Article

LANGUAGE
TESTING

Language Testing
28(1) 105–126
© The Author(s) 2011
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0265532210373605
http://ltj.sagepub.com
SAGE

The Word Associates Format: Validation evidence

Norbert Schmitt and Janice Wun Ching Ng
University of Nottingham, UK

John Garras
Matsuyama Shinonome College, Japan

Abstract

Although the Word Associates Format (WAF) is becoming more frequently used as a depth-of-knowledge measure, relatively little validation has been carried out on it. This report of two validation studies tackles various important WAF issues yet to be satisfactorily resolved.

Study 1 conducted introspective interviews regarding students' WAF test-taking behavior along with interviews on featured target words to determine how accurately the most common scoring system for the WAF reflects the examinees' actual knowledge of the words. Analysis is provided concerning WAF accuracy and item answering strategies and patterns.

Study 2 repeated the interview procedures from Study 1 with several modifications, including the addition of a receptive dimension in the word knowledge interview. The various WAF-scoring methods were compared, and the format types (6 and 8 option), distractor types, and distribution of answers were examined in depth.

Both studies indicate that the WAF reflects true lexical knowledge fairly well at the extremes of the scoring scale while scores in the middle do not lead to any reliable interpretation. Furthermore, there is the likelihood that the WAF may both underestimate and overestimate vocabulary knowledge. Suggestions regarding item construction and use of the WAF are given to improve its accuracy and reliability.

Keywords

depth of knowledge, item construction, validation, vocabulary testing, Word Associates Format, word associations

As part of the growing focus on vocabulary issues in applied linguistics and language pedagogy (e.g. Ellis, 2009), the field has developed a more sophisticated understanding of the nature of vocabulary knowledge. This includes the realization that vocabulary

Corresponding author:
Norbert Schmitt, School of English Studies, University of Nottingham, Nottingham NG7 2RD, UK
Email: norbert.schmitt@nottingham.ac.uk

learning is an incremental process (Paribakht and Wesche, 1997; Schmitt, 2000), and that it consists of a number of knowledge dimensions, including meaning, form, collocation, register, and several others (Nation, 2001). Based on this enhanced understanding, vocabulary knowledge can be conceptualized not only as the number of lexical items known (vocabulary *size*), but also as how well these items are mastered (vocabulary *depth* or *quality*). Depth of knowledge is important, because it dictates the manner in which lexical items can be utilized, e.g. receptively or productively, and how appropriately. This makes measurement of vocabulary depth a useful endeavor. However, while measuring vocabulary size is commonplace, ranging from classroom quizzes to more established assessment instruments (e.g. the *Peabody Picture Vocabulary Test* (Dunn and Dunn, 2009) for L1 children and the *Vocabulary Levels Test* (Schmitt, Schmitt, and Clapham, 2001) for L2 learners), the measurement of vocabulary depth is more problematic. There are two main methods of measuring depth, described by Read (2000) as *developmental* and *dimensions* approaches. The developmental approach uses scales to chart developing mastery of a lexical item (e.g. 0 = no knowledge to 5 = full mastery). The best known one is the *Vocabulary Knowledge Scale* (Paribakht and Wesche, 1997), but as with all scales of vocabulary knowledge, it has important limitations (see Read, 2000 and Schmitt, 2010 for detailed discussions).

The best-known test format based on the dimensions approach is the *Word Associates Format* (*WAF*), first developed by Read (1993, 1995). Unlike conventional vocabulary assessments (such as matching tests), the WAF has the potential to economically measure learners' familiarity with target words' meanings and some of their uses as well, which has led to its popularity as a depth of knowledge measure amongst researchers (e.g. Greidanus and Nienhus, 2001; Qian, 2002; Schoonen and Verhallen, 2008). Their studies have suggested that the WAF has utility for research, for pedagogical applications, and perhaps even for standardized testing, with Qian and Schedl (2004) proposing using WAF items on the TOEFL examination.

Although the WAF is becoming more frequently used as a depth measure, it is probably fair to say that there has been relatively little validation carried out on it, and this has shown both strengths and weaknesses. For example, Read (1993, 1998) identified the WAF's susceptibility to guessing as one of the main threats to its validity. This makes the interpretation of scores problematic given that scores may not truly reflect students' actual vocabulary knowledge. On the other hand, Schoonen and Verhallen (2008) found that the test appeared valid for use with their young learners (9–12 years old) on the basis of IRT evidence, with reliability of .75–.83, and concurrent correlations with a definition test of .82. They also found that their WAF could distinguish between students with previously-known differences in level of more advanced word knowledge.

