars represent ±3.74 points. The y-axis represents scores. Scores are sorted in ascending order.



To read the chart, interpret the spread shown by the error bars as showing the likely "true" score on the FEPT test (that is, how they would score if they took the test again, forgetting everything they learned between tests). This means that first 3 students from the right, for example, cannot be distinguished from each other. It means also that last 4 students cannot be distinguished from each other. The groups of the first four and the last four students can be distinguished from each other, but that all the students in the middle 20 cannot be distinguished from one another or from all but the most extreme students.

What does this mean? It means that if we put all of these students in the same class, we might have a proficiency mismatch between the first and last students, but that most of the students would be well matched with regard to proficiency, as far as it is measured by the FEPT.

A further, and disturbing implication of this small exercise is this: because I selected these 30 students a random, it is equivalent to making up a class of 30 students with no test whatsoever. Hence, it is not clear that the FEPT would function usefully if it were used as the only placement test. Given the reliability of the FEPT, following the FEPT with the OPI would seem essential for making placement decisions.

If the FEPT were more reliable, the OPI would not be a necessary second test. (The question of the reliability of the OPI has yet to be investigated.) Note as well that, given the range of possible scores in the FEPT (1-74), and the SEM given the reliability of the test, it is probably not possible to realistically distinguish more than 10 proficiency levels with the FEPT. Given the normal bell curve distribution of the test, as well, it follows that, even if we were able to distinguish 10 proficiency levels, the number of students in the lowest and highest levels would be very small, a situation that would make distributing students evenly into a given number of classes according to FEPT proficiency level impossible.

Are there really three subtests in the FEPT?

As can be seen in the factor analysis table, there are probably not three subtests in the FEPT corresponding to the three question groupings in the test. Further investigation might reveal, for example, that the rough groupings that might be forming are due to lexicogrammatic knowledge, not subskills of grammar, reading, or listening.

References

American Psychological Association (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: American Psychological Association.

Baker, F. B. (2001). The basics of item response theory (2nd ed.). Eric Clearinghouse on Assessment and Evaluation. (17 July 2002: http://ericae.net/irt/; http://ericae.net/scripts/ft/ftget.asp?want=http://ericae.net/irt/baker/final.pdf).

Brown, J. D. (1996). Testing in language programs. Upper Saddle River, NJ: Prentice Hall Regents.

Etcoff, N. (1999). Survival of the prettiest: The science of beauty. London: Abacus.

Forster, D. E., Kearney, M., & Ridge, P. (1999). Asia University Freshman English Placement Test. Tokyo: Center for English Language Education, Asia University.

Hatch, E. & Lazaraton, A. (1991). The research manual: Design and statistics for applied linguistics. Boston: Heinle & Heinle.

Jacoby, S., & McNamara, T. (1999). Locating competence. English for Specific Purposes, 18, 23-241.

Koelbleitner, C. (2003). Results of the 2002-2003 FEPT and 2001-2002 I-TOEFL tests. CELE Journal, 11, 125-129.

Koelbleitner, C., Gustavsen, E., & Alberding, M. (2003). An examination of the proposed use of the TOEIC at Asia University. CELE Journal, 11, 115-124.

Microsoft Excel 2000 [Computer software]. (1999). Redmond, WA: Microsoft.

SPSS version 10.0.1 [Computer software]. (1999). Chicago: SPSS.

SPSS version 9.01 [Computer software]. (1998). Chicago: SPSS.

Tabachnick, B. G., & Fidell, L. S. (2001). Using multivariate statistics (4th ed.). Needham Heights, MA: Allyn and Bacon.

Appendix A

Item functioning for the 75 items of the April 2003 administration of the FEPT.

Item	Correct answers	IF	ID	r_{pbi}
1	1329	0.90	0.18	0.27
2	1221	0.83	0.33	0.38
3	1325	0.90	0.23	0.35
4	747	0.51	0.38	0.34
5	1298	0.88	0.18	0.28
6	1302	0.88	0.18	0.27
7	1304	0.88	0.16	0.24
8	1326	0.90	0.18	0.27
9	1420	0.96	0.07	0.20
10	1190	0.81	0.31	0.33
11	1067	0.72	0.32	0.29
12	536	0.36	0.15	0.14
13	997	0.68	0.24	0.23
14	379	0.26	0.19	0.19
15	162	0.11	0.09	0.16
16	404	0.27	0.29	0.30
17	797	0.54	0.37	0.32
18	1203	0.82	0.19	0.23
19	1306	0.89	0.14	0.21
20	575	0.39	0.29	0.29
21	939	0.64	0.41	0.35
22	1339	0.91	0.20	0.31
23	397	0.27	0.22	0.25
24	725	0.49	0.37	0.32
25	1226	0.83	0.18	0.23
26	899	0.61	0.28	0.24
27	869	0.59	0.29	0.28
28	915	0.62	0.10	0.14

