《研究ノート》

Lack of Motivation as a Criterion in the Assessment of Results of Placement Tests.

T. P. P. Grose

Introduction

In 2004, Sapporo Gakuin University established a computerised system of tests to stream students into appropriate English language classes according to their levels of proficiency. The classes are a compulsory part of the university's General Education programme and are taken by English majors and non-English majors alike. As a result of the tests, teachers reported that classes that had previously been of mixed levels had become easier to plan and teach. The same tests are also used at the end of each academic year to try to evaluate any improvements in student competence. In this case they are renamed and called 'Progress tests'.

The testing system initially provided two examinations: one for English Department students and three other Humanities Departments who were deemed to have a slightly above average level of English proficiency (called the 'e' test) and the other test for the other four Departments (called the 'c' test).

The tests were built on a Moodle platform so they could be designed by SGU teachers and tailored to the appropriate levels of our students. The tests are one hour long and made up of 20 reading comprehension questions and 30 listening questions. Feedback in the form of item analysis provided by the system, gave insights into the validity of all questions (hereafter called 'items') and those deemed too hard or too easy could be changed accordingly. Details of the program are outlined in Hinkelman and Grose (2004).

Feedback comes in various forms. The most important of these in assessing the validity of the items are the Correct Facility (or Item Facility), Standard Deviation, the Discrimination Index and the Discrimination Coefficient.

The Correct Facility simply provides the percentage of correct answers to each item. In a placement test, which aims to rank students in order of proficiency, if all students get 100% (or

0%), it is impossible to tell who is more proficient. Therefore, the optimum range of correct items is considered to be between 40% and 60%. (Brown 1996) If more than 60% of students answer an item correctly, it may be considered too easy. Conversely, if fewer that 40% of students get the correct answer, it may be deemed too difficult. And so on, along the continuum. The higher (or lower) the percentage, the more difficult it becomes to distinguish between the students' abilities.

The Discrimination Index gives an indication of the randomness with which questions are answered. It does this by measuring the results of the top 33.3% of examinees against the bottom 33.3% for each item. The results of the middle 33.3% of students are discarded. For a test to effectively reflect students' overall abilities, the top 33.3% of students should consistently outscore the bottom 33.3%. If an item shows that this is not the case, it indicates that weaker students have outperformed more proficient students. Therefore, a varying degree of guesswork has taken place and an unacceptable level of randomness has influenced the overall results. The Discrimination Index produces results on a scale from -1 to +1. If the index goes below zero, it means that more of the weaker students answered the item correctly than did proficient students. Such items should be discarded. The closer the number is to +1, the less the item has been influenced by randomness.

The mathematical expression is:

$$DI = (Xtop - Xbottom)/N$$

Where Xtop is the sum of the fractional credit (achieved/maximum) obtained at this item by the 1/3 of users having the highest grades in the whole quiz (i.e. number of correct responses in this group). Xbottom is the analogue sum for users with the lower 1/3 grades for the whole quiz. (Moodle, 2008)

Similarly, the Discrimination Coefficient measures how effectively an individual item distinguishes proficient from weak learners. It is a correlation coefficient between a learner's score for a single item set against his/her scores for the quiz as a whole. It is expressed as:

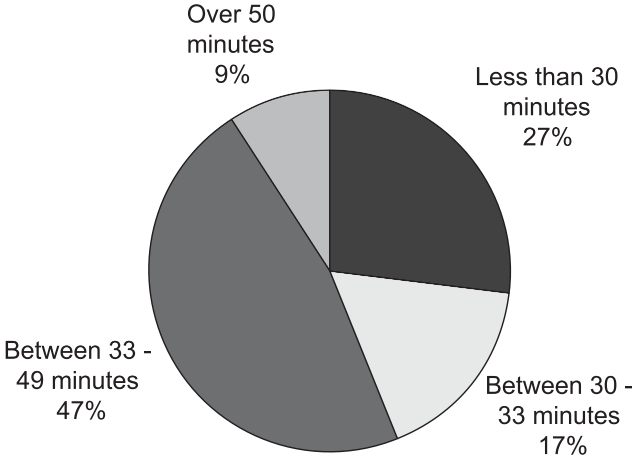
$$DC = Sum\ (xy)/(N * sx * sy)$$

Where Sum (xy) is the sum of the products of deviations for item scores and overall quiz scores, N is the number of responses given to this question, sx is the standard deviation of fractional scores for this question and, sy is the standard deviation of scores at the quiz as a whole. As with the Discrimination Index, results are produced on a scale from -1 to +1. Because it uses data from all examinees (unlike the Discrimination Index which only uses the top and bottom thirds of the body of learners), the Discrimination Coefficient may be seen as a more sensitive and complete form of evaluation. (Ibid, 2008). For this reason, the data from the Discrimination Coefficient has been used in this paper while data from the Discrimination Index has been omitted. It is available on request from the author.

