Defining and Tracking Student Performance on an EFL Placement Examination over a Three-year Period

Christopher Weaver

A placement examination can be an effective means of determining the level of English knowledge of new students entering an EFL program. Student performance on a placement examination can not only help ensure that students are placed into a class that best suits their current level of English knowledge, but also their test scores can be used to track the level of English knowledge of new students over time. Yet, underlying both of these benefits is the important issue of accurately defining students' current level of English knowledge.

Classical testing theory has established the practice of using raw test scores to define students' performance either directly or indirectly through the use of percentiles, t-scores, or z-scores (Brown, 1996). The use of raw scores has a number of merits. Raw scores are typically very easy to calculate. One simply counts the number of correct responses that a student has achieved on the placement examination and treats the sum total as a representation of the student's level of English knowledge. The student's test score can also be reported as the percentage of correct responses, which provides the additional information concerning the degree to which the student has mastered the content on the placement examination.

Another benefit of raw scores is that they are typically easy to understand and as such they can be readily used to make decisions concerning, for example, the placement of students in an EFL program. One approach might place students into a class with other students who have similar raw test scores. The end result would be a level system that divides students according to their current level of English knowledge as defined by their score on the placement examination.

Rasch measurement theory (Rasch, 1960/1980), however, challenges the assumption that raw test scores are a true reflection of students' level of knowledge. Rasch measurement theory argues that student performance on a test involves two interconnected parameters. The first parameter is the students' level of ability. The second parameter is the level of difficulty of the different items on the test. From the

Rasch measurement perspective, a statistical analysis of these two parameters produces a more accurate account of students' level of knowledge. Unlike raw scores, calculating students' level of knowledge is a more complicated procedure. The formula for dichotomous scored test items (i.e. items that are either scored as either being correct or incorrect) is:

$P_{ni}\ (x=1) = f(B_n\text{-}D_n)$

This formula states that the probability (P_n) of a student getting the correct response (x=1) for an item ($_i$) on an examination is the function (f) of the difference between a student's level of ability (B_n) and an item's level of difficulty (D_n). In other words, the Rasch measurement model defines a student's level of English knowledge as the probability of him or her correctly answering the different questions on the placement examination. The resulting probability score may initially seem to be very similar to the percentage of correct answers achieved on the placement examination. Yet, there is one important distinction between probability scores and percentages. The percentage of correct responses defines students' level of English knowledge as being the relationship between the number of correct answers and the number of questions on the placement examination. Probability scores, on the other hand, define students' level of English knowledge as being the interaction between the ability level of the students and the difficulty level of the items on the placement examination.

Thus, raw scores and probability scores both have their own distinct advantages and disadvantages when defining students' level of English knowledge. These advantages and disadvantages are also important considerations when comparing the level of students' English knowledge from one year to the next. The most significant challenge facing the use of raw scores to compare students' level of knowledge over a period of time is the lack of a common scale of reference. For example, if Student A receives a score of 30 on the placement examination one year and Student B receives the same score of 30 on the same placement examination the next year, it is difficult to say that these two students have the same level of English knowledge since they might have achieved the score of 30 in two entirely different ways. The potential for divergent paths arising from the same raw score highlights the importance of establishing the test items' level of difficulty before any comparisons can be made.

Rasch measurement theory addresses this issue by using a statistical technique called *anchoring* (Wright, 1996). This technique uses all of the students' responses

from multiple years to define the level of difficulty for each item on the placement examination. These item difficulty levels in turn are used to define the students' level of ability. Thus the shared item difficulty levels create a common frame of reference in which students' ability levels can be directly compared over a period of time.

Research Questions

The differences between defining students' level of English ability with raw test scores or probability scores creates a number of interesting research questions that have very practical implications for the use of a placement examination to stream students into an EFL program. The following research questions guide this investigation.

1. To what extent is there a difference between the probability of students correctly answering the different questions on the placement examination and the percentage of items that students correctly answered on the placement examination? And if they exist, what do these differences reveal about students' level of English knowledge?
2. To what extent is there a difference between the probability of students correctly answering the different questions on the different sections of the placement examination and the percentage of items that students correctly answered on the different sections of the placement examination? And if they exist, what do these differences reveal about students' level of English knowledge?
3. To what extent does the average probability of correctly answering the different questions on the placement examination and the average percentage of correct answers on the placement for the different departments at a university vary over a three-year period?