However, given the relative shortage of direct validation work, it is not surprising that there are still a number of important WAF issues which have either not been yet resolved, or even addressed at all. These include:

- What lexical knowledge does the WAF illustrate?
- What is the best way to score the WAF?
- What is the best way to interpret various WAF scores, especially 'split' scores?
- What strategies do examinees use when taking a WAF test?
- What effect does guessing behavior have on the WAF?

- What effects do distractor type and distribution of associates have on the WAF?
- What is the best WAF format: 6-option or 8-option?

It would not be good for a test format like the WAF to become widely accepted and used without a rigorous investigation of its properties. This report of two validation studies of the WAF will begin to fill this gap, by investigating the characteristics and interpretations of the WAF in various forms, with an emphasis on the issues highlighted above.

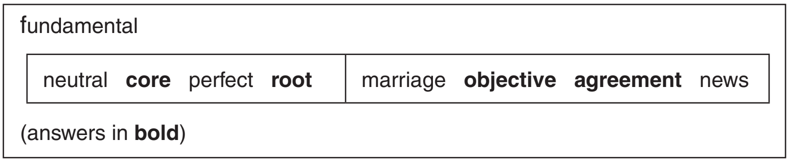
Study 1

The first study focuses on the relationship of WAF scores and examinees' true knowledge about the target words. It will also examine how the various WAF scores should be interpreted.

Methodology

Participants. The participants included 18 Japanese adults studying English for academic purposes, living in Kobe, Japan. The subjects included university undergraduate and graduate students along with several learners on a TOEFL study course. All participants expressed a desire to acquire academic vocabulary either to prepare for study at an English-medium university or to pursue other advanced study purposes. They had demonstrated mastery (90%) of the 2000 level vocabulary on the Vocabulary Levels Test, and mastery or near-mastery of the 3000 and Academic levels.

Instrument. The test was based on Read's 1998 version of the WAF where test takers are required to pick four associates related to a target word out of eight options in total. The test measured 50 adjectives taken from the Academic Word List (Coxhead, 2000). In the example item (Figure 1), the target word is *fundamental*. The associates are divided into two boxes: the box to the left contains words reflecting the meaning sense(s) of the target word (*core* and *root*) while the box on the right consists of common collocates (*fundamental objective*, *fundamental agreement*). To make the WAF less susceptible to guessing, Read (1998) varied the distribution of associates such that there can be one associate on the left and three to the right (1–3), three on the left and one to the right (3–1) or two in each box (2–2) as shown in the example.



fundamental

neutral **core** perfect **root** | marriage **objective** **agreement** news

(answers in **bold**)

Figure 1. WAF item

Procedure. Participants took a paper-and-pencil WAF test of 40 items (see Appendix 1 for more examples). Each participant was then personally interviewed. In the first stage of the interview, they were given another 10 items and asked to introspect about their thought processes and strategies as they answered them (the same 10 items were used for all participants). From this introspection, strategies were identified using an approach modeled by Johnstone, Bottsford-Miller, and Thompson (2006), and then categorized by an approach inspired by Paul, Stallman, and O'Rourke (1990). Answering behavior was codified into six strategies and a single dominant strategy was identified as the answering behavior of a student on a given item. Based on the think-aloud protocol, we identified the following WAF-taking strategies:

- *Strategy 1:* Student appears to know the target word and all or most associates.
- *Strategy 2:* Student appears to be making inferences from partial semantic knowledge of the target word and associates.
- *Strategy 3:* Student appears to be making inferences from word parts of the target word and paradigmatic associates.
- *Strategy 4:* Student appears to be making inferences based on very loose or incorrect semantic associations with target word.
- *Strategy 5:* Student appears to be making inferences among associates and distractors only – finding relationships or eliminating answers.
- *Strategy 6:* Strategy unclear – appears to be unsystematic guessing.

