Kyoei University English Placement Test Assessment

共栄大学プレースメントテストに対する評価

Steven LLOYD & Nicholas A. BUFTON
スティーブン　ロイド・ニコラス　A　バフトン

Abstract

This paper discusses the development, and analysis of the newly developed placement test for the 2010 International Business Management Department (IBMD) first year intake. It acknowledges that while a small number of items may need further examination or improvement, the statistical analysis provides solid empirical evidence about the performance of the test. It is clear that the test is suitable for the purpose for which it was designed and that the satisfactory reliability and validity evidence gathered proves that it is a dependable test. Moreover it demonstrates that this test can be used with confidence now, and as a model on which to build future versions.

Keywords: placement test, reliability, test analysis

概要

この論文は、初めての試みである 2010 年共栄大学プレースメント・テストにおける分析を主旨とする。このテストは、依然として多くの微調整や課題を要するものの、その統計的な分析の結果、その実践における確固たる有効性を証明することができた。このテストは、プレースメント・テストとしては、その信頼性においてもその目的を果たすものして適切なものであるということがこの論文における統計分析でよって証明ができたと言える。さらにこのテストをモデルとして、今後のプレースメント・テストにおける将来的な展望が大いに開けたと言えよう。

キーワード：プレースメント・テスト、信頼性、テスト分析

1 Background

The International Business Management Department's (IBMD) English Program at Kyoei University (KU) has been streaming incoming students since 2000. This streaming allows KU to place students in classes best suited to their current level of English proficiency. This is especially important when one considers the ethnicity and academic diversity of the student body. The English proficiency of each new intake ranges from the inability to read beyond that expected of a third grade junior high school student, to those who have TOIEC© scores of 600 and above. Until 2010 the program had used the commercially available General Tests of English Language Proficiency (G-TELP©) Level 3 test as the main method of assessing new students' EFL ability. However, from 2005 it was noted that the distribution of test scores were beginning to produce positively skewed distributions. The data showed an overall fall in the English proficiency of incoming students. This trend continued until the test could no longer provide data by which we could reliably assess and stream students in order to place them in a class appropriate to their current level of proficiency.

Although a commercial test made life easier for the university staff as the vendor undertook marking and data compilation, it was becoming expensive especially as we required the results quickly. In addition, the time needed to assess and adjust the results was increasing and as a result the delays in producing student placement lists for class assignment was increasingly putting pressure on both educators and administrators alike. Thus work began in 2009 to develop a test that would allow us to (1) gauge the incoming students' knowledge of what they should have learnt at high school, and (2) more accurately stream and place students according to our curriculum requirements.

2 Test Development

There are advantages and disadvantages to creating and administering a tailor-made test. Obviously the disadvantages are the time and effort required in constructing and marking a placement test, and risk of producing a test that provides no better data than the one it is replacing. On the other hand, the advantages of creating and administering our own placement test are manyfold. The main advantage though is that the creators have detailed knowledge of the KU English program and syllabuses, and accordingly can identify the key features of the different levels within the English program. Thus the considerable investment of time constructing an 'in-house' test should be rewarded by more accurate placement, and the time saved between administering, marking, adjusting and placing students. In addition to facilitating streaming of the students, the detailed data collected

will allow us to examine the efficacy of each item, the distracters and the test as a whole. This will also provide us with the required information to remove any extraneous items, refine questions and improve the placement process.

Before constructing the placement test, four high school English textbooks authorized by the Japanese Ministry of Education, Culture, Sports, and Technology (MEXT) were chosen for review. All texts chosen complied with either the Aural/Oral Communication II or the English II section of the *Course of Study for Foreign Languages* revised guidelines, 2003.

In Japan, authorized textbooks are used to support public high school English language learning. These authorized textbooks cover the essential elements of the MEXT *Course of Study for Foreign Languages* guidelines and thus are good guides to what students should have covered during their high school English language studies.