Method

Participants

This investigation involved the placement examinations of 2161 female students attending Jissen Women's University over a three-year period. The students included English and non-English majors who are required to take a semester long required English communication course. The course meets twice a week with one

meeting being taught by a group of native speakers of English using the textbook *Interchange* 1 from Cambridge University Press and the second meeting being taught by a group of Japanese teachers of English using the video workbook from the same series.

The placement examination

The placement examination is a commercially produced test originating from the Placement and Evaluation Package (Lesley, Hansen, & Zukowski-Faust, 2003) prepared for Cambridge University Press' *New Interchange series*. The examination has three sections. The first section focuses on listening. This section has 20 items that initially correspond to different spoken conversations involving two speakers. Yet as the student progress through the listening section, a greater number of items correspond to one spoken conversation. The length of the spoken interactions also increases in duration. Students have 15 minutes to complete the listening section of the placement examination.

The second section of the placement examination focuses upon students' reading skills. This section has 20 items. Similar to the listening section the number of items corresponding to each reading passage and the length of the reading passages increases as the students progress through the reading section of the placement examination. Students have 20 minutes to complete this section of the examination.

The third section of the placement examination focuses upon language use. This section has 30 discrete items designed to assess students' grammatical competence. Students have 15 minutes to complete this section.

All of the items on the placement examination are multiple-choice with four possible responses. Students write their answers on a mark sheet, which are machine-scored immediately after the placement examination. The examination has also been found to be a fairly reliable means of dividing students into two statistically distinct groups based upon their level of English knowledge (Weaver, Jones, & Bulach, 2007).

Procedure

The raw test scores for the 2161 students were first calculated for the three-year period. Then the average percentage of correct responses for each department was calculated for the entire test as well as the different sections of the placement

examination. The students' probability scores for the entire test and the different sections on the placement examination were calculated using three steps. First, all of the students' responses on the placement examination over the three-year period were used to determine the difficulty level of the different items on the placement examination. In the second step, these item difficulty levels were used to anchor the probability score calculated for each student. The final step involved calculating the average probability level for the entire test and the different sections of the placement examination for each department.

In 2005, the English Department did not take the placement examination. As a result, the average probability of success and the average percentage of correct responses on the placement examination as well as the different sections of the placement examination is not available (N/A) for the English Department for that academic year. In addition, students who wrote the placement examination more than once during the three-year period were grouped together under the category of reset students. The average probability scores and the average percentage scores for this group are based upon their responses on the placement examination the second time around.

Results

Student performance on the placement examination

Table 1 shows that the average probability of correctly answering the different questions on the placement examination and the average percentage of correct answers on the placement examination is quite close for the different departments over a three-year period. When differences exist, the average probability of correctly answering the different questions on the placement examination is slightly higher than the average percentage of questions answered correctly.

In terms of the average probability of correctly answering the different questions on the placement examination, some variability exists between the different departments. There are two basic patterns of variation over the three-year period. The first pattern involves a variable upward or downward movement. For example, in 2005 the average probability of success for Aesthetics & Art History students was 0.45. In 2006, their average probability of success increased to 0.48, but decreased to 0.44 in 2007. This pattern of variability is also apparent with Registered Dietician students and Human Sciences & Arts students.

Table 1
The average probability of correctly answering the different questions on the placement examination and the average percentage of correct responses on the placement examination for the different university departments over a three-year period

	2005		2006		2007	
Departments	Probability	Percent	Probability	Percent	Probability	Percent
1 Japanese Literature	0.45	45%	0.45	44%	0.44	42%
2 English	N/A	N/A	0.55	53%	0.53	47%
3 Aesthetics & Art History	0.46	45%	0.48	46%	0.44	44%
5 Registered Dietician	0.59	57%	0.55	53%	0.58	56%
6 Food Sciences	0.50	49%	0.46	46%	0.46	44%
7 Human Environmental Sciences	0.47	46%	0.45	44%	0.42	39%
8 Human Sciences & Arts	0.45	44%	0.47	46%	0.44	44%
Reset Students	0.42	41%				

The second pattern of variability found over the three-year period is a decreasing average probability of success. This downward trend varied in terms of severity amongst the different departments. The Japanese Literature and the English Departments, for example, had lower average probabilities of success in 2007 than in the previous years. Similarly, the Food Science Department's average probability of success was lower in 2006 and 2007 than its 2005 level. The Human Environmental Sciences Department was the only department that had a consistent decrease in the probability of correctly answering the different questions on the placement examination over the three-year period.

