The design and validation of an achievement test to measure improvements in vocabulary

Adrian Paterson
Akita University

Introduction

Assessment is an important part of any teacher’s job. However many teachers see educational testing as a distraction from teaching and technically difficult (Basanta, 1995; Stevenson & Riewe, 1981). This preconception means that there is a poor understanding among teaching practitioners of the principles of testing. Therefore it is important to make testing theory more accessible to teachers. This means either raising the level of understanding of teachers or lowering the level of the materials. One solution to this problem would be to make training in testing theory a required part of teacher training courses. However, this would require a large investment of time and resources. Another solution would be for assessment and testing practitioners to make the subject more accessible by developing easier testing resources that are based on testing principles and can be readily used and adapted by teachers with minimal training. This paper attempts to do this by showing how a popular vocabulary test was adapted for the specific needs of a course using a simple procedure that anyone with knowledge of spreadsheets can use.

Background of the study

This project was conducted at a Japanese national university. It was the result of a request to demonstrate the efficacy of reduced class sizes. The objective was to administer a pretest and a post-test in order to show a significant improvement in the English ability of students. In fact, several teachers expressed concern about the very real possibility that scores could actually decrease due to first-year students relaxing their study habits after having studied hard for university entrance exams. The student placement test was chosen to be the instrument for this, however, the vocabulary questions in the placement test were combined with the grammar questions, and they were considered too few and too difficult to interpret. It was felt by the teachers of the course that vocabulary was an important area that students were likely to make an improvement. Therefore, it was decided to create a new test specifically for vocabulary. However, as is often the case in such circumstances, extra time and resources were very limited. So it was decided to use a proven test design and adapt it to fit our requirements. In so doing, several assumptions were made about the materials used, including; the validity of the target vocabulary list provided in the textbook, the appropriateness of the definitions provided in the teacher’s book, and their suitability for use in this type of test design. The aim of this paper is to report the results of the test, and examine the validity of the assumptions made in its preparation.

Unit of measurement

One of the first, and most important, things one must do when attempting to measure vocabulary is to define the unit of measurement (Nation, 2007; Schmitt, 2010). This means defining exactly what we mean by “a word”. This may seem simple at first, but it actually quite complex. The most commonly used units are:

Token: The number of words in a text regardless of the number of times used.

Type: The form of word that is different from others, each inflected form is counted as a separate word.

Lemma: A base word plus all of its inflected forms. Dictionaries usually classify words by lemmas. They are differentiated semantically and by part of speech.

Word family: A headword, its inflected forms, and its closely related derived forms” (Nation, 2001). A group of lemmas related to a base word by the use of affixes that are similar in meaning but differ in part of speech.

Formulaic sequence: Fixed or semi-fixed multi-word phrases that behave more like words than grammatical constructions.

Tokens and types are not generally considered very useful when estimating vocabulary size, therefore, they are not widely used in vocabulary studies, although, they are easy to program into a computer, and the ratio
of types to tokens is sometimes used a rough measure of lexical richness. Formulaic sequences (Schmitt, 2004; Wood, 2010; Wray, 2002, 2008) are gaining considerable momentum in vocabulary research but are too complex for the current study.

The units of most interest to vocabulary researchers are lemmas and word families. The choice depends on whether the vocabulary of interest is productive or receptive; “… in studies of learners’ productive use of language, the lemma is the most valid unit of counting because different lemmas involve different collocations and grammatical constructions, and the use of word families would mask this important productive knowledge. When receptive uses are considered, however, the word family is a more valid unit” (Nation, 2007, p.39). The VLT is a test of receptive vocabulary.

Research questions

The aim of this project was to produce a vocabulary test that could be prepared, administered, and scored using a minimal amount of time and resources, and that could measure the gains made by first-year students during two semesters of required English courses. This meant that a high priority was given to practicality. It was made using “recipe” style instructions based on (Schmitt et al, 2001, p.59), and consequently one of the concerns was did these compromises impact the reliability and validity of the test. The research questions are:

1. Do Japanese university students make incidental vocabulary gains during their first year of university English study?
2. Can a vocabulary test be made with the main priority given to practicality without making too many compromises in reliability and validity?
3. Can such a test be made in a way that is reproducible?

Method

As the practicality of the test was given high priority, also because there were not enough time or resources for trialing the test, it was necessary to use a vocabulary test design with a well-proven record of reliability and validity. The choice of test, its adaptation, and implementation are outlined in this section.