There is a range of textbooks that follow the MEXT guidelines at various depths of detail, ranging from challenging to less challenging. Although high school English textbooks authorized by MEXT are sold throughout Japan, certain publications tend to be more popular in different areas of the country. After informal interviews with students at Kyoei University we decided on 4 textbooks to inform us on the basic vocabulary and the general level of reading difficulty that freshman students were likely to have been exposed to. The textbooks examined were: *Crown English Series II* (Sanseido, 2006) *Sunshine English Course II* (Kairyudo, 2006) *Expressways - Oral Communication I* Advance Edition (Kairyudo, 2006) and *Planet blue — Oral Communication I* Revised Edition (Yokyo: Oubunsha Press, 2006).

From the aforementioned textbooks two lists of common vocabulary were compiled. One list of vocabulary common to both *Crown English Series II* and *Sunshine English Course II,* and one list of vocabulary common to both *Expressways - Oral Communication I* and *Planet blue — Oral Communication I* were produced. Furthermore several lessons from both *Crown* and *Sunshine* were sampled to ascertain Flesch Reading Ease and Flesch-Kincaid Grade Level scores. The Flesch Reading Ease and Flesch-Kincaid Grade Level scores for *Crown* were 68.9 and 7.6 respectively, and 71.5 and 6.9 for the *Sunshine* textbook. Samples were also taken from *Expressways* and *Planet blue,* though it should be noted that both of these were written for oral communication and so the ease of reading scores varied considerably between samples. The average scores for the selections were: *Expressways,* 82 for the Flesch Reading Ease and 4.3 for the Flesch-Kincaid Grade Level, and 74.5 and 5.3 respectively for *Planet blue*.

While the use of readability formulas to match texts to student reading levels and the accuracy of such methods has been discussed in several publications (Brown, 1993, 1998; Chall & Dale,

1995; Fry, 1989) and that the performance criteria for English as a Foreign Language (EFL) readability has not yet been resolved, (Greenfield, 2004) the Flesch Reading Ease and Flesch-Kincaid Grade Level scores still remain reasonably reliable for ascertaining the reading difficulty of text. Moreover they are invaluable when setting a benchmark for the construction of an English language placement test.

A short beta version test was made and given to the previous year's intake of students. The results were analysed and were found to be similar to their earlier G-TELP© results. This beta version of the test was expanded and improved upon, resulting in the current placement test, which was given to the students in March 2010.

Like all tests, the KU English Placement Test administered in 2010 was a product of a compromise between available time, resources and objectives. The time available between administering the test, hand marking, analysing the results and submitting the resulting class lists to administration in time for the first week of term was three days.

The sixty-minute test is a norm-referenced test, (NRT) and made up of ninety items over three sections: grammar, reading and listening. This is the standard method used for placement tests, as the objective is to group the students by ability, rather than to pass or fail them on specified criteria.

3 Test Analysis

The following analysis — and the placement of the students — was only applied to 141 of the 200 students who took the test. The remainder, although taking the test, were pre-assigned a class on the basis of other academic and non-academic commitments regardless of English ability. In addition, for the purposes of this paper, the results of students scoring 25% or less (as most of the items had four options to choose from) were also removed on the basis that their English ability was so low that any answers they gave would largely be random and so give false results. This left us with N=125. The research found that the test was effective and that the placement of students was generally successful. The research and hence the terminology used in this paper will follow J. D. Brown's *Testing in Language Programs* (2005), and was carried out using the Apple *Numbers '09* spreadsheet application.

All Scores	Scores	Percentages
N	141.00	
k (item count)	90.00	
Mean	40.51	45.01%
Mode	22.00	24.44%
Median	40.00	44.44%
Midpoint	49.50	55.00%
High	84.00	93.33%
Low	15.00	16.67%
Range	70.00	
Standard Dev	14.70	

Table 1a Statistics for all students

Scores over 25%	Raw	Percentages
N	125.00	
k (item count)	90.00	
Mean	43.10	47.88%
Mode	43.00	47.78%
Median	42.00	46.67%
Midpoint	53.50	59.44%
High	84.00	93.33%
Low	23.00	25.56%
Range	60.00	
Standard Dev	13.58	

