Voice of Academia

Academic Series of Universiti Teknologi MARA Cawangan Kedah Vol. 9, No. 1, 2014 ISSN 1985 - 5079

UNIVERSITI TEKNOLOGI MARA
UiTM CAWANGAN KEDAH

© Voice of Academia is jointly published by the Universiti Teknologi MARA, Kedah and Penerbit UiTM (UiTM Press), Universiti Teknologi MARA Malaysia, Shah Alam, Selangor. Printed by Perpustakaan Sultan Badlishah, Universiti Teknologi MARA, Kedah.

Voice of Academia
Academic Series of Universiti Teknologi MARA Cawangan Kedah

CONTENTS **PAGE**

Making a Placement Test Based on Vocabulary Knowledge

Andrew Gorringe
John Anderson
Kwassui Women's University

ABSTRACT

The object of this paper is to describe how the authors used vocabulary to create a valid and reliable placement test. First, the writers will explain the reasons for using vocabulary for placement testing purposes by looking at the literature on knowledge of vocabulary and its relationship to language ability. Next the paper will explain how the test was constructed and delivered to the test takers using the Moodle quiz module. It will then show how the test and test items were analysed using both classical test theory and item response theory. It will also show the reliability of the test and how the test was validated. Finally the paper will show how accurately the test placed the test takers.

The results of this study demonstrate that using vocabulary knowledge as the basis for a placement test is an easy and efficient way for teachers to determine their students' relative language ability. The study also shows that the test itself is both valid and reliable and that it accurately places students according to their language ability.

Keywords: Testing, Vocabulary, CALL

Introduction

In Japanese universities, mandatory English classes for first-year students are widespread, as are the use of placement tests for English language ability grouping. The effectiveness of streaming over time is not at issue here. Practically speaking, it is easier to meet the needs of a class of students with similar second language ability, whether high or low, as opposed to a class of mixed, divergent abilities. How do we ensure then, that the placement test accurately separates students according to their English language ability?

Standardized tests measuring all four skill areas would likely give us the clearest picture of a student's overall English ability. The time, cost and staff required to administer such tests, however, is often prohibitive. Especially given the time constraints of an already-packed orientation schedule, an ideal placement test would be quick, effective and easy-to-administer.

This paper will describe the development of a valid and reliable multiple-choice test, based solely on vocabulary knowledge, which accurately separates students according to their general English language ability. Because the test is computerized, it can be administered to a large group of students at one time, and by only one person. It also can be completed in 45 minutes, with the results calculated automatically and instantaneously. The first part of the paper will show how vocabulary knowledge is related to language ability. Next there will be a description of how the test was created and delivered to the test takers. Then, it will show how the test was validated and demonstrate its reliability. The authors will then describe how the test was analyzed by using both classical test analysis and the one parameter Rasch item response model. Finally, the paper will show how the test has been improved and how it now accurately and reliably places students into ability groups.

Vocabulary and Language Ability

Before describing the actual development of the test, however, let us first address the underlying premise upon which it is based: vocabulary knowledge is an accurate predictor of general second language ability.

While learners do not need to understand every word they hear or read in

order to understand the gist of a particular passage in a second language, an increase in both one's breadth and depth of vocabulary leads to higher levels of listening and reading comprehension. This is backed up empirically by Stæhr (2009), who found a significant correlation between vocabulary knowledge and listening comprehension, and Qian (2002), who demonstrated a highly positive correlation between vocabulary size and depth and reading comprehension, as well as how vocabulary depth in particular is helpful in predicting reading comprehension. Willis (1990) also reminds us of how often the size of one's vocabulary is used as an indicator of one's language ability or progress in a second language, as seen in the use of graded readers.

Vocabulary knowledge as an accurate predictor of second language ability is not restricted to passive language skills. It is generally regarded among scholars that around 2000 word families are needed to engage in everyday spoken English discourse, a vocabulary size larger than previously thought (Adolphs and Schmidt, 2003), and Chastain (1988) reminds us of what we regularly observe in our own language classes: a lack of vocabulary is the main reason why students are unable to say what they want to in communication activities.

The Pilot Test

Before creating the test it was decided that the test should meet the following criteria:
a. The content of the test has to be at the students' level since it is to be used to place them into ability groups.
b. The test should be able to divide the students according to their relative abilities.
c. The test must be easy to take and deliver to the students, and be able to be changed as necessary.
The first of these criteria was met by looking at a previous unpublished study at Kwassui University. In the summer before the test was created first-year students took the V-check test. This test is a computer adaptive test based on vocabulary knowledge. The test showed that the students entering Kwassui had gaps in their knowledge of the 2,000 most frequently used English words. Those with fewer gaps in their knowledge of these 2,000 words were stronger at English than those with larger gaps. Therefore, it was decided that the test would be based on the 2,000 most com-

monly used words, and to do this the lists of words from Paul Nation's Range programme were used as the basis for the test.

The second criterion was met by making sure the test had sufficient items to help discriminate between the stronger and weaker students. The more items the test has the easier this is. Also, with more items a test is usually more reliable and therefore we can be certain of the results of the test (Bachman and Palmer).