29	1297	0.88	0.24	0.32
30	697	0.47	0.29	0.28
31	336	0.23	0.14	0.16
32	959	0.65	0.19	0.17
33	748	0.51	0.16	0.16
34	1124	0.76	0.34	0.36
35	761	0.52	0.25	0.24
36	694	0.47	0.34	0.28
37	598	0.41	0.13	0.13
38	251	0.17	0.06	0.07
39	584	0.40	0.30	0.28
40	857	0.58	0.38	0.32
41	668	0.45	0.45	0.39
42	551	0.37	0.31	0.30
43	866	0.59	0.49	0.41
44	544	0.37	0.10	0.13
45	323	0.22	0.01	0.01
46	851	0.58	0.50	0.44
47	601	0.41	0.37	0.36
48	453	0.31	0.41	0.41
49	852	0.58	0.15	0.19
50	649	0.44	0.24	0.25
51	638	0.43	0.40	0.38
52	525	0.36	0.27	0.25
53	968	0.66	0.27	0.26
54	421	0.29	0.08	0.11
55	1026	0.70	0.44	0.39
56	311	0.21	0.23	0.23
57	649	0.44	0.32	0.28
58	189	0.13	0.07	0.11
59	768	0.52	0.54	0.43
60	868	0.59	0.30	0.27
61	1024	0.69	0.23	0.23

62	623	0.42	0.18	0.17
63	919	0.62	0.36	0.32
64	418	0.28	0.27	0.26
65	356	0.24	0.10	0.10
66	364	0.25	0.10	0.10
67	612	0.41	0.40	0.37
68	816	0.55	0.38	0.31
69	763	0.52	0.46	0.38
70	312	0.21	0.22	0.23
71	398	0.27	0.22	0.22
72	1205	0.82	0.35	0.39
73	485	0.33	0.17	0.22
74	499	0.34	0.50	0.46
75	679	0.46	0.53	0.44

Appendix B

Rotated component matrix for a 3-component solution for the April 2003 administration of the FEPT. Principal components extraction and varimax rotation used. Extraction converged in 8 iterations. Cumulative variance accounted for: 13.83%.

Item	1	2	3
1	-0.04	0.37	0.25
2	0.00	0.51	0.3
3	-0.05	0.58	0.26
4	0.21	0.29	0.14
5	0.00	0.58	0.06
6	0.02	0.55	0.03
7	-0.03	0.51	0.05
8	0.03	0.55	0.03
9	0.00	0.50	-0.04
10	0.13	0.20	0.28
11	0.17	0.25	0.10
12	0.05	0.10	0.04
13	0.15	0.21	0.04
14	0.10	0.06	0.14
15	0.09	-0.18	0.27
16	0.23	-0.06	0.30
17	0.00	0.13	0.43
18	-0.15	0.13	0.45
19	-0.07	0.22	0.26
20	0.19	0.05	0.23
21	0.10	0.11	0.42
22	0.05	0.31	0.27
23	0.27	0.06	0.06
24	0.34	0.16	0.05
25	-0.05	0.06	0.39
26	0.16	0.01	0.22
27	0.11	0.16	0.2

28	-0.07	0.07	0.18
29	0.09	0.34	0.21
30	0.17	0.06	0.2
31	0.09	0.00	0.13
32	0.09	0.19	-0.01
33	0.08	0.08	0.05
34	0.06	0.20	0.41
35	0.05	0.06	0.24
36	0.02	0.04	0.40
37	0.04	0.05	0.07
38	0.09	-0.02	-0.05
39	0.10	0.12	0.27
40	0.22	0.05	0.25
41	0.43	0.00	0.21
42	0.34	-0.05	0.14
43	0.31	0.20	0.23
44	0.02	0.04	0.11
45	-0.02	-0.16	0.07
46	0.40	0.10	0.28
47	0.32	0.02	0.26
48	0.44	0.01	0.23
49	0.11	-0.03	0.18
50	0.17	0.00	0.20
51	0.34	0.04	0.25
52	0.19	-0.04	0.23
53	0.02	0.06	0.34
54	0.01	-0.03	0.13
55	0.28	0.27	0.18
56	0.26	0.15	-0.03
57	0.31	0.08	0.06
58	0.26	-0.10	-0.06
59	0.49	0.18	0.07
60	0.27	0.23	-0.03

61	0.12	0.15	0.14
62	0.13	0.05	0.06
63	0.22	0.10	0.21
64	0.33	0.04	0.03
65	0.14	0.02	-0.05
66	0.11	0.03	-0.05
67	0.40	0.17	0.07
68	0.28	0.00	0.24
69	0.36	0.12	0.16
70	0.39	-0.04	-0.01
71	0.29	-0.10	0.09
72	0.22	0.28	0.25
73	0.22	0.06	0.05
74	0.56	0.08	0.15
75	0.52	0.09	0.15