Table 1b Statistics for students scoring over 25%

As the differences in measurements of the central tendency in the above tables indicate, there is a positive skew in the results. The high standard deviation and range indicates the wide difference in ability of the students taking the test (and highlights the need for assigning the students to appropriate classes). The positive skew is indicative of the large number of low ability students. As this was not a pass-fail test, the low average score is disappointing rather than a problem. This is best illustrated with a figure of the frequencies:

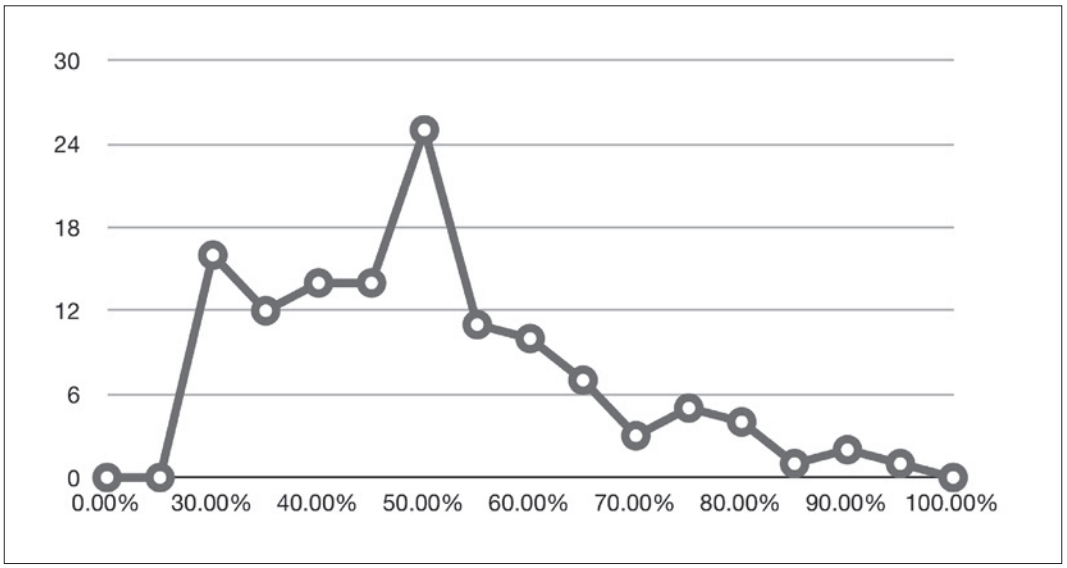

Figure 1 Frequency of scores for students scoring over 25%

In addition to the positive skew, Figure 1 also shows a bimodal distribution (sixteen students at 30% and 25 students at 50%).

Breaking the test into its constituent parts, the following results are produced:

			Sub-tests					
Statistics	Total	%	Grammar	%	Reading	%	Listening	%
N	125.00		125.00		125.00		125.00	
k	90.00		70.00		5.00		15.00	
mean	43.10	47.89%	33.97	48.53%	2.26	45.28%	6.87	45.81%
median	42.00	46.67%	32.00	45.71%	2.00	40.00%	7.00	46.67%
mode	43.00	47.78%	32.00	45.71%	3.00	60.00%	5.00	33.33%
midpoint	53.50	33.89%	26.50	37.86%	2.50	50.00%	6.50	43.33%
range	60.00		52.00		4.00		12.00	
high	84	93.33%	67.00	95.71%	5.00	100.00%	13.00	86.67%
low	23.00	25.56%	14.00	20.00%	0.00	0.00%	0.00	0.00%
s	13.57	15.08%	11.61	16.59%	1.23	24.68%	2.46	16.42%

Table 2 Statistics for overall results and sub-tests for students scoring over 25%

As Table 2 and Figure 2 show, on average the students are stronger in grammar than reading and listening. There were a small number of test-takers (11 students) who did not answer any of the reading items correctly — one of whom also failed to answer any of the listening items correctly. After examining the individual test papers, it is evident that some of these test-takers didn't even attempt to answer these items at all. However, these students were all good enough at the grammar questions to score more than 25%. *Vide infra* for further discussion.