The last criterion was met by using the quiz module in Moodle and by using multiple choice items. The advantages of using Moodle and its quiz module are many. First the Moodle software is free and open source which is a great advantage over commercial software. Secondly, using Moodle is efficient and easy to administer. The students can take the test in large numbers in a computer room and be supervised by only one teacher. The quiz module can randomise each item in the test and each of the distracters making it almost impossible for test takers to cheat. Finally, the quiz module in Moodle automatically analyses the test using classical item analysis so we can change and edit items as necessary.

Multiple choice items were chosen for the test because over the computer they are easy for the test takers to take and easy to grade. Also, as mentioned above, such items can be quickly analysed and improved as necessary.

It was decided, therefore, that the test would have 100 multiple-choice vocabulary items taken from the baseword lists in the Range programme. The items would test the vocabulary in context and follow this basic model:

Smoking is a ———————— to your health.

a. sickness b. wrong c. risk d. problem

Reliability and Validity

The pilot test was administered to all 57 first-year English Department students at Kwassui at the end of the summer semester. There was a 60 minute time limit to the test, and students were allowed to leave the testing room as soon as they had finished the test (the quickest took 24 minutes 47 seconds and the slowest 45 minutes 8 seconds). The reliability of the test was determined using the internal measure of reliability Kuder-Richardson 21 (KR-21), and its validity was determined by correlating it with the

students' results on the CASEC test (a commercial test used at many Japanese universities for placement purposes).

Mean = 64.32, Median = 64, Minimum= 38, Maximum = 85,
Standard deviation = 10.63
Reliability (KR-21) = 0.80
Standard Error of Measurement = 4.70
From the statistics we can see that the test has a normal distribution about the mean and has a "moderate" reliability level of 0.80.

The validity of the test was determined in two ways. The first and most important was to find the extent to which the test correlated with a reliable and dependable measure of the test takers ability in order to find its concurrent or criterion related validity (Hughes, 2003). The students had previously taken the CASEC test, and so the in-house test results were compared with those on the CASEC test using Pearson's r (correlation coefficient) and Spearman's Rank Correlation Coefficient. With Pearson's r the coefficient was 0.69 and with Spearman's Rank it was 0.64. The correlation coefficients show "moderate" correlation between the 2 tests. It should also be noted that CASEC tests have been positively correlated with the TOEIC test. This correlation with the CASEC test shows that the test has concurrent or criterion related validity with a test that measures overall language ability and one that has been proven to divide and place students of varying abilities into ability groups.

The test was also checked to see if it had content validity. This validity is determined by checking if the test has representative sample of what it is meant to test (Hughes, 2003). Certainly the placement test has content validity since all the items were taken from the baseword lists of the 2, 000 most commonly used words from the Range programme. The test is testing vocabulary based on those words so it has content validity.

Classical Item Analysis

As mentioned earlier, the test was administered through the Kwassui Moodle system using the quiz module. As part of the quiz module some statistical analysis is done automatically. The module gives us the following useful general statistics for the test: the mean, standard deviation, reliability (Cronbach's alpha) and the standard error. It also gives us the following

individual item analyses all based on classical test theory:

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 are two examples: — the first 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 "improve" and 68% 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 or edit items like these from the test then the test will be even more reliable.

Another way in which classical item analysis can help us to create a more valid and reliable test is to look at how the distracters are functioning. For example, this item from the placement test shows that one of the distracters is not functioning well.

She _____ money for the poor people.
collected (51%) asked (22%) handed (27%) rented (0%)

The correct answer collected has been answered by 51% of the test takers, the distracter asked by 22%, the distracter bought by 27%, and the final distracter rented has not been answered by anyone. We would therefore have to change the distracter rented for a different one.

Rasch Analysis

The test items were also analysed using Rasch Analysis with software from Assessment Systems. Unlike with traditional Classical Theory, where the results and the information from the test are dependent on the characteristics of the test takers and the test (Baker 1997), Rasch analysis can impose linearity on a set of scores. This means we can calibrate a test independent of the test takers, and calibrate the test takers independent of the test on the same scale. In practical terms it means we can give different tests to different students and, because the scores from the test are on the same scale, we can compare the results of the tests directly.

When using Rasch analysis on a set of items two conditions have to be met. The first is "unidimensionality", which means the items on the test should be testing the same construct. In the case of the placement test discussed in this paper, care was taken to ensure that each item was indeed testing

vocabulary. The second condition is "local independence". This means that each item must be independent of any other item. Each item on the placement test is separate and does not depend on any other item to be answered correctly so the test has "local independence".

