Vocabulary Learning and Instruction
Volume 1, Issue 1, August 2012
http://vli-journal.org

A Multiple-Choice Test of Active Vocabulary Knowledge

Jeffrey Stewart
Kyushu Sangyo University
doi: http://dx.doi.org/10.7820/vli.v01.1.stewart

Abstract

Most researchers distinguish between receptive (passive) and productive (active) word knowledge. Most vocabulary tests employed in second language acquisition (SLA), such as the Vocabulary Levels Test (VLT) and Vocabulary Size Test (VST), test receptive knowledge. This is unfortunate, as the multiple-choice format employed on most receptive tests inflates estimates of vocabulary size, and there are clear theoretical advantages to focusing instead on productive knowledge, which is associated with greater strength of knowledge as well as written and oral communication skills. This is in large part due to the logistical problems associated with such tests, as the full-word answers given must either be entered online or hand-marked. This paper will describe a multiple-choice format test of active vocabulary knowledge, in which learners confirm their knowledge of an English word by selecting its first letter. As there are 25 possible options, odds of guessing the correct answer by chance are reduced to 0.04. Findings of the study include that word difficulty estimates and scores are highly correlated to those of conventional, full-word active tests (>0.90), and that test reliability is higher on the proposed format than on that of a receptive test of the same words.

Keywords: vocabulary acquisition; productive vocabulary knowledge; language testing.

1 Background

Most researchers distinguish between receptive and productive word knowledge (e.g. Meara, 1990). Receptive, or “passive” knowledge, associated with the skills of reading and listening, involves retrieving meanings of words once they are presented, implying comprehension of input. Productive, or “active” knowledge, however, is associated with speaking and writing, and involves retrieval of form once the word is required in written or spoken contexts.

“Passive” tests of second language vocabulary knowledge are numerous, and include the Vocabulary Levels Test (VLT; Nation 1990) and the Vocabulary Size Test (VST; Nation & Beglar, 2007). Such tests typically operationalize assessment of receptive knowledge with a multiple-choice format, presenting word forms and asking learners to select the correct meanings from a list of options.

“Active” tests of knowledge, in which learners are asked to provide the word forms themselves, also exist, though appear to be used less frequently in practice. A drawback of such tests is that written answers typically must be hand-scored. In addition to being time-consuming, this procedure can also lead to inconsistencies between raters regarding accepted responses. To address these issues, such tests are

Please cite this article as: Stewart, J. (2012). A multiple-choice test of active vocabulary knowledge. *Vocabulary Learning and Instruction, 1*(1), 53–59. doi: 10.7820/vli.v01.1.stewart

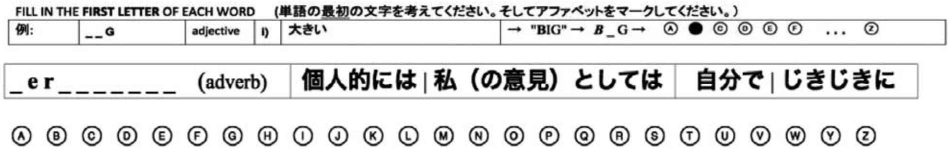
typically computer-delivered to expedite marking, and/or require correct spelling on answers. Though technology continues to become more widespread in educational institutions, at present the employment of computers in the testing of large groups still often poses logistical problems for educators who wish to test learners' vocabulary knowledge in applied contexts. Consequently, researchers and test-makers tend to employ multiple-choice formats, which are most suitable for assessment of receptive knowledge. This is unfortunate, as research indicates that although active knowledge is more difficult to test, the construct may play an important and distinct role in diagnosing language proficiency. Laufer and Goldstein (2004) administered tests of both active and passive vocabulary knowledge, and found that active tests of vocabulary were of considerably greater difficulty than those of passive knowledge. Consequently, the researchers concluded that the active test measured a construct of vocabulary "strength", which appeared to develop at higher stages of proficiency. This active construct of vocabulary knowledge could be of particular value in the assessment of sub-skills of language proficiency that extend beyond the receptive processing of input. In Japan, the bulk of English education involves the comprehension and translation of written texts, with few opportunities to produce language (see Barfield, 2012, this issue). While pressure to pass entrance exams has created a culture in which Japanese students are adept at writing multiple-choice tests, it is frequently observed that even students capable of high marks on English exams have considerably more difficulty expressing their own thoughts in spoken or written forms. Pedagogy that stresses the production of studied vocabulary could have greater washback on spoken and written proficiency.

