Paper

Placement testing for three university departments

Robert G. JAMES

Abstract

This study probed whether a single placement test used for streaming students in three departments at Jobu University was equally appropriate and effective for each group of students.

Samples of student answers on a common placement test from each department were compared by means of ANOVA to ascertain whether they constituted one or more groups in terms of proficiency. Further, an item analysis of the test was performed, to ascertain the difficulty of the test for each group, and also judge how well the test discriminated between strong and weak students in each department. These latter objectives are part of an ongoing project to improve placement testing at the institution.

Results of the ANOVA study suggest that there are only two groups of students across the three departments: Business and Information Sciences students on the one hand, and Nursing students on the other. Item analysis suggested that the test is of an appropriate difficulty level for the former group, while being overly easy for the latter. Additionally, the analysis suggested that for the higher-scoring group, the test appears to discriminate less well between strong and weak students.

Based on these conclusions, the item analysis provides a starting point for the development of more focused placement instruments, should the decision be made to use separate test for the two student population.

Key words and phrases

item analysis, language testing, placement, reliability, streaming

（Received 8 May 2009）

1. INTRODUCTION

Jobu University English Center provides lessons in English to students in three departments: Information Sciences, Business and Nursing. At the beginning of each academic year these students are streamed into proficiency levels by means of a placement test, and teachers in the three departments assigned to the various levels decide on course

content based on these scores as well as other considerations. The placement instrument used for the three departments, by tradition, is the same.

Teachers of nursing students agree that these students are at a generally higher proficiency level that the other two management-oriented departments. This is not so surprising. Nursing courses in the tertiary education system are not particularly common, since they require colleges to invest heavily in specialized equipment; and many of the courses that do exist are offered by junior and vocational colleges. A four-year university that offers nursing courses can compete very effectively with such institutions. On the other hand, management and computer courses are offered by many institutions in both the two-year and four-year sector, and competition for students is much stiffer. In short, our school can be selective with nursing applicants while it has to accept more or less all business department applicants, and this is reflected in the "quality" of the students. The nursing students are more proficient in English, seem more committed (in terms of class attendance rates) and work harder (in terms of class preparation and homework), than their colleagues in the management departments.

However, teacher intuitions about students' levels are usually based on impressions of communicative ability rather than measures of linguistic competence, and no formal cross-departmental measure of the linguistic abilities of the students had yet been conducted. This study was a first attempt to furnish such information.

The study was designed to address two research questions provoked by the streaming procedure mentioned in the first paragraph. Firstly, do the three departments reflect three different student populations from the point of view of English teaching, or are they sufficiently similar to be considered a single group? And secondly, to what extent is a single placement instrument appropriate for all three groups? The answers to these questions could determine whether the English Center continues to use a single placement instrument or develops more specialized instruments for one or more of the student groups.

To address the first question, the null hypothesis of no statistically significant difference between performances on the test for the three departments was tested by means of the one-way ANOVA statistical procedure. To address the second, reliability coefficients and Standard Error of Measurement statistics were calculated for the test. These statistics can indicate the usefulness of a test in terms of streaming into level bands. Additionally, for the purpose of both test improvement, and as first step in the potential development of additional instruments, item analysis was conducted on each item in the test. It was

believed that these more detailed statistics would provide more precise pointers to ways in which the groups differed, and thus what a more specialized test instrument might look like.

2. METHOD

Students from the three departments were given a placement test at the beginning of the 2008 academic year, and from the answer sheets, samples from each department were culled. Students in the Business and Information Sciences departments were predominantly male, while those in the Nursing department predominantly female. Additionally, about 20% of the student in Business and about 10% of students in the Information Sciences were non-Japanese Asians. All students in the Nursing department were Japanese. A small number of non-Japanese students who indicated that they had learned no English at all during their school years were omitted from the study.

A stratified sample of answer sheets was culled from the population of test-takers by ordering the sheets from each department by total score, and taking every Nth sheet for use in the study. To create broadly similar sample sizes from widely differing department enrollments, every second sheet from the Nursing department was used in the sample, every third sheet from Information Sciences was used, and every eighth sheet from the much larger Business department was used. Table 1 shows the number of students taking the placement test from each department in 2008, and the proportion used to create the samples.

Table 1: Test-taking population and sample size

Department	Test takers	Nth sheet used	Sample size
Nursing	67	2	33
Information Sciences	87	3	29
Business	256	8	32
Total	410		94

For the purpose of analysis, answer sheet data were inputted into the computer and an initial set of descriptive statistics generated. Microsoft Excel™ and an add-in statistical package (Analyse-it™) were used for all calculations. Table 2 shows sample characteristics as a whole and for each department.