The average percentage of correct answers over the three-year period for the different departments mirrors the trends found with the average probability of correctly answering the different questions on the placement examination. There are, however, some notable differences between the average probability levels and the average percentage levels. For example, a drop of 0.02 in terms of probability of

success in 2006 and 2007 for English students was accompanied by a 6% decrease in the number of correct answers. However, a similar drop of 0.02 in terms of probability of success in 2005 and 2006 for Human & Environmental Science students was accompanied by only a 2% drop in the number of correct answers on the placement examination.

Student performance on the listening section of the placement examination

Similar to the results for the entire placement examination, Table 2 shows that the average probability for success in the listening section exceeds the average percentage of correct questions achieved by the students from the different departments over the three-year period. The relation between the average probability of success and the average percentage of correct responses on the listening section was relatively close. Moreover, the degree of variability over the three-year period for the different departments was uniform.

Table 2
The average probability of correctly answering the different questions on the listening section of the placement examination and the average percentage of correct responses on the listening section of the placement examination for the different university departments over a three-year period

	2005		2006		2007	
Departments	Probability	Percent	Probability	Percent	Probability	Percent
1 Japanese Literature	0.44	43%	0.45	44%	0.48	46%
2 English	N/A	N/A	0.54	51%	0.53	50%
3 Aesthetics & Art History	0.49	47%	0.51	48%	0.48	46%
5 Registered Dietician	0.55	52%	0.54	51%	0.54	51%
6 Food Sciences	0.44	43%	0.44	43%	0.47	45%
7 Human Environmental Sciences	0.45	44%	0.43	43%	0.45	44%
8 Human Sciences & Arts	0.43	42%	0.48	46%	0.44	43%
Reset Students	0.45	44%				

Two patterns of variability emerge from the average probability of success on the listening section over the three-year period. Similar to the entire placement examination, the average probability of success for some departments increased or decreased from year to year. Students in the English, the Aesthetics & Art History, the Registered Dietician, the Human Environmental Sciences, and the Human Sciences & Arts Departments exhibited this variable pattern of change. The second pattern of variability is quite distinct in that it does not exist in any other section of the placement examination or in the placement test in its entirety. The average probability of success for the Japanese Literature students on the listening section of the placement examination consistently increased over the three year-period from 0.44 in 2005 to 0.48 to 2007. The Food Science Department also had a 0.03 increase in their average probably of success in 2007 compared to the two previous years.

Student performance on the reading section of the placement examination

Table 3 shows that the average probability of success on the reading section of the placement examination exceeded the average percentage of correct responses for the different departments over the three-year period. However unlike the listening section of the placement examination, the difference between the average probability of success and the average percentage of correct responses in the reading section of the placement examination is slightly greater. For example, the difference between the average probability of success and the average percentage of correct responses on the listening section of the placement examination for the Registered Dietician Department was 0.55 and 52% in 2005. Yet, the difference between the average probability of success and the average percentage of correct responses in the reading section for the Registered Dietician Department was 0.60 and 52% in 2005.

The variability patterns within the reading section of the placement examination over the three-year period are similar to that of the entire placement examination. For the majority of departments, their students' average probability of success fluctuates from year to year either upwards or downwards. One notable fluctuation is the 0.07 decrease from 2005 to 2006 in Registered Dietician Department followed by the 0.10 increase in 2007. The average probability of success for the Japanese Literature and the Aesthetics & Art History Departments decreased slightly from 2005 to 2006, but remained the same in 2007.