The main interest was to determine whether or not the target word featured prominently in the decision loop, and whether answering was based on semantic knowledge or not. Correct answers seemingly derived from assured meaning knowledge of the target words, either strong or partial, were distinguished from points gained through other types of inferencing and guessing; the former were considered to be in line with the intentions of the WAF and the latter were judged to deviate from the depth construct.

The second stage involved comparing participants' WAF answers against their actual knowledge of the target words on the test. Ten items were chosen from the original 40 items according to the ways students had marked answers, so the selection varied from subject to subject. Items earning the full range of point values were included (0–4), but the focus was placed on those with split scores – where participants got 1, 2, or 3 points for the WAF item. To correspond as closely as possible to the item knowledge and point scheme on the paper test, the interview test also probed both meaning and collocation knowledge and rated answers on each item from 0 to a maximum score of 4 points. The subjects were shown each word, given the proper pronunciation, and then asked a sequence of questions (see Appendix 2). They first had the opportunity to convey meanings in L1 (Japanese), and additional points were awarded for appropriate English synonyms, meaningful collocations, definitions, or sentences indicating meaningful use. The minimum measure for any awarding of points was some acceptable expression of the core meaning(s) of the word. Subjects were only awarded maximum points (4) if they also offered at least one acceptable collocation and, where applicable, more than one meaning of a polysemous word.

Table 1. WAF item scores vs. interview scores

WAF item score	Interview score					Total number of words tested
	4	3	2	1	0	
4	15	9	3	2	4	33
3	26	22	7	2	13	70
2	4	12	5	12	23	56
1	0	1	2	2	14	19
0	0	0	0	0	2	2
Total	45	44	17	18	56	180

Results and discussion

WAF scores vs. interview scores. Comparison of the WAF scores (One-Point method; see Study 2 for details) with the interview scores (which should represent as good a measure of the participants' true lexical knowledge as is possible) produced mixed results (Table 1). Examinees who selected all correct WAF options (4) generally did have good knowledge of the target word as indicated by a 3 or 4 on the interview score (73% of the cases). Likewise, poor WAF scores (0 or 1) generally coincided with similarly low (0 or 1) scores on the interview (86%). Thus, the WAF seems to reflect true lexical knowledge fairly well at the extremes of the scale (0 or 4). However, the 126 'split score' cases on the WAF (2 or 3) only corresponded with partial interview knowledge (2 or 3) in 37% of the cases. Rather, they often related to full knowledge judgments (4) on the interview (24%), or conversely, judgments of no (0) knowledge (29%). Clearly, this suggests that interpreting split scores on the WAF is problematic, as there is no clear indication from these results as to the true underlying lexical knowledge. Furthermore, the high extreme of the scale is not entirely safe. In 12% of the cases where the WAF score was 4, the students were judged as having no knowledge of the target word in the interview (0).

We can also look at the data by determining the relative accuracy of the WAF in indicating underlying knowledge (as indicated by the interview scores). The WAF scores and the interview scores were found to be equivalent 26% of the time, and this can be considered an 'accurate' estimation of knowledge. However, in 49% of the cases, the WAF scores exceeded the interview scores, thereby overestimating the actual knowledge of the target words. Nonetheless, there were also cases where the subjects obtained higher scores in the interview than they had on the WAF (25%), suggesting that the WAF can sometimes underestimate word knowledge as well. Overall, these results suggest that the WAF has a tendency to overestimate the vocabulary knowledge of examinees. This is also shown by the fact that of 56 cases where the interview led to a judgment of no knowledge (0), in 40 (71%) the students were able to obtain two or more points on the WAF.

Strategic behavior of examinees taking the WAF. In terms of test-taking behavior, it is most interesting to draw a distinction between answers that were semantically based in a

Table 2. Strategic behavior of answering WAF test items[a]

WAF core (frequency)	Answering behavior related to the meaning of the target word (Strategies 1 & 2) Frequency (%)	Answering behavior apparently unrelated to the meaning of the target word (Strategies 3–6) Frequency (%)
4 (29)	29 (100)	0 (0)
3 (85)	70 (82)	15 (18)
2 (57)	18 (32)	39 (68)
1 (8)	1 (13)	7 (87)
0 (1)	–	1 (100)
Total: 180	118 (66)	62 (34)

[a] 18 students × 10 items = 180 answering strategies

whole-word sense (Strategies 1 and 2) and ones that appeared to be coming from an analysis of word parts (Strategy 3) or some form of test-wiseness or guessing (Strategies 4–6). As such, Strategies 1 and 2 can be said to represent choices made by participants due to their possessing of what might be called positive knowledge of the target word, either strong or partial (See detailed description of strategies on page 4).

