Validating the Vocabulary Size Test
A Classical Test Theory Approach

BACKGROUND
The design of the VST

The Vocabulary Size Test (VST) (Nation & Beglar, 2007; Nation 2008) is a fairly new test of written, receptive English vocabulary size. It is a multiple-choice test that is designed to measure test-takers’ knowledge of the 14,000 most frequent word families of English. The test consists of 140 items, and there are 10 items for each of the 14 frequency bands from which the target words have been sampled.

An item in the VST looks like this:

33. candid: Please be candid.
   a. be careful
   b. show sympathy
   c. show fairness to both sides
   d. say what you really think

In each item, a test-taker is presented with the English target word, followed by a short non-defining sentence context where the word appears in bold. Below the targeted word, four options are given out of which one is the key.

The word lists that the VST is based on stem from the British National Corpus (BNC). Originally, frequency lists were developed based on the whole 100-million-word version, but these lists were subsequently revised using frequency figures from only the 10 million word spoken section of the BNC (Nation & Beglar, 2007). This was done to avoid the bias that the formal written nature of the BNC creates, coupled with the assumption that a spoken ordering more closely represents the order in which intended test-takers might learn English words. By multiplying a test-takers score on the VST by 100, an estimated vocabulary size measured in word families is assumed.

Initial reports on validity evidence of the test (Beglar, 2010), based on cross-sectional data from native speakers of English and four proficiency levels of Japanese learners of English, have argued that the test is largely a reliable and valid measure of vocabulary size. Based on a Rasch analysis, Beglar concluded that the VST is unidimensional to a strong degree, and that the vast majority of items function well as measures of the assumed construct. He also argued that measuring learners’ progress in vocabulary learning over time will likely be the greatest value of the test. In Beglar’s study, only a relatively small group of participants sat the whole test (< 50), and no longitudinal data were collected.

THE VALIDATION STUDY

Purpose

The purpose of the study was to investigate the reliability and validity of the monolingual version of the VST using a Classical Test Theory approach, based on data from a sizeable group of high intermediate and advanced Swedish learners of English. An additional purpose was to also use the test in a longitudinal design in order to investigate learners’ vocabulary size development over time.

Method

Participants

A total of 198 participants took part in the study. They were all university-level, full-time students of English at a university in southern Sweden. The great majority (n = 151) had just started their first term of study; students in their second term of study (n = 22) also sat the test, as did students in their third term of study (n = 25).

The instrument

The test was a paper-and-pencil version of the 140-item monolingual (English) Vocabulary Size Test. The test version used can be found in Nation & Beglar (2007) and Schmitt (2010).

Procedure

All participants (N = 198) sat the test for diagnostic purposes during the first week of the autumn term (September = T1) as part of a vocabulary module. Out of that original group, 94 participants sat the test again at the very end of the autumn term (January = T2). At T1, they were not told that a second administration was going to take place. The time between T1 and T2 was 4.5 months.

RESULTS FROM T1

Reliability and Item Discrimination

Cronbach’s Alpha: .905

Cronbach’s Alpha: .920

Item-total correlation values:

<table>
<thead>
<tr>
<th>ITC value</th>
<th>&lt;.0</th>
<th>.0-1</th>
<th>.1-3</th>
<th>.3-5</th>
<th>&gt;.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of items</td>
<td>11</td>
<td>19</td>
<td>45</td>
<td>83</td>
<td>2</td>
</tr>
</tbody>
</table>

RESULTS FROM T2 (cont.)

Mean Item Facility and Item Analysis

MeanItemFacilitywascomputedforeachofthe14frequencybands.Thehighestmeanwasobservedforfrequencyband2(MeanIF=.936)andthelowestforfrequencyband14(MeanIF=.419).

In terms of construct validity, a difficulty continuum was hypothesized, with an expected downwards slope from the highest frequency band (1K) to the lowest (14K). This is generally visible in Figure 1, but there are some anomalies (most notably bands 5, 8 and 11). There is also a specific problem in band 1 (related to item 10 basis). The main reason for the anomalies is believed to be the presence of cognates or near-cognates. A cognate analysis, shown in Table 1 below, gave support to this assumption.

Table 1. The number of English target words in each frequency band that has a cognate or near-cognate in Swedish

<table>
<thead>
<tr>
<th>Frequency band</th>
<th>Number of cognates</th>
<th>Number of near-cognates</th>
</tr>
</thead>
<tbody>
<tr>
<td>1K</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2K</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3K</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4K</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5K</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6K</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>7K</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8K</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9K</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10K</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>11K</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>12K</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>13K</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>14K</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

A distractor analysis furthermore revealed that the distractors in many of the cognate items did not attract many answers. A closer look at the items in bands 5, 8 and 11 showed that the following items were indeed very easy for the tested population:

1. (5K) the doorknob
2. (5K) the footbridge
3. (5K) the armchair
4. (5K) the towel
5. (5K) the magazine
6. (5K) the sheet
7. (5K) the table
8. (5K) the sofa
9. (5K) the bed
10. (5K) the desk
11. (5K) the chair
12. (5K) the book
13. (5K) the table
14. (5K) the chair
15. (5K) the desk
16. (5K) the book
17. (5K) the table
18. (5K) the chair
19. (5K) the desk
20. (5K) the book

CONCLUSIONS AND NEED FOR FURTHER RESEARCH

• The VST seems capable of yielding reliable scores, but some items are in need of revision
• A difficulty continuum based on target word frequency is visible but with some anomalies
• VST scores correlate highly with VLT scores (another widely used vocabulary size test)
• Classical Test Theory approaches to validation suffer from sample dependence
• Over a period of 4.5 months, Swedish university students of English increased their mean vocabulary size by 648 word families.
• There is a need for further validation studies that investigate whether learners truly know the proportion of words from the target domain to the extent suggested by VST scores.

REFERENCES


Henrik Gyllstad
Centre for Languages and Literature
Lund University
henrik.gyllstad@englund.lu.se

9th annual EALTA conference
Innsbruck, Austria
31 May – 3 June 2012