Language Testing in Asia
a SpringerOpen Journal

RESEARCH Open Access

Relationship between types of distractor and difficulty of multiple-choice vocabulary tests in sentential context

Yuko Hoshino

Correspondence:
yukohoshino@live.jp
Department of Management, Tokyo Fuji University, 1-7-7 Shimoochiai, Shinjuku, Tokyo 161-8556, Japan

Abstract

This study compares the effect of different kinds of distractors on the level of difficulty of multiple-choice (MC) vocabulary tests in sentential contexts. This type of test is widely used in practical testing but it has received little attention so far. Furthermore, although distractors, which represent the unique characteristics of MC tests, are known to influence test difficulty, studies have focused only on the semantic relationship between target words and distractors in vocabulary tests. By also considering the words that are syntagmatically related to the words in context, this study contrasted distractors relating to target words and those relating to context characteristics, and compared three MC vocabulary tests with different types of distractors: distractors with a paradigmatic relationship to the words in the correct answer, distractors with a syntagmatic relationship to contexts, and unrelated distractors. The results suggested that tests with syntagmatically related distractors were generally the most difficult, and tests with unrelated distractors, the easiest; the paradigmatically related distractors remained in the middle. However, this difference disappeared when the test takers could not use contextual information, which indicates that test takers strongly rely on contextual information in taking multiple-choice vocabulary tests in context.

Background

Vocabulary researchers have shown interest in different aspects of vocabulary knowledge. Some are curious about the number of words learners know and others, about learners' ability to find collocations between two words. To serve the researchers' needs, various vocabulary tests have been produced to measure each aspect of vocabulary knowledge for a more accurate assessment of the progress of learners' vocabulary acquisition. The most widely used tests are those on vocabulary size, which aim to examine how many L2 words the learners can match with their meanings (e.g., Nation, 2001; Schmitt, Schmitt, & Clapham, 2001). Depth of vocabulary tests are also gaining popularity, and the number of studies using them is increasing (e.g., Nurweni & Read, 1999; Qian, 1998; Qian & Schedl, 2004; Read, 1993).

However, though quite popular in research, vocabulary size and depth tests are not often employed in practical testing situations. Instead, many large-scale tests, such as the Cambridge English for Speakers of Other Languages (ESOL) examination, Test of English as a Foreign Language (TOEFL), Test of English for International

Communication (TOEIC), Society for Testing English Proficiency (STEP Eiken), and National Center for University Entrance Examinations in Japan, adopt vocabulary tests in various contexts because researchers have reached a close consensus on testing vocabulary in context. Presenting words in context increases the authenticity of the test; in fact, Read (2000) claims that words should almost always be shown in context. Yet, despite the trend toward contextualized vocabulary tests, they have not been researched in detail.

The multiple-choice (MC) format has been widely used in language testing, mainly because it is much easier to score than other test formats, such as open-ended tests. Since there is only one definite answer in MC, there is no need to worry about inter- and intra-rater reliability, and computer scoring can be readily adopted. However, despite its advantages, many researchers have criticized the MC test for two major reasons. One is that an MC test tends to induce a guessing strategy and is thus less authentic. Some researchers (e.g., Cohen 1988a; Cohen and Upton, 2006; Farr et al., 1990; Nevo, 1989; Nikolov, 2006; Paul et al., 1990) have found that test takers employ a choice-oriented process, which never occurs in conditions without choices.

The other reason is that making distractors, especially plausible ones, is extremely difficult. Although textbooks have suggested including plausible distractors in guidelines or rules for writing MC items (Haladyna and Downing, 1989; Haladyna, et al., 2002)—plausibility being one of the most important characteristics of distractors (e.g., Celce-Murcia et al., 1974)—there are no reliable references about distractors because little research has been conducted on this topic. Therefore, there is an urgent need to investigate the influence of distractor characteristics in MC vocabulary tests in context. This study focuses on the change of test difficulty by different types of distractors.

Literature review

Researchers have reached a close consensus on testing vocabulary in context. Supported by Read's (2000) claim that words should almost always be presented in context, most large-scale tests have adopted vocabulary tests in various contexts. Vocabulary tests in context have a long history, going back at least as far as the 1930s, and researchers have long been interested in the difference between isolation tests and context tests. Stalnaker and Kurath (1935) and Kurath and Stalnaker (1936) were the first researchers who tried to detect differences between the two, and they showed that both were equally valid. However, a critical shortcoming exists in their studies; that is, they used different formats for the isolation and context tests (i.e., MC in the isolation tests and open-ended questions in the context tests). Thus, they could not make a pure contrast between the isolation and contextualized tests. After these studies, the subject of vocabulary tests in context was neglected as a research issue. However, in 1976, the subject started to attract attention again because TOEFL changed the format of its vocabulary section from tests out of context to tests in context. Since then, vocabulary tests in context have become very popular worldwide, probably because presenting words in context can change students' learning style of memorizing only synonyms or translation, which often occurs when testing words in isolation (Read, 2000, p. 140). Moreover, the presence of context resolves any semantic ambiguity of the target words (Alderson et al. 1995, p. 48); hence, knowledge of the target word's meaning can be assessed.