Instrument

Time was a major consideration in the preparation of this test, so it was decided to use the most readily available resources where possible. This meant avoiding lengthy development of materials. The test format was the vocabulary levels test (VLT) (Nation, 2001; Nation, 1983; Schmitt, Schmitt et al, 2001). This format was chosen because it makes efficient use of vocabulary resources, and this researcher had prior experience of using and adapting it (Paterson, 2004).

The VLT was originally designed as a diagnostic tool to give teachers a rough indicator of the receptive vocabulary size of their students. Nation recommends that it be used as part of a battery of assessments to give a clearer indication of a student’s linguistic ability. It has also been used for research purposes in EFL (Lauf er, 1998; Beglar & Hunt, 1999). It is divided into five frequency-based sections; 2,000-word level, 3,000-word level, 5,000-word level, academic word level, and 10,000-word level. Combinations of levels can be changed to suit the ability level, vocabulary needs, and the goals of the course in question. For example, Beglar and Hunt (1999) used only the 2,000-word level and academic word level in their study of a course of English for academic purposes. The original test was validated by Read (1988), and later versions by Schmitt et al (2001) and Beglar & Hunt (1999) were validated by their respective authors.

The distinctive design feature of the test is its use of clusters of six words with three definitions, or synonyms next to them. The definitions use only words from higher frequency levels, so that they should be easier for learners to understand than the words being tested. The vocabulary items in each section are word families taken from the same word frequency band, and the words in each cluster are of the same part of speech. Learners are required to match the definitions with three of the six words, commonly referred to as test items, and write the relevant number next to the meaning. The advantage of this design is that it requires fewer distracters, while overall the test requires one distracter per test item, when a test taker is matching the words to the meanings there are always effectively between three and five distracters. Correct answers are usually given one point, incorrect or blank answers receive zero. The original version (Nation, 1983, 1990) had six clusters, but Schmitt et al (2001, and also reproduced in Nation, 2001) increased that to ten clusters.

Test design

For the purposes of the current study, the VLT was adapted in several ways. Firstly, it was reduced to a single level. This was mainly due to the relatively small
size of the vocabulary sample, and because word frequency data was not readily available. In order to make scores easier to interpret the test was scored out of 100 points, but to reduce the time required for test administration and marking, each correct item received two points. This meant that there were 50 items grouped into 16 clusters of three items with one cluster of two items.

The test items were randomly selected from the word list in the teacher’s book (MacIntyre & Hubley, 2010), which consisted of 240 words. The number of clusters assigned to each part of speech was determined by calculating its proportion of the total, as shown in Table 1:

<table>
<thead>
<tr>
<th>Part of speech</th>
<th>Number in list</th>
<th>Percentage</th>
<th>No. of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun</td>
<td>87</td>
<td>36.2%</td>
<td>6</td>
</tr>
<tr>
<td>verb</td>
<td>94</td>
<td>39.2%</td>
<td>6.67</td>
</tr>
<tr>
<td>adjective</td>
<td>52</td>
<td>21.7%</td>
<td>3</td>
</tr>
<tr>
<td>adverb</td>
<td>7</td>
<td>2.9%</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>240</strong></td>
<td><strong>16.67</strong></td>
<td></td>
</tr>
</tbody>
</table>

In order to prepare a test using the method outlined in this paper, the following are required; 1) spreadsheet software, which in the current study was Microsoft Excel and 2) a list in electronic table form of target L2 words with part of speech and either L2 definitions or L1 translations.

The textbook used by the classes involved in the current study was Reading Explorer 2 (MacIntyre, 2009), which included a list entitled “Target Vocabulary”, the words in this case were lemmas. The companion Teacher’s Guide (MacIntyre & Hubley, 2010) had the same list but included, their part of speech, and an English definition based on the context in which it was used in the Student’s book. The exact source of the target vocabulary is not clearly specified. However, it does state that they are “high-frequency words in academic and non-fiction reading” (MacIntyre & Hubley, 2010, p.12) and they go on to refer readers to Nation (2008), implying that it is probably based on the Academic Word List (AWL) (Coxhead, 2000), which has become the standard for EFL textbooks focused on academic skills.

Unfortunately, the list was not available in a usable electronic form, so it had to be scanned and then converted to digital text using optical character recognition (OCR) software. Once it had been checked for errors and copied into a spreadsheet, it was ready to use. The final list contained 240 words (lemmas), of which 87 (36.25%) were nouns, 94 (39.16%) were verbs, 52 (21.67%) were adjectives, and 7 (2.91%) were adverbs.