From the outset of the implementation of the test, it was noted that the scores for all of the above criteria were poor, especially in the 'c' test. It was suggested that the test was too difficult and attempts were made to simplify it. However, simplifying the test did not result in any significant overall improvements.

At the same time it was noticed that many students were not trying. A substantial number (Figure 1) were finishing the test as quickly as possible and leaving the room. Some students were finishing within 2 minutes. The physical act of reading the texts and listening to the recordings in the test takes a native speaker approximately 30 minutes so students finishing in less time may be deemed to be not trying. Indeed, forty-five minutes may be considered a reasonable minimum time for a non-native elementary/intermediate student to have a realistic chance of attempting all questions in a comprehensive manner (especially bearing in mind that they can listen to the listening comprehension questions as many times as they like). In order to encourage students to at least have a realistic try at the test, it was therefore decided to disallow students from leaving the examination room for the first thirty minutes of the test. In the year-end test, students were also 'bribed' with 5% of a course grade for just attending the test, and up to 10% depending on how well they performed.

Figure 1: Times Taken By Examinees



Preliminary results indicate these strategies have not worked. Many students still finish as quickly as possible, as before, then do nothing until they are allowed to leave the room. This is supported not only by anecdotal reports from invigilators but also from the data provided by Moodle.

Figure 1 shows the time taken by 449 first-year students who took the 'c' Placement Test on April 7th, 2008. It shows that 121 students (27%) finished the test in under 30 minutes. This is in spite of the fact that they were clearly informed that the scores of those who left the examination room before thirty minutes had elapsed would be discounted. That is to say, their scores would be individually invalid and they would be placed in the bottom class. This caveat notwithstanding, these students' scores have always been incorporated into the overall feedback process. Therefore, for statistical purposes, they are deemed to have taken the test and their scores contribute to the item analysis procedures and to any conclusions drawn therefrom.

It is not clear whether these students had misunderstood, forgotten or simply ignored the instructions (a few expressed regret when they realised they had finished too early) but, for whatever reason, it certainly indicates a general lack of involvement or personal commitment to the process. A substantial minority of 74 students (17%) finished the test between 30 and 33 minutes indicating that many of them quit the test at the first permissible opportunity. Only 39 students (9%) stayed for longer than 50 minutes.

This clearly has serious ramifications when trying to evaluate the validity of questions. Under these circumstances, no matter how simple a question may be, there is a chance that it will give poor Item Facility and Discrimination Coefficient results.

For the purpose of item evaluation, it was therefore decided to remove the scores of students who had left the classroom in under forty-five minutes. It was hoped, by so doing, the validity of questions could be more accurately measured. Figure 2 shows the item analysis prior to the removal of the data of early finishers.

Results

Figure 2: Item Analysis Random Sample (402 students)

Items	Correct Facility	SD	Discrimination Coefficient
1．	50％	0.501	0.34
2．	66％	0.474	0.40
3．	41％	0.492	0.45
4．	49％	0.501	0.35
5．	43％	0.496	0.35
6．	47％	0.500	0.26
7．	57％	0.496	0.35

8.	45％	0.498	0.44
9.	31％	0.464	0.34
10.	50％	0.501	0.27
11.	61％	0.489	0.34
12.	5％	0.211	0.06
13.	40％	0.490	0.19
14.	57％	0.495	0.44
15.	31％	0.461	0.35
16.	52％	0.500	0.43
17	53％	0.500	0.46
18.	40％	0.490	0.36
19.	31％	0.463	0.23
20.	26％	0.438	0.18
21.	48％	0.500	0.17
22.	41％	0.492	0.39
23.	39％	0.488	0.41
24.	56％	0.497	0.32
25.	46％	0.499	0.41
26.	71％	0.452	0.26
27.	40％	0.490	0.39
28.	75％	0.436	0.31
29.	34％	0.475	0.32
30.	53％	0.500	0.45
31.	61％	0.489	0.30
32.	63％	0.483	0.32
33.	38％	0.486	0.28
34.	40％	0.491	0.49
35.	36％	0.481	0.27
36.	50％	0.501	0.25
37.	36％	0.480	0.20
38.	27％	0.445	0.13
39.	36％	0.482	0.44
40.	49％	0.500	0.14
41.	24％	0.427	0.35
42.	21％	0.406	0.44
43.	36％	0.482	0.37
44.	24％	0.427	0.10
45.	34％	0.473	0.30
46.	36％	0.480	0.34
47.	20％	0.404	0.16
48.	21％	0.407	0.11
49.	31％	0.464	0.07
50.	36％	0.480	0.11