After the TOEFL revision, Pike (1979) conducted a large study on vocabulary tests in context, wherein he examined the correlation coefficients between sections of TOEFL. He used the following vocabulary test formats: words in isolation, words in context, and sentence completion, in which one of the words in the target context is replaced by a pair of brackets. Overall, there were at least moderate correlation coefficients in all of the listening, structure, reading, and writing sections; the vocabulary tests had high intercorrelations. Henning (1991) used a larger variety of vocabulary tests in context than Pike did by changing the following characteristics: the length of context relating to whether the test takers can infer the answer from context, the formats (matching or supplying), and the location of options (passage-embedded or not passage-embedded). The comparison of eight vocabulary tests showed moderate-to -high correlations between all of the tests, and the difficulty of the tests did not differ significantly. Hence, both Pike and Henning found that vocabulary tests in isolation and in context had similar characteristics.

However, a series of studies by Mori (1999; 2002; 2003; Mori & Nagy, 1999) produced different results from those of Pike (1979) and Henning (1991). She used Japanese vocabulary tests and compared three conditions: kanji only, context only, and kanji and context. Using Pike's and Henning's terms, the three conditions can represent words in context (kanji only); sentence completion or supplying format (context only); and words in context or matching format (kanji and context). Contrary to Pike and Henning, Mori (2002; 2003) found that the scores of the kanji and context condition were twice as high as those of the context-only condition. The differences between the test conditions are due to the number of possible clues that the test formats contain. For instance, a matching format includes two clues (the target word and the context), whereas a supplying format has only one (the context). Because Mori used identical contexts for the context-only and kanji-and-context conditions, it can be concluded that the context contained sufficient clues; otherwise, her participants would not have been able to obtain correct answers in the context-only condition. That is, the target words in Mori's kanji-and-context condition are easily inferable from the contexts provided. In this sense, Henning did not control whether his contexts in matching format contained useful clues for the participants; therefore, the results of Mori and Henning were not consistent.

Results similar to Mori's were achieved by Hoshino (2008). She used three types of English vocabulary tests in context: matching format whose context cannot be clues, matching format whose context can be clues, and supplying format (similar to Mori's kanji-only, kanji-and-context, and context conditions, respectively). Her results supported Mori's findings that tests containing two clues were easier than tests with only one clue. Hoshino also calculated standardized partial regression coefficients from knowledge about vocabulary meaning, paradigmatic knowledge, syntagmatic knowledge, and reading ability for the three types of vocabulary tests in context introduced above. Her findings indicated that tests using the matching format were significantly influenced by the knowledge of word meanings and paradigms, whereas tests in supplying format were significantly affected by syntagmatic knowledge. Hence, when target words exist in the context (i.e., matching format), knowledge of the word meanings is influential. On the other hand, when target words do not exist in the context and participants have to find the word that fits the context (i.e., supplying format), the tests have a relationship with syntagmatic knowledge, as they require test takers to find a collocation between the choice and the words in the context.

Integrating the results above, it can be safely said that vocabulary tests in isolation and in context differ in difficulty when we control inferability from context. In addition, among the vocabulary tests in context, matching and supplying formats are possibly affected by different factors.

The studies introduced so far dealt with context and word factors (i.e., words in context) but did not take distractors into consideration, although most of the studies above used the MC format. Distractors are specific to tests with choices, so it is inevitable to associate them with MC tests. Research about distractors has been done on the following topics in various MC tests: the number of options (e.g., Bruno & Dirkzwager, 1995; Green et al., 1982; Haladyna & Downing, 1993; Rogers & Harley, 1999; Shizuka et al., 2006; Trevisan et al., 1991); position of correct options among the choices (Attali & Bar-Hillel, 2003); and surface characteristics of distractors, such as the length or number of content words (e.g., Davey, 1988; Drum et al., 1981; Freedle and Kostin, 1993).