Table 2. Descriptive statistics for total sample and each department

Statistic	Total Sample	Info. Sc. dept.	Business dept.	Nursing
Mean	61.3	55.4	51.8	75.8
Median	62	54	51	78
Mode	62	46	62	84
St. Dev.	17.8	15	15.8	12.2
Skew	-0.1	0.1	-0.5	-0.1

The descriptive statistics in Table 2 for the total sample show close similarity for mean, median and mode, suggesting a fairly well-formed distribution, though with a slight negative skew. The score distribution for the individual departments is not so smooth, in particular the modal score differs markedly from the other two measures of central tendency, suggesting one or more sub-groups within each department. Nevertheless, it was decided that the sample conformed adequately to the requirements of normative test analysis, and the assumptions required for the analyses conducted.

The test instrument used was the English Center's placement test. This instrument has been developed over a number of years (see James, 2007), and contains 50 multiple-choice items, grouped into three sub-parts: grammar (20 items), listening (15 items), and vocabulary (15 items). Given the general level of proficiency of the students that this university typically attracts, items for the test reflect grammar and vocabulary found in elementary and intermediate general English course books.

A reliability coefficient for the test (KR-20) was calculated based on the total sample of 94 answer sheets. Since the test is used primarily for streaming purposes, a Standard Error of Measurement (see Brown 2005, p.188) was also calculated. The result of this calculation was a KR-20 score of .88 and SEM of 3.07. The reliability coefficient was judged to indicate adequate reliability for the test, although for placement purposes the SEM would ideally have been smaller.

To test the first hypothesis of no relationship between the three group scores, a one-way ANOVA design was used to compare group means. A significance level of .95 was set.

To address the second issue of how performance on the test for each group differed, an item analysis was conducted for each of the 50 items on the test. An item facility (IF) score and a discrimination index (DI) score was calculated for each item. The former score gives an indication of the difficulty or ease of a particular item, since it represents the percentage of students who correctly answer the question. An item facility score of 1 represents 100% of subjects answering the question correctly. Traditionally, an IF of 0.5 is regarded as ideal, indicating half the subjects answered correctly and half

incorrectly. The discrimination index indicates how well an item discriminates between strong and weak students, as reflected by students' total scores. A DI of 1 indicates that all those students defined as strong (the upper third of students by total score) answered the question correctly and none of the weak subjects (defined as the lowest third of the subjects by total score) answered the item correctly. An ideal DI score would thus be 1. Brown (2005, p.75) cites Abel (1979, p.267) suggestion that acceptable scores for DI range down from 1 to a minimally acceptable low score of 0.2. Scores below this figure or even a negatives score (indicating that more weak subjects answered a question correctly than strong ones) are a signal that there is something confusing about an item, and it should be replaced. DI scores higher that 0.2 are acceptable but for scores lower than 0.4, effort should be made to improve the item if the test is reused later.

Average IF and DI scores for the items in each sub-section of the test (grammar, listening and vocabulary) were calculated, and these scores for each department were scrutinized to see how the three groups differed in their performance in the three skill areas.

The test was administered on the same date and time for the two management-oriented departments, and the following day for the Nursing department. While it would be practically possible for collusion to occur between the earlier and later test takers, this was not regarded as likely, due to the lack of social interaction between the groups at this early stage in their university career, the separate buildings used for the respective department's classes, the low priority given to the placement test within the overall curriculum (when compiling class lists, English language placement results can be overridden for a variety of administrative reasons), and no prior knowledge by the students that the same test would be used.

3. RESULTS

The ANOVA study results are shown in Table 3. As these results show, the null hypothesis of no difference between the three group means could be rejected.

Table 3. Significance of difference between means for the 3 departments

Sources of variation	SSq	DF	MSq	F	p
Department	2698.67	2	1349.33	26.11	<0.0001
Within cells	4702.44	91	51.68		
Total	7401.11	93			
n = 94					

Table 4 below shows the results of a further Scheffe test to isolate the source of difference. That test showed that the difference between the average score for the Business department and that of the Information Sciences department was not statistically significant, while the difference between the Nurses and both management departments was statistically significant.

Table 4. Scheffe test of source of significance between means

Contrast	Difference	95% Confidence Interval	Comment
Business v Info. Sc.	-1.78	-63.67 to 2.8	not significant
Business v Nursing	-11.97	-16.41 to -7.53	significant
Info. Sc. v Nursing	-10.19	-14.74 to -5.64	significant

These results show that the population of 410 students taking Jobu University's Placement test in April 2008, while assigned to three different departments, was, in terms of English proficiency, only two groups: a management-oriented group on one hand (comprised of business and information science students), and a nursing group on the other.