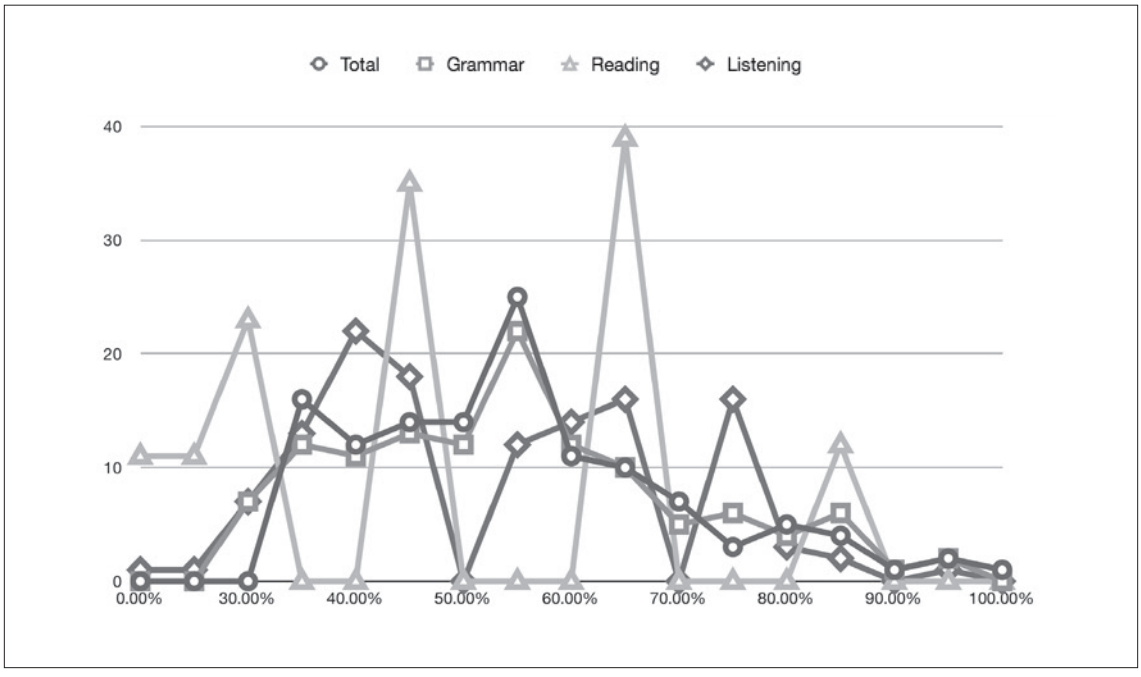

Figure 2 Frequency of scores for students scoring over 25% by sub-test

Figure 2 shows that grammar corresponds with the total scores. The peaks and troughs in the reading scores are to be expected given that there are only five items.

There are various statistical tools available for examining the internal consistency and reliability of the test. Firstly, the Spearman-Brown prophecy formula ($r_{xx'}$) is applied to a Pearson product-moment correlation coefficient (r), which is calculated using the split-half method. To calculate Pearson r, the following formula is applied:

$$r_{xx} = \frac{\sum (X_{odd} - M_{odd})(X_{even} - M_{even})}{NS_{odd}S_{even}}$$

r_{xx} = Pearson product-moment correlation coefficient
X_{odd} = each student's score for odd-numbered items
M_{odd} = mean of odd-numbered items
S_{odd} = standard deviation of odd-numbered items
X_{even} = each student's score for even-numbered items
M_{even} = mean of even-numbered items
S_{even} = standard deviation of even-numbered items
N = number of students

$$r_{xx} = \frac{5{,}351.68}{6{,}154.79} = .87$$

The result of 0.87 indicates a high correlation between the two halves of the exam. There is a 99% certainty that this number is statistically significant (Brown, p.155). The coefficient of determination (that is to say r_{xx}^2) is .76, and the error variation (1- r_{xx}^2) is .24. In other words, 76% of variance in the results is shared by both halves of the tests, and only 24% is unique or due to random chance. Given that the results are somewhat skewed, the correlation coefficient given is probably lower than is actually the case. This all strongly suggests a good and meaningful correlation between the two halves of the exam.