A further consideration is the degree to which the multiple-choice format commonly employed in receptive-knowledge tests compromises measurement. Multiple-choice tests have become popular due to their practicality and cost-effectiveness rather than their statistical properties, as multiple-choice answer sheets can be collected and quickly scanned by a computer using optical mark recognition software. Yet despite these practical advantages, they remain most suitable for assessment of lower-order skills (Phelps, 1996), and pose various challenges to measurement. It has been acknowledged that guessing on multiple-choice tests can adversely affect test reliability (Zimmerman & Williams, 1965), and the multiple-choice format employed on vocabulary tests such as the VLT has been shown to inflate estimates of learner vocabulary size (Stewart & White, 2011).

Ideally, then, it would be desirable to find a vocabulary test format that held not only the theoretical and statistical advantages of a test of active recall, but also the practical advantages of a multiple-choice format that can be processed using a scanner and OMR software, without requiring either a computer for each student or a complex hand-marking procedure carried out by human raters on collected paper tests. The purpose of this paper is to describe such a test format, and compare its reliability and difficulty to both a conventional active recall test and to a popular measure of receptive vocabulary knowledge, the VST.

1.1 The KSU Active Multiple-Choice Test Format

The KSU Active Multiple-Choice Vocabulary test (hereafter referred to as the "Active MC") is a pencil-and-paper, OMR-scannable multiple-choice test of active vocabulary knowledge. The novel multiple-choice format is made possible with Remark OMR software, which allows custom design of answer sheets. Prompted by L1 definitions, the part of speech, the second and third letters in the word and the number of letters contained in the entire word, students are asked to provide the first letter of the target word by selecting the correct letter from multiple-choice bubbles that list every letter in the English alphabet (with the exception of the letter x, which is removed for brevity). Student instructions for the test format and an example item are depicted in Figure 1.



FILL IN THE **FIRST LETTER** OF EACH WORD (単語の最初の文字を考えてください。そしてアファベットをマークしてください。)

例:	__G	adjective	i)	大きい	→ "BIG" → *B*_G → Ⓐ ● Ⓒ Ⓓ Ⓔ Ⓕ ... Ⓩ

_er_______	(adverb)	個人的には | 私（の意見）としては	自分で | じきじきに

Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ Ⓗ Ⓘ Ⓙ Ⓚ Ⓛ Ⓜ Ⓝ Ⓞ Ⓟ Ⓠ Ⓡ Ⓢ Ⓣ Ⓤ Ⓥ Ⓦ Ⓨ Ⓩ

Figure 1. Example item and student instructions for the KSU Active Multiple-Choice Test.

Without the L1 definitions, the target word in the example shown in Figure 1 is essentially unrecognizable from the hints given about its form. This is by design. As the intention of the test is to assess active recall of L2 vocabulary, it is essential that the hints provided prompt recall of the word form, rather than recognition. The additional information simply prevents the selection of synonyms that could also correspond to the L1 translations provided.

2 Research Questions

This format could provide a variety of advantages for test-makers. As conventional distractors are not employed, the process of item-writing is greatly simplified, yet the format retains the advantages of a conventional paper-and-pencil multiple-choice test, as the format is machine-readable, and tests can be scored by computer. However, despite these advantages, a number of questions remain before the format can be used with confidence:

To what degree does selection of a word's first letter correlate with knowledge of the entire word?

The reader may note that simply providing the first letter of a word may not be equivalent to providing the entire word. While the correlation between the two forms of knowledge is unlikely to be perfect, we wish success on the first task to approximate success on the second as closely as possible.

How do statistical reliability and estimates of word difficulty under the proposed format differ to those of conventional test formats?

This format also holds the potential of reducing the measurement error associated with conventional multiple-choice formats: though options are given, the greatly extended number of choices restricts the probability of a correct answer due to random guessing to 0.04. Even if test takers become sophisticated in their strategies and choose more frequently occurring letters, the number of plausible options remains far higher than conventional formats. To test this hypothesis, however, it is necessary to compare test reliability to that of a test of the same words that uses a conventional multiple-choice format.

3 Results

3.1 Correlation to a Conventional Test of Active Knowledge

In order to compare scores on the proposed Active MC format to scores on a conventional active recall test in which the entire word must be provided (henceforth referred to referred to as the "Active Full Word" format), second year English conversation students at a Japanese university ($n = 205$) wrote a test of the second 1000 most common words in English ($k = 40$) employing the Active MC format. Upon completion, students were asked to then write the words in full. These full answers were then hand-marked. Only words written with correct spelling were marked as correct, as this practice approximates the scoring procedure of popular tests of active recall such as Laufer's CATSS test (Laufer & Goldstein, 2004). Tests in which answers were not provided for one of the formats were omitted from the sample. The resulting scores from the two formats were then compared.

The Active MC test had a mean of 33.70 points with a standard deviation of 5.88. The Active Full Word test had a slightly lower mean of 30.71 points, with a standard deviation of 7.67. The two test formats were highly correlated at 0.93, despite a somewhat restricted range of scores. This suggests that while the relationship is not perfect, the proposed Active MC test format closely approximates word knowledge under the active recall construct of vocabulary knowledge (Figure 2).