Another popular topic of studies with vocabulary tests is the relationship between distractors and test difficulty. For example, Greidanus and Nienhuis (2001), Jenkins et al. (1989), Marshalek (1981), and Nagy et al. (1985) compared distractors that were semantically related to the target words and distractors that were not; all of them concluded that test takers had more difficulty in answering the tests with semantically related distractors. However, all of the studies above used vocabulary tests in isolation, not in context, so that it is necessary to confirm whether their results also apply to vocabulary tests in context. One exception is Goodrich (1977), who conducted research on attractive distractors in MC vocabulary tests in context in matching format. He compared eight distractors and discovered that synonyms, contextually related words, and antonyms were more attractive than other distractors. Goodrich's study differed from the others, as he found that semantically related words (synonyms and antonyms) and words relating to context were plausible distractors. Hence, the difficulty of vocabulary tests in context could be controlled by using either semantically related distractors or distractors relating to the contexts. However, no study has examined this issue to date. What is more, distractor effects on vocabulary tests in context in supplying format have not been investigated. As can be seen in Hoshino's study (2008), tests using matching and supplying formats are affected by different factors. Hence, different results from those of Goodrich might be obtained. Furthermore, the supplying format has been widely used (e.g., TOEIC, Cambridge ESOL exam, STEP Eiken), and the results of the supplying format will also be beneficial to cloze tests and gap-filling tests, both of which require test takers to fill in the blanks with suitable words.

Hence, the focus of this study was to investigate whether there was a difference in the difficulty of vocabulary tests in context in supplying format. Three types of distractors were used: paradigmatically related distractors (PARA), syntagmatically related distractors (SYN), and control distractors that had no semantic or syntagmatic relationship (CON [see Materials Section for examples]). In a paradigmatic relationship, a word is classified under the same part of speech as the target word and has a semantic relationship with it (e.g., Wolter, 2001); therefore, Goodrich's (1977) synonyms and antonyms, and the semantically related distractors in Greidanus and Nienhuis (2001), Jenkins et al. (1989), Marshalek (1981), and Nagy et al. (1985) can be included in PARA. Regarding SYN, this study used words that co-occurred with one of the words in the context (see Materials Section). Although the definition of a syntagmatic relationship is narrower than that of a contextual relationship in Goodrich, the former enjoys an advantage over the latter in that an objective criterion can be used since the strength of syntagmatic association can be calculated by using a large corpus. Therefore, it is easily replicable. Most studies have compared only semantically related (i.e., PARA here) and unrelated distractors (i.e., CON). However, by including SYN, we could contrast between distractors related to the target words and distractors related to the context (to be more specific, collocation).

Method

Participants

A total of 372 students from seven universities in Japan participated in this study. They had learned English for at least six years at the time the study was conducted. Every year, around 500,000 students take the university admission test made by National Center for University Entrance Examinations in Japan (National Center for University Entrance Examination, n.d.). The test includes various subjects, such as Japanese, Mathematics, Sciences, and foreign languages (there are five languages to choose from, and most test takers select English). Since the test assesses vocabulary knowledge in sentential contexts in supplying format, it was assumed that the participants were used to the test format adopted by this study. The participants were randomly grouped into (a) those who took the test with paradigmatically related distractors (PARA group: 123 participants), (b) those who took the test with syntagmatically related distractors (SYN group: 131 participants), and (c) those who took the test with distractors with no relationship to (a) or (b) (CON group: 118 participants).

Materials

A vocabulary proficiency test and experimental test were created for this study. In the vocabulary proficiency test, 28 items were used from the vocabulary and grammar sections in the 2nd, pre-2nd, and 3rd grades of the STEP Eiken test conducted in January 2005. The test formats were MC tests in context in supplying format, and their distractors were similar to those in STEP Eiken. Originally there were 55 items, but 11 items that measured grammatical knowledge were excluded after having been assessed by the author and another researcher (Cronbach's $\alpha = 1.00$), leaving 44 items. After that, 118 university students who did not participate in the main study answered the remaining vocabulary items to select the best items for the vocabulary proficiency test. Misfit items (including underfit and overfit items) were excluded by the rationale of Bond and Fox (2007, p. 240), who determined that mean squares of over 1.3 were underfit and below 0.75, overfit. Consequently, 28 items (13, 9, and 6 in the 2nd, pre-2nd, and 3rd grades, respectively) were chosen for the vocabulary proficiency test. All participants who joined the main experiments took the proficiency test.

The format of the experimental test was almost identical to that of the vocabulary proficiency test; that is, there was a blank in the sentential context (supplying format) and participants would select one of four choices. The words in the choices were of high frequency so that the tests would match the estimated English ability of the participants. All test words were chosen from levels 1 and 2 of the JACET 8000 word frequency list (JACET, 2003), which was created especially for Japanese learners of English. Since each level has 1,000 words, the experimental test was composed of the most frequent 2,000 words.