The choice of lemmas as the unit of vocabulary in this textbook is problematic. As Nation points out; “…in studies of learners’ productive use of language, the lemma is the most valid unit of counting because different lemmas involve different collocations and grammatical constructions, and the use of word families would mask this important productive knowledge. When receptive uses are considered, however, the word family is a more valid unit” (Nation, 2007, p.39). As this a reading textbook, the more appropriate choice of vocabulary unit would be the word family rather than the lemma. This raises questions about the content validity and consequently the construct validity of the test, which will be addressed later in this discussion.

A copy of the list was sorted by part of speech, and random numbers were generated and used to sort the words within their part of speech. Then an empty row was inserted after every sixth word to break them up into clusters. The parts of speech were in alphabetical order, and the number of clusters in each was determined by its proportion of the total, so there were 3 adjective clusters, 1 adverb cluster, 6 noun clusters, and 7 verb clusters respectively. The remaining words for each part of speech were discarded. The first three words in each cluster became the test items and the last three the distracters, except for the last verb cluster which only had 2 test items but had four distracters. The definitions for the distracters were deleted. Then the option words in each cluster were sorted alphabetically in order to randomize the number sequences of the answers. Finally, each cluster was checked to ensure that it met the following principles from Schmitt et al (2001, p.59).

1) The options in this format are words instead of definitions.
2) The definitions are kept short, so that there is a minimum of reading, allowing for more items to be taken within a given period of time.
3) Words are learned incrementally, and tests should aim to tap into partial lexical knowledge (Nagy et al., 1985). The Levels Test was designed to do this. The option words in each cluster are chosen so that they have very different meanings. Thus, even if learners have only a minimal impression of a
target word’s meaning, they should be able to make the correct match.

4) The clusters are designed to minimize aids to guessing. The target words are in alphabetical order, and the definitions are in order of length. In addition, the target words to be defined were selected randomly.

5) The words used in the definitions are always more frequent than the target words. The 2000 level words are defined with 1000 level words and, wherever possible, the target words at other levels are defined with words from the GSL (essentially the 2000 level) (for more details, see Nation, 1990: 264). This is obviously important as it is necessary to ensure that the ability to demonstrate knowledge of the target words is not compromised by a lack of knowledge of the defining words.

6) The word counts from which the target words were sampled typically give base forms. However, derived forms are sometimes the most frequent members of a word family. Therefore, the frequency of the members of each target word family was checked, and the most frequent one attached to the test. In the case of derivatives, affixes up to and including Level 5 of Bauer and Nation’s (1993) hierarchy were allowed.

7) As much as possible, target words in each cluster begin with different letters and do not have similar orthographic forms. Likewise, similarities between the target words and words in their respective definitions were avoided whenever possible.

The decision to use the definitions in the list unmodified meant that some of the above principles were violated. While most of the definitions are short, there are a few longer ones, violating principle 2. In order to avoid violations of principles 3 and 7, if two words in a cluster had similar meanings or spellings they were swapped with a word of similar function in a cluster with the same part of speech (i.e. nouns were swapped with nouns, distracters were swapped with distracters, etc.). When a test item and a distracter were in conflict, the distracter was moved. In accordance with principle 4, the option words were sorted alphabetically, the definitions were not as they were already been sorted using random numbers, and experience with previous tests using this format had shown that interfering with them produced regularity in the patterns of answers (Paterson, 2004).

Subjects

The participants in this study were 938 Japanese university students enrolled in compulsory first-year English courses. First-year students are required to take two one-semester general English courses, the pre-test was done at the beginning of the first semester course and the post-test was done at the end of the second semester one. They were from all of the majors offered by the three faculties of the university, except for trainee doctors who take a specialized medical English course. The cohort was divided into classes by major using a commercially available placement test (the name and validation information of this test are not available to this researcher at time of writing). Of these 938 students, 915 did the pre-test, 826 did the post-test, and 810 did both. Students who were absent were not asked to do the test later. Students who only did one of the tests were included in the class means, but not included in the pre- and post-test comparisons.

There were 28 non-first-year students who took the tests. These were second-year students and above repeating the courses because they did not pass or could not take them in their first-year. Of these repeaters, only one did both tests, and the rest did only the pre-test or the post-test. These repeaters were included in the statistical analysis because every year there are several of them in each class and so including them offers a truer reflection of the progress made by the class as a whole. Also, because they make up less than three percent of the class they had a negligible effect on the final results.