Table 3
The average probability of correctly answering the different questions on the reading section of the placement examination and the average percentage of correct responses on the reading section of the placement examination for the different university departments over a three-year period

	2005		2006		2007	
Departments	Probability	Percent	Probability	Percent	Probability	Percent
1 Japanese Literature	0.46	40%	0.44	39%	0.44	38%
2 English	N/A	N/A	0.54	48%	0.56	50%
3 Aesthetics & Art History	0.46	40%	0.45	38%	0.45	39%
5 Registered Dietician	0.60	52%	0.53	47%	0.63	56%
6 Food Sciences	0.47	42%	0.49	42%	0.48	44%
7 Human Environmental Sciences	0.44	39%	0.47	42%	0.43	39%
8 Human Sciences & Arts	0.40	36%	0.46	39%	0.45	41%
Reset Students	0.42	37%				

Student performance on the language use section of the placement examination

Table 4 shows that the language use section of the placement examination has a number of unique characteristics that distinguish it from the other sections of the examination. Unlike the listening and reading sections of the placement examination, there are sixteen instances where the average percentage of correct responses exceeds the average probability of success. This reversal is most apparent with the Aesthetics & Art History and the Human Environmental Science students over the three-year period.

Table 4 The average probability of correctly answering the different questions on the language use section of the placement examination and the average percentage of correct responses on the reading section of the placement examination for the different university departments over a three-year period

	2005		2006		2007	
Departments	Probability	Percent	Probability	Percent	Probability	Percent
1 Japanese Literature	0.46	49%	0.44	48%	0.42	42%
2 English	N/A	N/A	0.59	59%	0.50	43%
3 Aesthetics & Art History	0.43	47%	0.48	50%	0.41	45%
5 Registered Dietician	0.63	62%	0.58	59%	0.59	60%
6 Food Sciences	0.55	57%	0.46	49%	0.45	44%
7 Human Environmental Sciences	0.49	52%	0.43	46%	0.36	37%
8 Human Sciences & Arts	0.49	52%	0.48	50%	0.42	47%
Reset Students	0.39	42%				

Over the three-year period, two patterns of variability are evident. The first is a variable pattern in which the average probability of success or the average percentage of correct responses on the language use section of the placement examination fluctuates either upwards or downwards over the three-year period. The Aesthetics & Art History and the Registered Dietician Departments exhibit this pattern of variability. The remaining six departments, however, have a trend of decreasing averages in both the probability of success and the percentage of correct responses on the language use section of the placement examination over the three-year period. The rate of decline varies considerably amongst these six departments. The decline in the Japanese Literature Department, for example, has been fairly consistent from one year to the next. The Human Environmental Sciences and the Human Sciences & Arts Departments have also seen a consistent decline over the three-year period, but with larger incremental drops. The remaining two departments have seen a substantial drop in the average probability of success in

one of out the three years. The average probability of success in the English Department dropped from 0.59 in 2006 to 0.50 in 2007; while, the Food Science Department had a similar drop in the average probability of success from 0.55 in 2005 to 0.46 in 2006.

Discussion

The relationship between the average probability of success and the average percentage of correct answers on the placement examination

Probability and percentage scores reveal slightly different information about student performance on the placement examination, which influences the types of inferences that can be made about students' level of English knowledge. Overall, there is a fairly harmonious relationship between the average probability scores and the average percentage scores for the entire test. As Table 1 shows, the average probability scores are slightly higher than the average percentage scores. This difference means that the students' level of English knowledge is slightly higher than what the percentage scores suggest. Once again, the difference between probability scores and percentage scores is that probability scores take into account the difficulty level of the different items on the placement examination. The seeming parity between the probability scores and the percentage scores also suggests that raw test scores are a fairly reasonable and efficient means of making placement decisions. However, the lack of a common frame of reference seriously undermines the use of raw test scores to compare students' level of English knowledge over time.

The fairly close relationship between probability scores and percentage scores is less apparent when the focus of the investigation turns to the different sections of the placement examination. Only the listening section maintained a close relationship between probability and percentage scores. The impact of item difficulty becomes more apparent in the reading and the language use sections of the placement examination. For example in the reading section, the 2005 Japanese Literature students had a considerably higher average probability score compared to their average percentage score. This difference means that although these students on average correctly answered only 40% of the questions on the reading section, they were successful on questions that had a high level of difficulty, which in turn resulted in the higher average probability score of 0.46. In the language use section of the placement examination, there are a number of instances where the relationship

between the average probability scores and the average percentage scores has reversed. In the case of the 2007 Human Sciences & Arts students, they correctly answered 47% of the language use questions on average, but the level of difficulty of these questions was relatively low, which produced the lower average probability score of 0.42.