The experimental test differed from the vocabulary proficiency test in regard to the distractors. To compare the difficulty of tests with different types of distractors, three tests were prepared with three kinds of distractors: (a) distractors with a paradigmatic relationship to the words in the correct answer, (b) those with a syntagmatic relationship to the words in the sentential contexts, and (c) those with no relationship to (a) or (b). To select PARA, a word whose meaning was related to the correct option was chosen through WordNet 3.0 (Princeton University Cognitive Science Laboratory, 2006), which displays the target words' cognitive synonyms, antonyms, troponyms, hypernyms, verb groups, derivationally related forms, sentence frame, entailment, and sister terms. SYN was a word collocating with a word that was in the same proposition as the choice words. In other words, SYN collocated with the content word that had the strongest semantic connection with the word in the context, since a proposition is a basic language unit (Kintsch, 1998, p. 37). The examples below provide further clarification.

Tests with PARA

I had started to feel very tired and I decided to try and find somewhere to ( ) a sleep.

a. experience
b. have (correct)
c. own
d. receive

Tests with SYN

I had started to feel very tired and I decided to try and find somewhere to ( ) a sleep.

a. lose
b. have (correct)
c. need
d. want

Tests with CON

I had started to feel very tired and I decided to try and find somewhere to ( ) a sleep.

a. guess
b. have (correct)
c. order
d. vote

Experience, *own*, and *receive* appear in the index of *have* (correct answer) in WordNet 3.0, so *have* was chosen as PARA. For SYN, a verb which strongly collocates with *sleep* was selected. When the correct answer was put in the blank, the proposition of [POSSESS I SLEEP] (meaning, *I have sleep*) is formed, based on Kintsch (1998) and Bovair and Kieras (1985); therefore, among the words in the sentence above, *sleep* is supposed to have the strongest relationship with the word that should be in the blank. The formed proposition was confirmed by two people with knowledge of this proposition.

Hence, in order to identify the words strongly collocated with *sleep*, the log-log score within five words of *sleep* was calculated, and *lose*, *need*, and *want* were selected. According to Smadja (1993, p. 151), 'most of the lexical relations involving a word *w* can be retrieved by examining the neighborhood of *w*, wherever it occurs, within a span of five (−5 and +5 around *w*) words"; hence, five words were considered enough to assess the strength of collocations. The log-log score was used instead of the mutual-information (MI) score to measure the strength of collocation because the MI score tends to give higher scores for low-frequency words. The target words in this study of high frequency, so it was better to avoid using the MI score. All SYN were selected from the 100 strongest collocations with the words in contexts (i.e., *sleep* in the example above). For CON, *guess*, *order*, and *vote* were chosen after it was confirmed that they had neither a paradigmatic nor syntagmatic relationship by the use of the above criteria. All of the distractors had the same word class as the correct answer, as well as the same frequency level in the JACET 8000 (JACET, 2003). Care was taken to ensure that PARA and SYN were exclusive; that is, PARA did not have the characteristics of a syntagmatic relationship and SYN did not have the characteristics of a paradigmatic relationship. If some distractors had both characteristics, the results would be contaminated. After all the distractors had been selected, three native English speakers checked the tests and confirmed that only the words chosen as the correct answers could fit into the contexts. Accordingly, three tests with 40 items each were created, and the contexts in all three were identical. In other words, the three tests were the same except for the distractors (see Additional file 1).

Procedures and data analysis

The vocabulary proficiency test was conducted in 15 minutes and the experimental test, 25 minutes. The experimental test was randomly distributed to the participants, so each participant took one of the three types. The test with PARA was taken by 123 participants; the test with SYN, 131 participants; and the test with CON, 118 participants. Time allotments were decided by a pilot study; more than 95% of the participants finished the tests within the allotted time.

The dependent variable was the test score and the independent variables, the test format and participants' ability. Therefore, this study has a 3 x 3 design (i.e., distractors: PARA, SYN, and CON x proficiency: upper, middle, and lower).

Results and discussion

Before conducting an analysis of variance (ANOVA), validity was examined through reliability (Cronbach's α) and unidimensionality. The reliability for the vocabulary proficiency test was .82; the test with PARA, .71; the test with SYN, .79; and the test with CON .84. The result of the test with paradigmatically related distractors was not high as that of an MC test, but it could be regarded as having moderate reliability. In terms of unidimensionality, the results by Winsteps (Linacre, 2006) indicated that all of the tests were unidimensional since most of the items passed the criterion of Bond and Fox's (2007) 0.75 to 1.30 in the infit mean square. Only one item in the test with unrelated distractors was underfit (1.44); the other items passed the criterion. From the results, therefore, the tests were regarded as having sufficient validity from the viewpoint of reliability and unidimensionality.