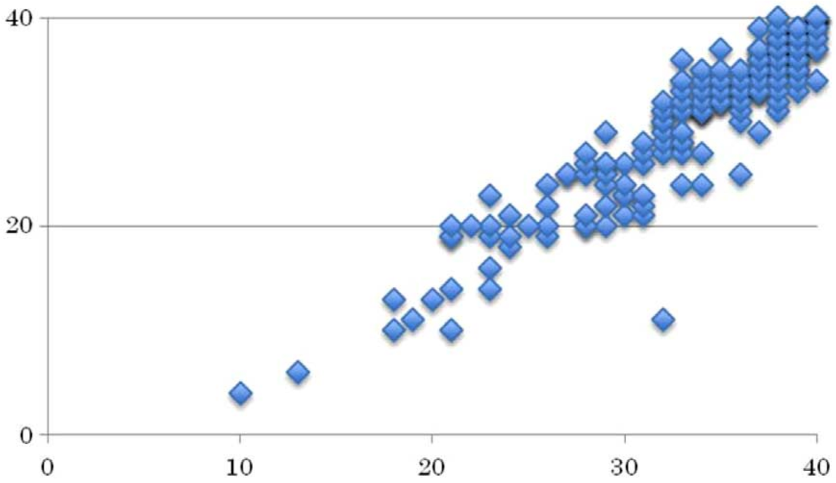


Figure 2. Correlations between Active MC format (x) and full-word active recall format (y).

3.2 Comparing Reliability and Estimates of Word Difficulty Between Formats

The first three 1000-word levels ($k = 30$) of a Japanese-bilingual version of the VST (Nation & Beglar, 2007) was adapted to the active recall format by using the correct Japanese definitions on the VST as prompts for recall of word. This created two tests of the same words, allowing for comparisons of the two formats.

The active recall test was given to a sample of first and second-year English conversation students at a Japanese university ($n = 119$). The tested group's scores on the TOEIC Bridge Test ranged from 90 to 140. The group was asked to select the first letters of the tested words and to then write the words in full. These full answers were then hand-marked with the entire answer examined rather than simply the first letter, resulting in a conventional measure of active vocabulary knowledge ("Active Full Word" format). Once again, only words written with correct spelling were marked as correct, and tests in which answers were not provided for one of the formats were omitted from the sample.

This allowed for direct comparisons between active recall formats, but a challenge of the study was that the VST test format provides the word form and requests a definition, whereas the active recall formats provide the definition and request the word form. Therefore, it was not possible to administer both tests to the same students, as one test would reveal the answers of the other. Consequently, the Japanese-bilingual VST-format test was given to a second group of students of comparable proficiency ($n = 125$). Correlations, item difficulties and split-half reliabilities between the resulting three test formats were then compared. Means, standard deviations and Cronbach alpha reliabilities for the three tested formats are listed in Table 1.

Table 1. Descriptive Statistics for VST and Active Test Formats ($k = 30$)

	Mean	*SD*	Reliability
VST	19.54	3.11	0.60
Active MC	12.00	3.48	0.70
Active Full Word	7.73	3.77	0.81

At 19.54 points with a Cronbach alpha value of 0.60, the VST-format test had the highest mean and the lowest reliability. The Active Full Word test of the same words had a dramatically lower mean of only 7.73, though reliability was substantially higher, with a Cronbach alpha of 0.81. The primary test format under investigation, the Active MC, ranks between the two both in terms of difficulty and internal reliability.

Score correlations between the VST format and Active format tests were not possible, as the two formats were written by different groups of participants. However, it is possible to correlate item difficulties for each word, to determine if word difficulties are consistent between test formats (see Table 2). Though lower between the original VST and the Active Recall formats, the correlation between the Active Recall tests was again very high, at 0.90.

Table 2. Item Difficulty Correlations Between Test Formats

	VST	Active MC	Active Full Word
VST	1		
Active MC	0.61	1	
Active Full Word	0.63	0.90	1

In order to examine relationships between test formats and estimated word difficulty in detail, item facilities for the tested words under all three formats are graphed in Figure 3. An item facility of 1.00 indicates that all respondents answered correctly; an item facility of 0 indicates none did.

There are clear differences between difficulty estimates under the VST format and the two Active formats. For the first 10 items, which test the first 1000 most common words in English, the VST and Active formats behave similarly. But from the second 1000-word level, results begin to diverge. With the exception of "reward" which was very difficult under all formats, few words on the VST format demonstrated item facilities below 0.2, indicating that chance guessing on the multiple-choice VST format does, in fact, affect item difficulty estimates. Indeed, some words that almost no students could guess under the Active format, such as "nil", still had high item facilities under the VST format. This suggests that when learners encounter a word they do not know under the VST format, they tend to guess, inflating estimates of words known receptively.