Test administration

Classroom teachers administered the tests during class time. They were told to give students adequate time to complete the tests, about 20 minutes was recommended depending on the level of the class. These were then hand graded by the classroom teachers, and the graded papers were handed in to the course coordinator. The students’ raw answers were then entered into a spreadsheet by this researcher. The spreadsheet automatically scored the answers and calculated the score. Any discrepancies between the hand-marked score and the computer-generated score were reconciled, and the necessary amendments made. While this meant a duplication of effort, it also ensured accuracy of both the grading and data input.

Results and Discussion

In this section, I will present a summary of the
results of the test. Then I will attempt to provide evidence of its validity.

Summary statistics

The number of students who took each of the tests are summarized in Table 2:

<table>
<thead>
<tr>
<th>Table 2. Test taker numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
</tr>
<tr>
<td>Pre-test</td>
</tr>
<tr>
<td>Post-test</td>
</tr>
<tr>
<td>Both pre- and post-test</td>
</tr>
</tbody>
</table>

As is fairly typical of first-year university courses, there is a very low rate of absenteeism at the beginning of the year. However, over the course of the year this increased, and when combined with students who dropped out, we see that there was a high rate of attrition by the time of the post-test. This is a concern for the integrity of the measurement because it is often the students of lower ability who drop out or have high absence rates and this could inflate the scores on the post-test. However, when we compare the means and standard deviations in Table 3 for all students who sat each test with those for students who sat both tests (pre-test m = 59.87, SD = 16.921, post-test m = 67.17, SD = 16.812), we see that there were negligible differences.

<table>
<thead>
<tr>
<th>Table 3. Summary statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>Pre-test</td>
</tr>
<tr>
<td>Post-test</td>
</tr>
</tbody>
</table>

A paired-samples t-test was conducted to evaluate the improvements in scores between the two test administrations. There was a statistically significant increase in scores between the pre-test (m = 59.87, SD = 16.921) and the post-test (m = 67.17, SD = 16.812, t(809) = 18.773, p<.005). The eta squared statistic (.303) indicated a large effect size. However, this result should be interpreted with caution as it represents an increase of six points, or three words.

Practicality

This was given a high priority in the preparation of the test. Once the word list had been scanned and the data imported into the spreadsheet, each version of the test took approximately one hour to prepare, including checking the clusters for words with similarities of form or meaning. Also, the test took up about 20-30 minutes of class time, and with the marking template provided to teachers, marking and scoring takes about one minute per paper because they are checking numeric answers. The data input was not required for the test itself, but it was done for the purposes of this validation study and took about a week for one person to complete due to class commitments. This suggests that the test is very practical.

Validity

Validity is the degree to which a language test measures what it is supposed to and how well it predicts real-world language use. It draws on information about the underlying concept of language (construct) and how that is manifested in the content of the test and the tasks test takers are required to perform. It also considers how the test is perceived by stakeholders (face validity).

Construct Validity

The construct validity of a language test “is an indication of how representative it is of an underlying theory of language learning” (Davies et al., 1999, p.33). This the most important validity measure because if a test does not actually measure what it claims to, all other claims about it become irrelevant.

The test is based on a course textbook, so the construct is largely determined by the list of target vocabulary in the textbook. The problem with the list is that some of the definitions use words with lower frequency than the words they are defining, for example using figures from (Leech, et al., 2001) aware (108 occurrences per million) is defined using conscious (31 occurrences per million). This suggests that word frequency and difficulty were not considered when compiling the definitions for the word list, and yet frequency is generally considered by vocabulary researchers to be one of the leading predictors of a word being known (P. Nation, 2001; Read, 2000; Norbert Schmitt, 2010). Nation (1990, p.264) recommends that words used in definitions should be of a higher frequency level than the target words. This raises the question of whether a student getting a wrong answer means that they did not know the word, did not understand the definition, or both?

One way to verify the validity of a test is to compare scores with an established test of the same skill. The first-year students’ scores on the vocabulary section of the placement test were provided to this researcher. However, this did not include information
about how it was assessed or how its validity had been established. Therefore, a comparison with the vocabulary section of the placement test does not provide useful information about the vocabulary test.