These results may be compared with the item analysis for students who stayed for longer than 45 minutes.

A quick glance at the raw data reveals that（not surprisingly）students who spent more time on the test scored better than students who finished early. Correct Facility scores were better

Figure 3: Item Analysis for students who took over 44 minutes（69 students）

Items	Correct Facility	SD	Discrimination Coefficient
1.	55％	0.501	0.46
2.	81％	0.394	0.34
3.	54％	0.502	0.44
4.	65％	0.480	0.33
5.	45％	0.501	0.26
6.	51％	0.504	0.12
7.	62％	0.488	0.30
8.	58％	0.497	0.43
9.	43％	0.499	0.32
10.	61％	0.492	0.32
11.	72％	0.450	0.25
12.	6％	0.235	0.03
13.	36％	0.484	0.03
14.	80％	0.405	0.19
15.	43％	0.499	0.37
16.	68％	0.469	0.47
17	65％	0.480	0.30
18.	61％	0.492	0.42
19.	43％	0.499	0.37
20.	30％	0.464	0.41
21.	54％	0.502	0.13
22.	62％	0.488	0.25
23.	57％	0.499	0.50
24.	59％	0.495	0.12
25.	58％	0.497	0.30
26.	71％	0.457	0.24
27.	57％	0.499	0.35
28.	78％	0.415	0.25
29.	43％	0.499	0.29
30.	72％	0.450	0.40
31.	70％	0.464	0.14
32.	75％	0.434	0.30
33.	49％	0.504	0.26
34.	54％	0.502	0.55
35.	45％	0.501	0.38
36.	61％	0.492	0.25
37.	36％	0.484	0.31
38.	35％	0.480	0.39
39.	65％	0.480	0.25
40.	57％	0.499	0.17
41.	35％	0.480	0.46
42.	33％	0.475	0.56
43.	49％	0.504	0.57
44.	29％	0.457	0.44
45.	46％	0.502	0.50
46.	58％	0.497	0.42
47.	26％	0.442	0.29
48.	23％	0.425	0.35
49.	36％	0.484	0.10
50.	42％	0.497	-0.09

for 47 out of 50 items. Two items received the same score and one item received a lower score. Improvement varied from a minimum of a single percent to a maximum improvement of 29%. The average improvement was 9.31%. Because some item scores became ‘too easy’, Discrimination Coefficient scores consequently showed less consistent improvements.

For the purposes of item assessment, data from Figures 2 and 3 may be divided into two categories. The first of these are those items in Figure 3 that reinforce the assessments we can make from some items in Figure 2. For example, in the first group, the Correct Faculty for Item 12 in both Figures 2 and 3 are 5% and 6% respectively. The lowness of these numbers is a very clear indication of the need to change or revise it. In the same way, low figures for Item 20（26% and 30% respectively）and Item 38（27% and 35% respectively）also indicate a need for change. Furthermore, consistently low figures for Items 41-50 both in terms of Correct Facility and Discrimination Coefficient indicate a problem with the entire genre of questions. Data from Figure 3 shows that there are problems with the validity of 4 items in terms of Correct Faculty and with 3 items in terms of Discrimination Coefficient. Questions 41-50 are sustained ‘lecture-type’ listening comprehension questions. As such, they are certainly the most challenging test of our students’ abilities. This evidence indicates it is too challenging so perhaps a ‘cross-the-board’ change should be considered. Also, low Correct Facility scores and/or low Discrimination Coefficient scores in both Figure 2 and Figure 3 for Items 13, 21, 26 make a strong case for modifying these items.

The second category are those items in Figure 3 that indicate the need for a reassessment of the way in which we may have evaluated the validity of some items in Figure 2. For example, in terms of Correct Facility, there are 9 items whose scores move from unacceptably low to within an acceptable range. They therefore move from being ‘invalid’ to ‘valid’. These are items 9, 15, 19, 23, 29, 35, 43, 45 and 46. This list excludes Items 33, 39 and 50 because, although their Correct Facility scores move into an acceptable range, their Discrimination Index scores are unacceptably low.

