Item Analysis of a Computer Based Placement Test

ANDREW GORRINGE

1. Introduction

The object of this paper is to describe how the author analysed the items on a computer based placement test that he designed and implemented at Kwassui (Gorringe, 2008). The origin of this study comes from the recognition that an in-house placement test would be beneficial to the college and students at Kwassui, since such a test would be cheaper to implement and use than any commercial based alternatives, and would also have the advantage of being more carefully attuned to the level of the students at the College. In June 2007 the test was piloted with 57 first-year students and it was found that there was moderate correlation (Pearson's r Correlation Coefficient 0.69 and Spearman's Rank Correlation Coefficient 0.64) between the placement test and the CASEC - the test being used to place students at the time. In April 2008, the test was again tested with the latest intake of 44 first-year students and the same moderate correlation with the CASEC was found. Also, in both implementations of the test it was found to have a reliability of 0.80 using the Kuder-Richardson 21 (KR-21) formula. This is a reasonable level of reliability, but the author feels that this can be improved upon, and to do this we can look at the individual items and determine whether they are working as they should.

2. The Items

When the placement test was first proposed certain criteria were decided upon and these were:

a. The object of this test is to stream first-year students at Kwassui, so content of the test should be at or around the ability level of the students taking it.
b. The test should be able to discriminate between students of differing language knowledge and ability.

c. It should also be easy to take and administer, and be easily changed as circumstances dictate. (Gorringe, 2008)

Because of their ease of use and the fact that they could be so easily integrated into *Moodle*, which was being used to carry the test, multiple choice items were the preferred choice. In writing these items the author attempted to follow as closely as possible the basic criteria, set out by Henning (1987), for writing multiple choice items and tried to avoid the most common mistakes outlined below (adapted from Henning, 1987).

Example 1 (mixed response)

Sam …?... the tickets for ten pounds yesterday.

1. brought 2. buys 3. brings 4. bought

The problem above is that we cannot be sure which skill we are testing. Is it vocabulary or grammar?

Example 2 (mixed response)

Simon usually…?... aerobics at the weekend.

1. is doing 2. does 3. do 4. have done

This item is written to test tense, but 3 and 4 can be rejected because of subject/verb agreement.

Example 3 (inconsistent)

I enjoy...?... cakes.

1. eats 2. healthily 3. eating 4. has eaten

Here ‘‘healthily’’ is not a verb and can be rejected.

Example 4 (convergence)

1. red 2. read 3. green 4. yellow

With distracters like the above test takers can assume (because of the addition of read) that “red” is probably the correct answer.

Example 5 (nonsense)

The monkeys ...?...

5. drinked 6. drunked 7. had drinked 8. drank

''Nonsense'' distracters should not be used because they have a bad ''washback'' effect (students may think they are correct English and use them later).

Even though the author tried to make sure the items were written in accordance with Henning's criteria there will always be items that do not function as hoped, and to determine which items did or did not function as anticipated, the author analysed each item in the test.

3. Item Analysis

As mentioned earlier the test, was administered through the Kwassui *Moodle* system (see Mazzarelli and Gorringe, 2006 for an overview of *Moodle* at Kwassui) using the testing module. As part of the testing module some statistical analysis is done automatically on the items in the test. The testing module on *Moodle* has the following analyses all based on classical test theory (see Gorringe, 1999):

a. Facility Index (determines the difficulty of an item relative to all the other items)
b. Standard Deviation (determines the spread of answers among the test takers)
c. Discrimination Index (determines the extent to which an item discriminates between more able and less able test takers)
d. Discrimination Coefficient (determines the extent to which an item discriminates between more able and less able test takers- similar to Discrimination Index, but more sensitive)

The Discrimination Index and Coefficient are important in determining how reliable a test is- by removing items that are not performing well we can make the test more reliable. Also note that, as the scale of the Discrimination index and Coefficient is from -1 to +1, the closer to +1 the more an item discriminates. Sometimes, however, an item is not necessarily bad if it has a low coefficient. For example, items which are very difficult or very easy will have low coefficients but we would not necessarily reject them from the test, since easy items are useful at the beginning of a test to give test takers confidence, and more difficult items may discriminate between the very best test takers. If

an item has a negative number then the item has a serious problem since those test takers who are doing best on the test are somehow getting that item wrong, and those doing worse are getting it correct.

In order to show how such analysis works here a two examples - one an example of a ''good'' item and the second an example of a "bad'' item.

After seeing the doctor his health did not _____ and he died.

improve (68%) injure (16%) behave (7%) income (9%)

The correct answer is of course "improve'' and 67% of the test takers chose that answer. The distracters all worked well with a reasonable spread of test-takers choosing different options. The Facility Index for this item was obviously 68%, so not difficult. The Standard Deviation was 0.469. The Discrimination Index was 0.76, and the Discrimination Coefficient was 0.53 which shows that the item is discriminating well.

What's all this _____ in your room?

stuff (18%) brand (29%) range (24%) good (29%)

The correct answer here is ''stuff'' and 18% of the test takers chose the answer. Once again the distracters worked well. However this time the Discrimination Index was 0.00 and the Coefficient was 0.11. This item is therefore not working well and is not discriminating between the better and less able test takers. If we remove items like these from the test then the test will be even more reliable.

4. Revising the Test

Once the analysis of the test results is done through the *Moodle* test module, it is a simple, but time consuming task to go through the items to determine which of them should be removed or modified. Obviously those items which have negative discrimination Coefficients will be rejected. Also, those items where distracters are not being chosen will have to be modified, and non-performing distracters replaced. Finally, after looking carefully at those items with low discrimination coefficients some of these items will also be removed.

In the case of the placement test designed by the author and being used at Kwassui the following revisions will be taking place before the next implementation of the test. Before any revisions of the test had taken place there were a total of 144 questions in the item bank. After the two successful pilots of the test (for a total of 101 test takers), it was found that 7 items had to

be rejected because of negative coefficients and 16 items for very low coefficients. Also, 16 items needed to be revised for the next implementation of the test because of non-performing distracters.

5. Conclusion

The object of this article was to describe the way in which the placement test that is being used at Kwassui was being revised in order to make it more reliable. In order to do this the author utilised the item analysis package that comes with the testing module on the *Moodle* software that is currently being used at Kwassui. By doing so the author could determine which items in the test were performing as they should and which were not. The revised test will be used in April 2009 with the next intake of students into Kwassui. In further articles the author will report on the progress of the test and the extent to which the revisions explained in this article have helped to improve the test's reliability and its usefulness at Kwassui.

Bibliography

CASEC (Computerized Assessment System for English Communication). 〈http://casec.evidus.com/〉.

Gorringe, Andrew. The Rasch Model in Language Testing. Kwassui Review 42 (1999): 201-217.

- - - Gorringe, Andrew. The Design and Implementation of a Computer Based Placement Test. Kwassui Review 51 (2008): 43-48.

Henning, Grant. A Guide to Language Testing: Development, Evaluation, Research. Cambridge, Mass.: Newbury House. 1987.

Hughes, Arthur. Testing for Language Teachers. Cambridge: CUP. 1989.

Kwassui Women's College *Moodle* Site. 〈http://moodle.kwassui.ac.jp/moodle/〉.

Mazzarelli, Sergio and Andrew Gorringe. ‘‘Piloting *Moodle*: The Experience of Two Kwassui Teachers and Their Students.’’ Kwassui Review 49 (2006): 63-85.

Moodle.org. 〈http://www.moodle.org〉.

Received January 31, 2009