Next, the participants were classified into three groups on the basis of their scores in the vocabulary proficiency test: 21 and above, upper group; 16 to 20, middle group; and 15 and below, lower group. The descriptive statistics of the vocabulary proficiency test are shown in Table 1. The results of a one-way ANOVA show a significant difference between the upper and middle, middle and lower, and upper and lower groups, implying that the proficiency levels were clearly separated by the vocabulary proficiency test.

To compare the difficulty of the test formats, a 3 (distractor) x 3 (proficiency) two-way ANOVA was conducted on the experimental test. A summary of the descriptive statistics is presented in Table 2.

Both main effects were significant. With regard to proficiency, there was a significant difference between each group and, on the basis of Cohen's (1998b) framework, the effect size was large: F (2, 372) = 166.28, p = .00, η^2 = .41. As predicted, the upper group obtained the highest score and the lower group, the lowest. In terms of distractors, the effect size was medium: F (2, 372) = 31.60, p = .00, η^2 = .08. Tukey's HSD (honestly significant difference) test showed that the test with CON was significantly easier than the test with PARA, which, in turn, was significantly easier than the test with SYN. However, this tendency was not maintained in the upper group, as can be seen in the descriptive statistics in Table 2; the test with SYN was easier than the test with PARA, although the difference was extremely small. This was because the interaction between the two factors approached significance, although it was not statistically significant and the effect size was small: F (4, 372) = 2.19, p = .07, η^2 = .02. The overall results suggested that the test with CON was the easiest, supporting the findings of research using vocabulary tests in isolation. Additionally, the PARA's being easier than SYN (a trend seen in the middle and lower groups) was similar to the results of Greidanus and Nienhuis (2001). Furthermore, these findings indirectly support the premise that there

Table 1 Descriptive statistics of the proficiency test for each proficiency level

		n	Mean	*SD*
PARA group	Upper	29	23.21	1.99
	Middle	43	18.21	1.49
	Lower	51	12.33	2.69
	Total	123	16.95	4.84
SYN group	Upper	32	24.25	2.03
	Middle	48	17.90	1.45
	Lower	51	11.69	2.75
	Total	131	17.03	5.38
CON group	Upper	45	23.93	2.03
	Middle	32	18.19	1.62
	Lower	41	12.02	2.51
	Total	118	18.24	5.51

Table 2 Descriptive statistics of the experimental test for each proficiency level

		n	Mean	*SD*
PARA	Upper	29	23.00	4.63
	Middle	43	18.47	4.44
	Lower	51	14.39	3.11
	Total	123	17.85	5.21
SYN	Upper	32	23.28	5.95
	Middle	48	15.69	4.38
	Lower	51	12.64	3.46
	Total	131	16.36	6.11
CON	Upper	45	25.31	5.01
	Middle	32	17.69	3.60
	Lower	41	14.48	4.54
	Total	118	18.63	6.76

are paradigmatic and syntagmatic links in learners' mental lexicons (Meara, 2009), because these links are thought to impede the ability to answer correctly.

The findings of this study support the theory that vocabulary tests in context in supplying format are clearly context-dependent tests and that the participants used collocation knowledge when they took the test. If they had not, SYN would have been no more plausible than unrelated distractors. The test with SYN was more difficult than the test with PARA for the middle and the lower students; that is, distractors relating to context (SYN) were more plausible than distractors relating to the target word (PARA). Why, then, were test takers more attracted to SYN than to PARA? The reason was that MC vocabulary tests in sentential contexts would themselves require syntagmatic rather than paradigmatic knowledge. In Hoshino (2008), it was found that syntagmatic knowledge and reading ability were significant predictors in a supplying format (which this study adopted), but this was not the case in a matching format. This means that the participants relied on collocation knowledge when they answered supplying-formatted MC vocabulary tests. In Goodrich (1977), who used matching-format MC vocabulary tests in context, participants were attracted to contextually related distractors, false synonyms, and antonyms—the latter two being classified as having a paradigmatic relationship. Hence, in matching formats, participants tend to be attracted to both paradigmatically and contextually related distractors. In supplying formats, participants are also attracted to both distractors, but are more attracted to syntagmatically related distractors. Therefore, tests with SYN were more difficult than tests with PARA because test takers would rely on contextual information when taking MC vocabulary tests in supplying formats.

To confirm this hypothesis, further analysis was conducted. If participants could not utilize information about collocation, the effects of SYN might not be as strong as they were. There was a high possibility that participants would fail to find collocations when there were unknown expressions in the proposition. Words in SYN were chosen by the criteria that the test takers collocated with the word in the same proposition (see Materials Section), and when they did not know the expressions in the proposition, they would likely fail to detect the relationship between the word in the proposition and the words in SYN. Let us consider the example below.