Probably the most serious challenge to the construct validity of the test is use of lemmas as the unit of vocabulary. As Nation points out; “…in studies of learners’ productive use of language, the lemma is the most valid unit of counting because different lemmas involve different collocations and grammatical constructions, and the use of word families would mask this important productive knowledge. When receptive uses are considered, however, the word family is a more valid unit” (Nation, 2007, p.39). The course is based on a reading textbook, the more appropriate choice of vocabulary unit would be the word family rather than the lemma. This raises questions about the content validity and consequently the construct validity of the test. However, Nation is not clear on whether “more valid” means that a violation of this invalidates the measure.

Face Validity

The next criterion by which to judge the test is face validity. Face validity is defined as: “The degree to which a test appears to measure the knowledge or abilities it claims to measure, as judged by an untrained observer” (Davies et al., 1999). While the design of the test is a little unconventional, it is quite clearly a vocabulary test because it involves matching words with their meanings. Also, it can be claimed to be an achievement test because the words and definitions are taken directly from the textbook that the test takers use in the course it is evaluating.

Another criterion is the authenticity of the task. It is difficult to argue that this type of test is truly authentic, as with most other multi-choice language tests, it has little relevance to real-world language tasks. However, the counter argument is usually that it provides an indirect measure of the underlying skill. In this case, it indicates an ability to retrieve the meaning of a word.

Reliability

Reliability is “the actual level of agreement between the results of one test with itself or with another test” (Davies et al., 1999, p.168). This can be measured using several correlation statistics.

The first of these is internal reliability, which is measured using Cronbach’s alpha. This measures how consistently the items are measuring the target skill. Alpha ranges between 0 and 1.0, the higher the coefficient is, the more consistent the measure. A Cronbach’s alpha above 0.7 is considered high. As shown in Table 4, the alphas for both tests are similar and quite high. This means that the items in the test are very consistent.

Table 4. Reliability statistics

<table>
<thead>
<tr>
<th>Test</th>
<th>Cronbach’s alpha</th>
<th>N of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td>.878</td>
<td>50</td>
</tr>
<tr>
<td>Post-test</td>
<td>.890</td>
<td>50</td>
</tr>
</tbody>
</table>

Squaring alpha gives an indication of the percentage of the variance in scores is accounted for by the trait, in this case 77% and 79% respectively.

The other measure is how well the scores on both of the tests correlate with each other. In other words, how well a score on the pre-test predicts the score on the post-test. This is measured using the Pearson product-moment correlation coefficient (often shortened to Pearson’s r). We can do this in two ways; one is by comparing students’ scores on both tests, and the other is by comparing the Item Facilities, or the percentage of students who got a particular item correct.

The correlation between the pre-test and post-test scores of the 810 students who did both tests was fairly strong (r = 0.785, n = 810, p < 0.0005). Again squaring r gives us the coefficient of determination, which tells us that 62% of the variance in students’ post-test scores is predicted by their pre-test scores.

In order to compare test items, it was first necessary to identify ones that occurred in both tests. Twelve items were identified and their Item Facilities were calculated and then these were correlated. This correlation was very strong (r = 0.966, n = 12, p < 0.0005). The coefficient of determination showed that the item facility of an item on the pre-test predicted 93% of the variance for the same item on the post-test.

While they could be better, the above measures show that the two versions of the test are equivalent and have not been seriously compromised in terms of reliability.

Conclusion

These tests provide evidence of knowledge of the form-meaning relationship of lemmas presented as target vocabulary in the course textbook. The comparison of the means of the pre-test and the post-test showed a statistically significant increase of six percentage points. However this only represents about
three words, and while this is better than the very real possibility that scores might have dropped, it could be much better. If we extrapolate that percentage out to the whole list of 240 words, we can estimate that they might have learned about 14 words. If we stretch this assumption even further, which we cannot actually do based on the current evidence because knowledge of one word does not imply knowledge of another, and assume that students bring to the course the minimum 1800-word vocabulary required by the Japanese Ministry of Education for junior and senior high school students, then it comes out at an increase of just over 100 words. However, when we consider Nation’s (2001) estimate of an average native speaker acquiring approximately 1,000 words per year, this number seems woefully inadequate.

We can at least conclude that the test design was robust enough to not make major compromises in terms of validity and reliability when priority was given to practicality. We can also conclude that by carefully following simple guidelines, two different versions of the test could be made that both worked equally well. In other words, the method for producing the test is practical and reproducible without compromising reliability or validity.

References