Although the Placement Test has been criticised for being too difficult, the above data brings this view into question. For students who spent time on the test, the above-cited items are valid items. This is also supported by data of overall average scores for the test. The average score on the ‘c’ test of April 7th, 2008 was 41.1%. The average for the lowest scoring department was 37.2% and the average for the highest scoring department was 46.2%. Such low figures have been cited as compromising the effectiveness of the placement process and reinforcing low self esteem among students. However, the average score for the students who spent more than 45 minutes on the test（Figure 3）was 53.5%, a score within a very acceptable range. Moreover, the scores for 6 items (14, 16, 30, 31 and 32) increased from an acceptable/marginally acceptable level *into a range that is unacceptably easy.* The Discrimination Coefficient scores were also correspondingly lower in Figure 3 than for Figure 2 for two of these items. Item 14 scores went from 0.44 to 0.19; Item 31, from 0.30 to 0.14;

Conclusion

Decisions on modifying items must be based on a variety of criteria. Clearly, items whose scores are in an unacceptable range in both sets of data should be modified. Also, consistently poor results for the final section of the test (Items 41-50) indicate that sustained listening comprehension is too difficult for these students so a wholesale change in this section should be considered.

Decisions concerning data that varies between the first set of figures (Figure 2) and the second set (Figure 3) are more problematical. However, the argument that the test is too difficult (according to the standards of students represented in Figure 2) is not persuasive enough to justify changes to items that are valid (according to the standards of students represented in Figure 3) thereby making them invalid by creating unacceptably easy questions. These items certainly need to be monitored over a period of a number of tests and changes, if warranted, may be made on a case-by-case basis.

It is important to remember that, for all its faults, the placement system works and that staff (and as far as can be determined, students too) are satisfied with the results that it produces. While the figures demonstrate that a lack of motivation compromise the integrity of the item analysis, it may also be the case that there are degrees of poor motivation and that some students may try a little whereas others may not try at all. This may be enough to provide sufficient data to enable effective placement to take place. In order to determine more precisely whether or not this is the case, there is a need for more close analysis of individual student's results. Also, in order to verify the existence and performance of 'moderately motivated' students, a series of item analyses set at various stages of the test (i.e. analysis of results of students who finish within 30, 35 or 40 minutes) may provide more detailed insights. In addition, a questionnaire designed to obtain feedback from examinees concerning their level of commitment to the placement test may provide valuable information in this area.

Deciding what to do with items that are unacceptably easy is also problematical. Making the items harder and thereby restoring their validity may adversely affect those students from Figure 2 who are marginally interested in the placement process. If issues such as low-esteem are indeed a factor to be considered, making the items more difficult may not be a sensible approach. As we have seen, the lowest departmental average score was only 37.2%. To make this figure even lower may have adverse effects on students'（already low）levels of enthusiasm. This suggests that a pragmatic approach to the revision of certain items may be a sensible option and that practical considerations may outweigh considerations of validity.

'If it ain't broke, don't fix it' is a common-sense attitude to be considered. The test, as we have seen is by no means perfect, but at the same time, neither is it 'broke'. It works. We have identified areas which clearly need to be fixed and this should be done. For the remaining items of questionable validity, a conservative policy of judicious change seems to be advisable. More data, over a longer period of times and covering a wider range of students' responses should provide more detailed insights into the effectiveness of the test. To rush to judgement and make precipitous alterations based on data that includes students with little or no motivation will only compromise the true validity of items to a greater degree than at present. Trial-and-error changes to these items should, in the long run, provide a test that is both valid and effective.

Acknowledgements

A special thanks to Don Hinkelman for all of his valuable input into the process.

References

Brown, J. D.（1996）. Testing in language programs. Upper Saddle River, NJ: Prentice Hall（ISBN: 0131241575）.

Hinkelman, D., Grose, T（2004）*Placement Testing and Audio Quiz-Making with Open Source Software.* PacCALL Journal Volume 1 No.1 Summer 2005, Pp.69-79

Moodle（2008）Quiz Reports: Descriptions defined. Available at:
http://docs.moodle.org/en/Quiz_reports Accessed on March 28th, 2008）

（ティモシー・グローズ　本学人文学部准教授　言語学専攻）