Entries for the awards ( ) at the beginning of October.

a. close (correct)
b. judge
c. read
d. send

The correct answer in the item above is *close*, and other choices (i.e., *judge*, *read* and *send*) had a syntagmatic relationship with *entry* because the proposition [Close Entry] meaning *entry closes* based on Kintsch (1998) and Bovair and Kieras (1985) is formed. However, it is highly possible that the participants missed finding these collocations when they did not know *entry*. In this case, it is hypothesized that SYN did not attract test takers; as a result, tests with SYN would not be significantly more difficult. To investigate this issue, data from another 143 participants was collected. The additional participants took 13 items each from tests with PARA, SYN, and unrelated distractors. These items were different from each other, so none of the contexts overlapped. Care was taken that the items selected had a difficulty level that represented the original 40 items. After answering the test, the participants marked unknown words or expressions that they had encountered, and one-way ANOVAs were conducted to compare the items with different kinds of distractors when there were unknown expressions within the same proposition as the blank (i.e., assuming the words in the correct option were in) and when there were none. Taking the example above, the comparison was between those who knew *entry* and those who did not. Descriptive statistics are shown in Table 3.

As expected, the results showed that item difficulty did not change when there were unknown expressions in the proposition: F (2, 109) = .05, p = .95, η^2 = .00; whereas, when there were none, a significant difference appeared: F (2, 665) = 10.89, p = .00, η^2 = .02. Items with CON were significantly easier than items with PARA and SYN (p = .00 for both), but there was no significant difference between items with PARA and SYN (p = .44). These results indicated two things. First, when there were unknown expressions in the proposition, the test takers had more difficulty in finding the collocations. In that situation, SYN did not work as it normally does, so there was no difference of proportions correct between items with different distractors. On the other hand, when there were no unknown expressions in the propositions, there was a greater possibility of finding the collocations. In this situation, SYN became more attractive than unrelated distractors, leading to greater difficulty in

Table 3 Descriptive statistics with different types of distractors when unknown expressions did and did not exist in the proposition (proportion)

	Distractor	*n* of items	*M*	*SD*
With unknown expressions in proposition	PARA	13	43.18	50.11
	SYN	13	44.90	50.25
	CON	13	47.37	51.30
Without unknown expressions in proposition	PARA	13	62.96	48.40
	SYN	13	68.25	46.67
	CON	13	81.74	38.71

items with SYN. This finding strongly supports the theory that test takers use collocation knowledge in vocabulary tests in context in supplying format.

Second, items with PARA followed the same pattern as those with SYN. That is, there was no significant difference when unknown expressions existed, but they were significantly more difficult than in the tests with CON when there were no unknown expressions in the propositions. If PARA were weakly related with the surrounding context, it was believed that it would follow the pattern of CON, but this was not the case. Therefore, like SYN, PARA seem to have an association with the words in the proposition in context. The possible reason is that when choices have similar meanings, many choices, if not all, are consistent with the meaning of the surrounding context, at least to some degree. Therefore, test takers have to utilize grammatical knowledge (e.g., whether a specific verb can take an object or not) and collocation knowledge in judging which choices fit into the context. Let us consider a sample item from the test with PARA.

We surely have all faced the ( ) that the world is in real danger of overpopulation.

a. detail
b. example
c. fact (correct)
d. information

In this example, the distractors, especially *information*, match the meaning of the context, but the test takers have to find a collocation between *face* and *fact* in order to pick the correct answer. Those who lack collocation knowledge or do not know the meaning of *face* might consider both *fact* and *information* to be right, and as a result, they might fail to choose the correct answer. For this reason, tests with PARA are affected by the same factor (i.e., knowing or not knowing the expressions in the proposition) that affects tests with SYN. Hence, even with PARA, test takers have to use collocation knowledge.

Conclusion

The purpose of this study was to investigate the effect of distractors on the difficulty of MC vocabulary tests in sentential contexts. The results showed that the distractors relating paradigmatically to the words in correct answers and distractors relating syntagmatically to the words in context were both more plausible than unrelated distractors. Therefore, the tests with these distractors were more difficult than the test with unrelated distractors. Furthermore, SYN was more attractive than PARA, especially for the middle- and lower-proficiency learners. The results of SYN and CON are strong proof that the participants relied on knowledge about collocation when they took this kind of vocabulary test. The disappearance of the significant difference when there was an unknown expression in the proposition is further evidence of this finding.