This coefficient was inserted into the Spearman-Brown prophecy formula to calculate the reliability of the test as a whole.

$$r_{xx'} = \frac{(n)r}{(n - 1)r + 1}$$

$r_{xx'}$ = full-test reliability
r = correlation between the two test halves (ie. r_{xx})
n = number of times the test length has to be increased

$$r_{xx'} = \frac{1.74}{1.87} = .93$$

A coefficient of .93 suggests a very high degree of reliability and to confirm this, the Cronbach alpha formula, the Kuder-Richardson formula 20 (K-R20) and Kuder-Richardson formula 21 (K-R21) were also calculated. The Cronbach alpha coefficient is also calculated using the split-half method of odd and even-numbered items.

$$\alpha = 2\left(1 - \frac{S^2_{odd} + S^2_{even}}{S^2_{total}}\right)$$

α = Cronbach alpha coefficient
S_{odd} = standard deviation for odd-numbered items
S_{even} = standard deviation for even-numbered items
S_{total} = standard deviation for the total test scores

$$\alpha = 2\left(1 - \frac{98.50}{184.13}\right) = .93$$

A coefficient of .93 is again extremely high. The Kuder-Richardson formulas tend to be more conservative, so a lower coefficient can be expected. The K-R20, the less conservative of the two estimates, uses the concept of item variance.

$$\text{K-R20} = \frac{k}{k - 1}\left(1 - \frac{\sum S_i^2}{S_t^2}\right)$$

K-R20 = Kuder-Richardson formula 20
k = number of items
S_i^2 = item variance
S_t^2 = test score variance

$$\text{K-R20} = 1.0112\left(1 - \frac{19.50}{184.13}\right) = .90$$

A coefficient of .90 is a respectable result. The K-R21 formula generally produces a more conservative coefficient, but .89 still suggests that the test is reliable.

$$\text{K-R21} = \frac{k}{k-1}\left(1 - \frac{M(k-M)}{kS^2}\right)$$

K-R21 = Kuder-Richardson formula 21
k = number of items
M = mean of test scores
S = standard deviation of test scores

$$\text{K-R21} = 1.0112\left(1 - \frac{2{,}021.40}{16{,}571.26}\right) = .89$$

Bearing in mind that all of these calculations of reliability produce underestimates, even the lowest of .89 strongly suggests the test is very reliable.

To verify and ensure reliability and consistency, the standard error of measurement (SEM) was also calculated using the following formula:

$$\text{SEM} = S\sqrt{1 - \text{K-R21}}$$

SEM = standard error of measurement
S = standard deviation of the test
K-R21 = Kuder-Richardson formula 21

$$\text{SEM} = 13.57\sqrt{.1121} = 4.54$$

This coefficient indicates that, on average, a student will score within a band of plus/minus 4.54 points around the actual score achieved. That is to say, a student scoring 70 could quite easily have scored between 66 and 75 points (± 4.54). This is a margin of about ±5%, which we estimate to be acceptable for our purposes.

Table 3 summarizes the reliability estimates and the SEM for each:

	Reliability estimate	SEM
Split-half adjusted by Spearman-Brown	0.93	3.58
Cronbach α	0.93	3.59
K-R20	0.90	4.20
K-R21	0.89	4.54

Table 3 Reliability estimates and standard error of measurement

All the reliability estimates are high, suggesting the test can be considered to be reliable (Hughes, 1989; p.39). For example, the KR-20 for the TOEIC is around .95 (ETS IV-1), so a reliability of .90 on the first attempt of the KU placement test is acceptable. The standard errors of measurement are not as narrow as we would like (±3.98% - ±5.0%) and suggest room for improvement, but are more than acceptable for the purpose of placing our students.

4 Discussion

The strong reliability estimates reported in Table 3 suggest that the test is internally consistent and reliable, and accurately measures the English abilities of the students. The positive skew of the results is a disappointment, but this arguably is a reflection on the level of the students' ability. As described above, the test was made after close examination of high school textbooks, and no new grammar, vocabulary or other structures were added into the test.