Thus while the relationship between probability scores and percentage scores is fairly close for the placement examination as a whole, disparities between the two types of scores on the different sections of the placement examination are an important consideration, especially if students' performance on the sub-tests is going to be factored into placement decisions. The overarching issue is whether or not placement decisions should be based upon the number of correct responses or the number of correct responses factoring in the difficulty level of the questions. An interesting research topic for a future investigation would be the extent to which students are placed in different classes based upon their probability scores versus their percentage scores.

The patterns of variability in students' level of English knowledge over the three-year period

As previously explained, probability scores are the only reliable means of tracking students' level of English knowledge over time. Over the three-year period, three patterns of variability emerged. The most common pattern was the up-and-down fluctuations of average probability scores from year to year. Since the present investigation only examined three years of placement examinations, it is difficult to determine whether or not these fluctuations are part of a larger pattern of change. Ideally, this investigation should serve as a foundation for annual reviews of student performance on the placement examination. An investigation involving a longer period of time would provide a more reliable account of the trends that emerge from the students' performances on the placement examination.

The second most common pattern of variability was the consistent decrease in the students' average probability of success. The rate of this decrease varied significantly amongst the different departments and the different sections of the placement examination. Although a decrease in the students' level of English knowledge might be an expected by-product of the changing demographics of the university student population in Japan, careful attention should be paid to detecting sudden drops in students' level of English knowledge. These drops pose a significant

challenge for an EFL curriculum. For example, the language use section of the placement saw the largest drops in students' level of knowledge over the three-year period. Out of the eight departments, five of them had consistently decreasing average probability scores with the English Department seeing a 0.09 drop from 2006 to 2007. This decreasing level of English grammatical knowledge suggests the need to highlight this area of communicative competence in the Jissen University EFL curriculum.

The third pattern of variability was the year-on-year increase in the students' level of English knowledge. The Japanese Literature students' performance on the listening section of the placement examination was the only example of this upward trend. This increase might be the result of an increasing shift in the high school English curriculum to develop communicatively competent students (Ministry of Education Culture Sports Science and Technology, 2003) and/or a greater exposure to Assistant Language Teachers, who use English as their first language.

Conclusion

The purpose of this empirical investigation was to clarify the relationship between probability scores and raw test scores. At the test level, the two types of scores are fairly similar. This parity suggests that raw test scores, which are easier to calculate, are a reasonable standard for making placement decisions. Yet, the disparity between the probability scores and the percentage scores on the different sections of the placement examination suggests that the continued close relationship between these two types of scores at the test level should not be assumed. Continued monitoring of students' performances would not only address this issue, but also provide a more reliable account of the different trends that emerge from the students' performances on the placement examination. In addition, the next logical step would be to investigate the relationship between the placement examination and the final test that students write once they complete the required English course. This type of investigation would be a valuable means of evaluating the performance of the EFL program at Jissen Women's University as long as there is a common frame of reference between the two examinations.

References

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

Lesley, T., Hansen, C., & Zukowski-Faust, J. (2003). *New Interchange: Passages Placement and Evaluation Package.* Cambridge: Cambridge University Press.

Ministry of Education Culture Sports Science and Technology. (2003). *Regarding the establishment of an action plan to cultivate "Japanese with English abilities".* Retrieved April 1, 2004, from www.mext.go.jp/english/topics/03072801.htm

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Copenhagen, Danish Institute for Educational Research, expanded edition (1980) with foreword and afterword by B. Wright ed.). Chicago: The University of Chicago Press.

Weaver, C., Jones, A., & Bulach, J. (2007). *Customizing a commercially produced placement examination.* Paper presented at the JALT 2007, Tokyo.

Wright, B. (1996). Time 1 to Time 2 Comparison. *Rasch Measurement Transactions, 10,* 235-251.