The most important implication of the results of this study is that it is important to investigate the construct of vocabulary tests in context. This study found that some of the participants used collocation information when taking such tests, but no tests adopting this test format mention collocation in test rubrics. Rather, they remain ambiguous on this point by stating that they measure reading ability (e.g., TOEFL), vocabulary knowledge (e.g., STEP Eiken), or vocabulary knowledge included in the reading section (e.g., KET in Cambridge ESOL and TOEIC). However, since vocabulary knowledge and reading ability contain a wide range of concepts, it is necessary to determine the specific construct. When researchers want to clarify what these tests measure, they should keep in mind that the construct may change depending on the types of distractors and context information that is necessary in selecting correct answers, even if the two are closely related. For example, Bachman (1985), who used cloze tests, classified items into four types, according to the amount of information needed for answering: within clause, across clause within sentence, across sentences, and extratextual. In this case, “within clause” might require collocation knowledge more than the other three because it is possible to find collocations, say, between a noun and a verb, and thus give the correct answer. In terms of MC cloze and the tests this study adopted, test takers can select the correct answer without reading the full sentence. They only need to read a specific clause if the distractors do not make sense when they are replaced by brackets. Therefore, defining the construct of these versions of the vocabulary tests will provide useful information because of the popularity of vocabulary tests in context.

The results of this study are not only helpful in creating MC vocabulary tests in context, but also contribute to assessing partial vocabulary knowledge. Vocabulary knowledge is known to be continuous but not dichotomous, but no best method has yet been investigated to measure such knowledge. By using several types of distractors, such as presenting syntagmatically related distractors first and unrelated distractors later, it is possible to scale the knowledge about vocabulary meaning, as Laufer et al. (2004) did. Another approach to using distractors is to include several kinds of choices in one item. Schwanenflugel et al. (1997) prepared five options in one item: a correct definition, a partial definition, two incorrect definitions, and a ‘don’t know’ option. Those who selected an option with a partial definition were regarded as having only partial knowledge, while those who selected an option with a correct definition were considered to have a more refined vocabulary knowledge. If similar tests can be created by using the three distractors this study adopted, MC tests can offer far more information than is obtained from “know” and “do not know” answers, and can become efficient diagnostic tests. Although Schwanenflugel et al. used a matching format and this study used a supplying format so the constructs of the tests in these two formats would differ, their study and the current study offer the possibility to use the information of distractors in vocabulary tests. This study is a step towards clarifying the function of distractors in vocabulary tests in context and also suggests further possibilities of MCs. Future research can replicate this study using cloze tests with more types of distractors.

Additional file

Additional file 1: The three types of test.

Competing interests
The author declares that he has no competing interests.

Received: 11 May 2013 Accepted: 9 August 2013
Published: 13 September 2013

References

Alderson, JC, Clapham, C, & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Attali, Y, & Bar-Hillel, M. (2003). Guess where: the position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement, 40*, 109–128.

Bachman, LF. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19*, 535–556.

Bond, TG, & Fox, CM. (2007). *Applying the Rasch model*. Mahwah, NJ: Lawrence Erlbaum.

Bovair, S, & Kieras, DE. (1985). A guide to propositional analysis for research on technical prose. In BK Britton & JB Black (Eds.), *Understanding expository text: a theoretical and practical handbook for analyzing explanatory text* (pp. 315–362). NJ: Lawrence Erlbaum.

Bruno, JE, & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: an information theoretic perspective. *Educational and Psychological Measurement, 55*, 959–966.

Celce-Murcia, M, Kooshian, GB, Jr, & Gosak, AJ. (1974). Goal: Good multiple-choice language-test items. *English Language Teaching, 28*, 257–262.

Cohen, J. (1988a). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, AD. (1998b). Strategies and processes in test taking and SLA. In LF Bachman & AD Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 90–111). Cambridge University Press.

Cohen, AD, & Upton, TA. (2006). *Strategies in responding to the new TOEFL reading tasks. TOEFL monograph series report No. 33*. Princeton, NJ: Educational Testing Service.

Davey, B. (1988). Factors affecting the difficulty of reading comprehension items for successful and unsuccessful readers. *Journal of Experimental Education, 56*, 67–76.

Drum, PA, Calfee, RC, & Cook, LK. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly, 16*, 486–514.

Farr, R, Pritchard, R, & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27*, 209–226.

Freedle, R, & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: implications for constructive validity. *Language Testing, 10*, 133–170.

Goodrich, HC. (1977). Distractor efficiency in foreign language testing. *TESOL Quarterly, 11*, 69–78.

Green, K, Sax, G, & Michael, WB. (1982). Validity and reliability of tests having differing number of options for students of differing levels of ability. *Educational and Psychological Measurement, 42*, 239–245.

Greidanus, T, & Nienhuis, L. (2001). Testing the quality of word knowledge in a second language by means of word associations: Types of distractors and types of associations. *Modern Language Journal, 85*, 467–477.