Given this conclusion, it made sense to conduct the follow-up item analysis in terms of the two groups identified. The item analysis scores were considered in terms of the three sub-tests that comprised the test instrument. Table 5 shows the results, which are broken down into average scores for each of the two groups isolated by the ANOVA study.

Table 5. Average IF and DI scores on test sub-sections for the two proficiency groups

	Business/Information Sciences		Nursing	
Placement test Sub-sections	Item Facility	Discrimination Index	Item Facility	Discrimination Index
Grammar	.55	.37	.77	.23
Listening	.58	.36	.79	.14
Vocabulary	.47	.29	.72	.25

The IF scores for the nursing students are all consistently over 70%, suggesting an undemanding test, with the listening section appearing to be the easiest of the three sub-sections, with a success rate of 79%. For the business-oriented group, the test was tougher; with success rates for each section ranged around the 50% level (again listening was somewhat easier, with a success rate of 58%). The discrimination index showed that all the sub-sections discriminated to an acceptable degree between the strong and weak

students in the business-oriented group. The least effective discriminator for the business group was the vocabulary section, with a score of .29, but even this score was above the minimal acceptable level cited in the previous section. The test discriminated less well between the subjects in the Nursing department. The best discriminator for this group was the vocabulary section, but the score of .25 was below even that of the business group. The least successful discriminator for nursing students was the listening section, which at .14 was below the minimum acceptable level of 0.2 suggested in the previous section. A general interpretation of the results in Table 5 would be that the test was of an appropriate level of difficulty for the business-oriented group, but was somewhat too easy for Nursing students, with a corresponding weaker ability to discriminate between good and bad students in this latter group.

4. DISCUSSION

Regarding the first research question, The ANOVA study showed the incoming students at Jobu University in 2008 to be made up of two groups in terms of English proficiency: nursing students and business/information science students. The difference between nursing students' scores and students from each of the other two departments was statistically significant, while the score difference between the two business-oriented departments was not statistically significant. This does not necessarily mean that a different placement instrument is needed for the Nursing department, but it does indicate that attempts to improve the test may benefit one group while not the other. On the other hand, a statistical argument could now be offered in support of any decision to pursue separate tests for each of the two groups.

Regarding the second research question, the item analysis summarized in Table 5 confirmed the conclusions drawn from the ANOVA study, showing nursing students scoring better on all sub-sections of the test. Conversely, the nurses' DI scores were consistently weaker than those for business students, showing a reduced ability to distinguish between higher and lower level nursing students. In practice this means that there were probably too many items that were too easy to usefully distinguish between nurses' English ability, but were useful for the same function when used to assess business students.

For the business students, the IF index suggest the test is an appropriate difficulty level. Therefore, any improvement would focus on increasing the DI scores through improving or replacing items. For the nurses, an improvement on the test could begin with retaining

only those items that showed both good discrimination and an appropriate IF score. Items with both weak IF and DI scores could be replaced, while others could be improved, if possible. The listening sub-section of the test was even easier for the nurses, with an average IF score of .79, while, as noted earlier, the DI was a disappointing .14. Again, a test aimed specifically at nurses would require more challenging items that discriminated better between stronger and weak students. The vocabulary sub-section was clearly the most difficult part for business students, with an IF of only .47. For the nurses, the section was also the most challenging, but was still well within their ability range, with an IF of .72: not too difficult. On the other hand, the DI was the weakest for business students at .29 and also weak for nursing students at .25. So, regardless of whether two separate placement instruments are developed, this sub-section could be improved. For the weaker group, items need to be made easier, while for the stronger nurses, effort should be made to clarify items and improve the DI scores.

In conclusion, the study reported here confirmed teacher intuitions that nursing students were significantly better at English than students in the other two departments. In addition, it is now clear that the test is not particularly demanding for the nursing students, and discriminates less than ideally. Improvements would in general need to focus on improving DI scores for all sections and, if a separate instrument were developed for the nursing students, more difficult items could be included.

REFERENCES

Analyse-it Software (2003). Analyse-it for Microsoft Excel, Leeds, UK.

Abel, R.L. (1979). Essentials of educational measurement (3rd Ed.). Englewood Cliffs, NJ: Prentice-Hall.

Brown, J.D. (1988). Understanding research in second language learning: A teacher's guide to statistics and research design. London: Cambridge University.

Brown, J.D. (2005). Testing in language programs: a comprehensive guide to English language assessment. New York: McGraw-Hill.

James, R.G. (2007). Placement Test: Analysis and Improvements. (Bulletin of Faculty of Management Information Sciences, Jobu University, Volume 30, October, 2007).

Microsoft (2003). Excel™.