After the test was administered to the students, the results were analysed using Rasch Analysis. The programme used analysed the items and the test takers and the most useful part of the analysis for the placement test's purpose looks like this:

Item	Difficulty	SE	Chi Square	D of F
1	0.066	0.284	25.989	11
2	0.299	0.276	11.229	11
3	0.299	0.276	12.107	11
4	-0.271	0.301	14.423	11
5	0.299	0.276	9.676	11

The above shows the first 5 items of the hundred item Kwassui placement test. For each item there is a numerical value representing difficulty, standard error, chi square and degrees of freedom. The difficulty of each item is measured in logits usually on a scale of between -4 and +4, with +4 being the most difficult item and -4 the easiest. In the same way test takers are also given a logit level signifying their relative ability at the construct being tested (in this case vocabulary knowledge). A test-taker with an ability level of 0.899 logits, for example, would have a 50% chance of answering an item with a difficulty level of the same amount of logits (0.899). For each item, and its associated level of difficulty, there is a standard error of measurement which is linked to the number of items and test takers. The more items and test takers the lower the standard error for each item. The Pearson Chi square Lack-of-fit statistic is also printed together with its associated degrees of freedom. This shows the extent to which an item "fits" the Rasch model. High Chi squares suggest that an item cannot be described properly by the Rasch model. With 11 degrees of freedom anything over 19.68 may signify lack-of-fit (at the 0.05 level). Item number 1 shows this lack-of-fit and so the item has either to be edited or removed.

There are advantages to using Rasch to analyse any test since, once we know the relative difficulty level of each item, we can easily load the test

with items of specific difficulty in order to divide the students more accurately. In the case of the placement test discussed in this article, most of the classes needed to be divided into two ability groups. By loading the placement test with more high difficulty items and more low difficulty items rather than average difficulty items the test can more accurately divide the test takers between "top" and "bottom" groups.

Discussion

The pilot test was found to have a reliability of 0.80 (KR-21) and a Standard Error of measurement of 4.70. Its validity was determined by comparing it to the CASEC test and it was found, using Pearson's r coefficient, that there was a correlation of 0.69 between the two tests. Out of the 100 items on the pilot test 24 were removed from the test either because they did not fit the Rasch model or they were not functioning correctly after being analysed through the quiz module in Moodle (Appendix I). A further 8 items were edited and their discriminators were changed (Appendix II). In subsequent iterations of the test many more items were removed or edited, and eventually the number of items was reduced to 80 and, by using Rasch analysis, more difficult items (over + 2 logits) and easier items (under -2 logits) were added to the test in order to make the test as a whole more discriminating between stronger and weaker students.

Today the test (of 80 items) is being used to place first year students from all departments into ability grouped English classes at Kwassui Women's University. The reliability of the most recently used test is 0.89 (Cronbach's alpha) with a Standard error of measurement of 3.642.

Conclusion

This paper has shown that using vocabulary knowledge to place students into ability groups is efficient, reliable and valid. The placement test at Kwassui is easy to deliver to students via computer and easy for the students to take. The results of the test are available for teachers immediately after the students have taken the test and the results can be easily and quickly analysed. The test has a high reliability and was shown to have criterion related validity with its reasonable correlation with the commercial CASEC test. More importantly the teachers at Kwassui Women's University are satisfied with the way the test divides students into ability groups.

References

Adolphs, S. & Schmidt, N. (2003). Lexical Coverage of Spoken Discourse. *Applied Linguistics¸24/4*, 425-438.

Baker, R. (1997). Classical Test Theory and Item Response Theory in Test Analysis. *Special Report No 2: Language Testing Update*. Lancaster: Centre for Research in Language Education, Lancaster University.

Bachman, L. & Palmer, A. S. (1997). *Language Testing in Practice*. Oxford: OUP.

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

Chastain, K. (1988). *Developing second language skills*. Orlando, FL: Harcourt Brace.

Hughes, A. (2003). Testing for Language Teachers (2nd Ed.). Cambridge: CUP. *Kwassui Women's College Moodle Site*. http://moodle.kwassui.ac.jp/moodle/.

Moodle.org. http://www.moodle.org.

Qian, D.D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52, 513–536.

Range Programme. http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, 31/4, 1-31.

V- Check. http://www.wordengine.jp/vcheck.

Willis, D. (1990). *The Lexical Syllabus*. London: Collins

Appendices

Appendix I (Vocabulary test items removed after the pilot test)

1. Vocabulary test items not fitting the Rasch model with Chi square values larger than 19.68 (11 degrees of freedom at 0.05).

    Activity
    Already
    Factory
    Familiar
    Field

Hang
Original
Sheet
Value

2. Vocabulary test items with very low (below 0.25) or negative discrimination coefficients.

Addition
Breathe
Cool
Determine
Frozen
Increase
Intend
Net
Package
Relatively
Scarce
Square
Still
Wet
Whatever

Appendix II (Edited vocabulary test items after the pilot test)
Items with non-functioning discriminators (no test takers chose the discriminator)

Item	Old Discriminator	New Discriminator
Bank	Change	Bill
Cost	Payment	Check
Dream	Try	Goal
Kind	Perfect	Sweet
Meal	Plate	Dish
Performed	Acted	Tried
Result	Final	End
Screen	Light	Cover


Voice of Academia

Voice of Academia

Academic Series of Universiti Teknologi MARA Cawangan Kedah

UNIVERSITI TEKNOLOGI MARA
UiTM CAWANGAN
KEDAH