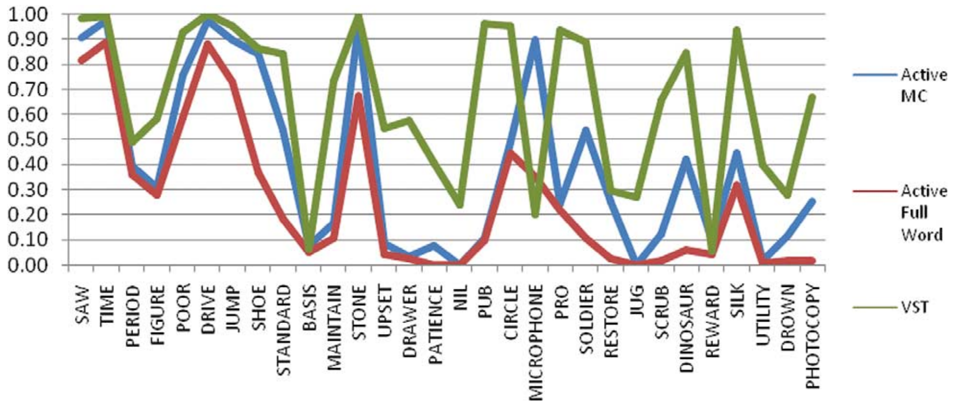


Figure 3. Word difficulty estimates under Active MC, Active Full Word and VST test formats.

The Active MC and Full Word item facilities correlated highly at 0.90, but differences between the two revealed hazards of the proposed Active MC format. One word, "Microphone", was markedly easier under the Active MC format. This is likely because the Japanese word for "microphone" is simply an abbreviation of the English term ("マイク", or "Mic"), and therefore, respondents could simply select the letter "m" without knowing the word's full form. This indicates that the Active MC format is not appropriate for testing of loan words.

There were, however, situations in which word difficulty estimates could arguably be said to be more accurate under the Active MC format than the Active Full Word format, because the Active MC format did not penalize learners for incorrect spellings. The word "dinosaur" serves as an example: though under the Active Full Word format few students knew it, an examination of responses revealed that a high proportion likely did, but simply could not master the word's more challenging spelling. Some alternative responses included "dinosaw". Though these answers were technically incorrect, it is likely that in spoken contexts the learner would be considered to have mastery over the words, albeit with differences from native speaker pronunciation.

4 Conclusions

The proposed Active MC test format returned high correlations (≥ 0.90) with the conventional, Active Full Word format. Though the Active MC format's reliability of 0.70 was lower than the Active Full Word reliability of 0.81, it was also markedly higher than that of the VST format (0.60). This suggests that although the Active Full Word test format remains the most statistically reliable, in contexts in which hand-marking tests is not feasible, the Active MC format provides an alternative which, in addition to being highly correlated to conventional active vocabulary knowledge estimates, is easy to construct, easy to mark, and offers higher reliability than a conventional multiple-choice test of the same words.

Though the Active Full Word format was slightly more difficult than the Active MC format, it is unclear how much of this is due to guesswork on the Active MC format, and how much is due to the additional difficulty imposed on the conventional active recall format by requiring test takers to correctly spell all words. As provision of words' correct spellings is a necessary condition of many such tests, it is possible that these tests can underestimate vocabulary knowledge on spoken productive measures, by conflating spelling with knowledge. A closer analysis of the Active Full Word format is required to determine if this is the case.

References

Barfield, A. (2012). Lexical development and learners' practices in a content-based learning course. *Vocabulary Learning and Instruction*, *1*(1), 10–19. doi:10.7820/vli.v01.1.barfild

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436. doi:10.1111/j.0023-8333.2004.00260.x

Meara, P. (1990). A note on passive vocabulary. *Second Language Research*, *6*(2), 150–154. doi:10.1177/026765839000600204

Nation, I.S.P. (1990). *Teaching and learning vocabulary.* Boston, MA: Heinle and Heinle.

Nation, I.S.P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

Phelps, R.P. (1996). Are U.S. students the most heavily tested on earth? *Educational Measurement: Issues and Practice*, *15*(3), 19–27. doi:10.1111/j.1745-3992.1996.tb00819.x

Stewart, J., & White, D.A. (2011). Estimating guessing effects on the vocabulary levels test for differing degrees of word knowledge. *TESOL Quarterly*, *45*(2), 370–380. doi:10.5054/tq.2011.254523

Zimmerman, D.W., & Williams, R.H. (1965). Effect of chance success due to guessing on error of measurement in multiple-choice tests. *Psychological Reports*, *16*(3c), 1193–1196. doi:10.2466/pr0.1965.16.3c.1193