Haladyna, TM, & Downing, SM. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*, 37–50.

Haladyna, TM, & Downing, SM. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement, 53*, 999–1010.

Haladyna, TM, Downing, SM, & Rodriguez, MC. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309–334.

Henning, G. (1991). *A study of the effects of contextualization and familiarization on responses to the TOEFL vocabulary test items*. Princeton: TOEFL Research Report, Educational Testing Services.

Hoshino, Y. (2008). *Factors affecting performance of Japanese EFL learners in multiple-choice vocabulary tests in sentential context*. University of Tsukuba, Ibaraki, Japan: Unpublished doctoral dissertation.

JACET. (2003). *JACET list of 8000 basic words [JACET 8000]*. Tokyo: JACET.

Jenkins, JR, Matlock, B, & Slocum, TA. (1989). Two approaches to vocabulary instruction: the teaching of individual word meanings and practice in deriving word meaning from context. *Reading Research Quarterly, 24*, 215–235.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kurath, W, & Stalnaker, JM. (1936). Two German vocabulary tests. *Modern Language Journal, 21*(2), 95–102.

Laufer, B, Elder, C, Hill, K, & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language Testing, 21*, 202–226.

Linacre, M. (2006). *Winsteps [Computer software]*. Chicago, IL: Winsteps.

Marshalek, B. (1981). *Trait and process aspects of vocabulary knowledge and verbal ability (Tech. Rep. No. 15)*. CA: Stanford University, School of Education.

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition*. Amsterdam: John Benjamins.

Mori, Y. (1999). Beliefs about language learning and their relationship to the ability to integrate information from word parts and context in interpreting novel kanji words. *Modern Language Journal, 83*, 534–547.

Mori, Y. (2002). Individual differences in the integration of information from context and word parts in interpreting unknown kanji words. *Applied Psycholinguistics, 23*, 375–397.

Mori, Y. (2003). The roles of context and word morphology in learning new kanji words. *Modern Language Journal, 87*, 404–420.

Mori, Y, & Nagy, W. (1999). Integration of information from context and word elements in interpreting novel kanji compounds. *Reading Research Quarterly, 34*, 80–101.

Nagy, W, Herman, PA, & Anderson, RC. (1985). Learning words from context. *Reading Research Quarterly, 20*, 233–253.

Nation, ISP. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

National Center for University Entrance Examination. (n.d.). (2999). Retrieved from http://www.dnc.ac.jp/.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing, 6*, 199–215.

Nikolov, M. (2006). Test-taking strategies of 12- and 13-year-old Hungarian learners of EFL: why whales have migraines. *Language Learning, 56*, 1–51.

Nurweni, A, & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes, 18*, 161–175.

Paul, PV, Stallman, AC, & O'Rourke, JP. (1990). *Using three test formats to assess good and poor readers' word knowledge. Technical Report No. 509 of the Center for the Study of Reading*. University of Illinois at Urbana-Champaign.

Pike, LW. (1979). *An evaluation of alternative item formats for Testing English as a Foreign Language*. TOEFL Research Report (RR79-6).

Qian, DD. (1998). *Depth of vocabulary knowledge: Assessing its role in adults' reading comprehension in English as a second language. Unpublished doctoral dissertation*. Toronto, Ontario, Canada: University of Toronto.

Qian, DD, & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing, 21*, 28–52.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*, 355–371.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Rogers, WT, & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement, 59*, 234–247.

Schmitt, N, Schmitt, D, & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*, 55–88.

Schwanenflugel, PJ, Stahl, SA, & McFalls, EL. (1997). Partial word knowledge and vocabulary growth during reading comprehension. *Journal of Literacy Research, 29*, 531–553.

Shizuka, T, Takeuchi, O, Yashima, T, & Yoshizawa, K. (2006). A comparison of three- and four-option English tests for university entrance selection purposes in Japan. *Language Testing, 23*, 35–57.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics, 19*, 143–177.

Stalnaker, JM, & Kurath, W. (1935). A comparison of two types of foreign language vocabulary test. *Journal of Educational Psychology, 26*, 435–442.

Trevisan, MS, Sax, G, & Michael, WB. (1991). The effects of the number of options per item and student ability on test validity and reliability. *Educational and Psychological Measurement, 51*, 829–837.

Wolter, B. (2001). Comparing the L1 and L2 mental lexicon. *Studies in Second Language Acquisition, 23*, 41–69.

doi:10.1186/2229-0443-3-16
Cite this article as: Hoshino: **Relationship between types of distractor and difficulty of multiple-choice vocabulary tests in sentential context.** *Language Testing in Asia* 2013 **3**:16.