As mentioned above, eleven students failed to score any of the reading items correctly, and one failed to answer a single listening item correctly. Although this data is not reported above because they failed to attain an overall score of 25%, thus not included in the analysis, but only another three students scored zero for reading, and an additional student also failed to score any listening items correctly. There are a variety of possible reasons for this, as well as the general low scores in the listening and reading sections. Firstly, some items may be too difficult or badly designed leading to the test-taker panicking briefly. Secondly, the students' ability was too low, which might be a consequence of either a failure on the part of the students to study, or of their high schools to provide them with adequate instruction and learning opportunities as required by MEXT. Either way, it has been agreed that the next placement test should include expanded listening and reading subtests, starting at either *Eiken* STEP TEST© grade 3 or possibly grade 4.

Several issues with the test need to be addressed. Some items will have to be re-written, and some replaced with easier items aimed at lower level students before administering the test in 2011. In particular the reading and listening sections will be expanded and the number of grammar items will be reduced to keep the test within 60 minutes.

5 Conclusion

As the above analysis strongly suggests, the first attempt at designing, constructing and validating an in-house English placement test for the Kyoei University International Business Management Department has proven successful. While it is noted that some items may need further exami-

nation, the study nonetheless provides solid empirical evidence about the performance of the test. It is clear that the test is suitable for the purpose for which it was designed and that the satisfactory reliability and validity evidence gathered proves that it is a good test. As this test clearly stands on its own, it can be used with confidence as a model on which to build future versions.

References

Brown, J. D. (1993). What are the characteristics of natural cloze tests? *Language Testing*, 10, 93-116.

Brown, J. D. (1998). An EFL readability index. *JALT Journal*, 29, 7-36.

Brown, J. D. (2005) *Testing in Language Programs.* McGraw-Hill

Chall, J., & Dale, E. (1995). *Readability revisited: The New Dale-Chall Readability Formula.* Cambridge. MA: Brookline Books.

Fry, E. (1989). Reading Formulas: Maligned but valid. *Journal of Reading*, 32, 292-297.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psycholgy*, 32, 221-233.

Greenfield, G. (2004). Readability Formulas For EFL. *JALT Journal*, 26, 5-24.

Kincaid, J. P., Fishbourne, R. P., Jr, Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability Formulas (Automated Reading Index, Fog Count and Flesh Reading Ease Formula) for navy enlisted personel. Research Branch Report 8-75. Millington, TN: Naval Technical Training, U.S. Naval Air Station, Memphis.

EIKEN Test in Practical English Proficiency, STEP (The Society for Testing English Proficiency, Inc.)

ETS (2010) *Technical Manual.* retrieved from http://www.google.com/search?source=ig&hl=en&rlz=1G1GGLQ_ENJP275&=&q=toeic+cronbach+alpha+&btnG=Google+Search&aq=f&oq=#sclient=psy&hl=en&rlz=1G1GGLQ_ENJP275&q=toeic+SEM&aq=f&aqi=&aql=&oq=&gs_rfai=&pbx=1&fp=269988b335e72e7d

Hughes, A. (2003) *Testing for Language Teachers.* (2nd Edition) Cambridge UP

Japanese Ministry of Education, Culture, Sports, and Technology (MEXT) (2003) *Course of Study for Foreign Languages.* retrieved from http://www.mext.go.jp/english/shotou/030301.htm

Takahashi, S. *et al* (2006) Crown English Series II. Sanseido.

Sanno, M., Yamaoka, T., Matsumoto, S., Satou, Y. (2006) Sunshine English Course II. Kairyudo.

Kobayashi, K., House, J.c., & Mitsui, T. (2006) *Expressways - Oral Communication I* (Advance Edition). Kairyudo.

Negishi, M., Yoshitomi, A., Kanou, M., shizuka, T., & Takayama, Y. (2006) *Planet blue ― Oral Communication I* (Revised Edition). Tokyo: Oubunsha Press.

Samdahl, D. M. (2010) *Introduction to Statistics*. retrieved from http://www.coe.uga.edu/~dsamdahl/Intro_to_Stats.doc