An analysis of Table 2 shows that there is a clear relationship between the way examinees approach the WAF items and the resulting scores. In the vast majority of cases where Strategies 1 or 2 were used, this resulted in a score of 3 or 4 on the WAF item. On the other hand, it was relatively rare that these strategies resulted in WAF scores of 2 or less. If we examine the cases where Strategies 3–6 were used, we find the converse. These strategies seldom led to WAF scores of 3 or 4, but usually resulted in scores of only 1 or 2. It is noteworthy that when these strategies were used, they produced a WAF score of (2) about 63% of the time (39/62), which should ring warning bells about how much knowledge such split scores really show.

Overall, the introspection results suggest that WAF scores of 3–4 probably reflect semantic knowledge about the target items, but scores of 0–2 do not.

Problematic answering patterns. Since the introspective exercise and interview only covered 20 of the words on the WAF, all 50 items on each test paper were then examined to see what examinees' answering patterns on split items might indicate regarding knowledge of the target words. As in the interviews, attempting to find reasonable evidence of meaning knowledge was the priority. The process uncovered two particularly problematic answering patterns:

> *Cancelling meanings:* In these items, the test taker chooses at least one correct meaning associate and at least one incorrect meaning distractor, so anyone marking the test would have no way of knowing if the meaning of the target is actually known because the meaning-related choices contradict each other. This pattern occurs in all split items, but most were found in 2-point items (45%), followed by 1-point items (29%) and 3-point ones (27%).

> *No meaning:* In a number of items, students were unable to identify any meaning associates at all but still earned points by selecting correct collocates. Given that it is unlikely that students would know the meanings of many of these collocations without knowing the meanings of the adjectives that help form them, it seems reasonable to conclude that these target words are not known. Once again, 1 and 2 point items comprise the majority of the items found to be missing an indication of meaning knowledge at 66% and 22% respectively. Some 3-point items were also found to have this problem, but comparatively few (4%).

Awarding points on WAF items with little or no evidence of meaning knowledge is questionable because knowing the form–meaning link seems to be the most basic type of word knowledge (Schmitt, 2008), and so presumably represents the initial level of depth of knowledge. It is worrying that WAF points can be gained without this essential meaning knowledge being in place. Combining the data from the two answering pattern categories (cancelling meanings/no meaning) reveals that examinees' indication of meaning knowledge is either ambiguous or completely non-existent in 31% of all 3-point items, 67% of all 2-point items, and 95% of all 1-point items. The presence of so many problematic items in the scoring would appear to be at direct odds with the WAF's purpose of being a depth test.

Study 2

Study 2 will investigate WAF-interview correspondences and examinee test-taking strategies as in Study 1. In addition, it will examine a number of other WAF issues. Different WAF scoring methods will be compared. Since there are currently two versions of the test format in use – 8-option and 6-option – the study also examines the two versions. Along with the number of options given, other facets involved in the construction of WAF items, i.e. distractor type and distribution of associates, are evaluated to determine their impact on the test-taker's behavior.

Methodology

Participants. The sample consisted of 28 international students who were studying at the University of Nottingham. Fifteen were Chinese speakers, and the rest were spread across eight L1s. The sample included students taking a pre-sessional university preparatory course and a mixture of postgraduates and undergraduates. The participants had studied English for at least 8 years, and were either in or preparing for university-level studies, and so were relatively advanced L2 learners. Participants were given the option of speaking in their L1 or English. However, as the researcher was only familiar with the Chinese language, this option was available only to the Chinese participants. Nonetheless, the participants were competent enough to be admitted to a good English-medium university, and the researcher found that they could communicate effectively with her regardless of the interview language used.

Instrument. The WAF constructed for this study was based on Read's (1998) revised format. Academic vocabulary was chosen for the targets due to their relevance to participants. The difficulty of the words was varied so that participants' behavior could be observed across a variety of situations (i.e. when the target word is known, partially known or not known at all). Piloting with participants similar to those in Study 2 narrowed the 84 potential target adjectives down to 34. These were divided into two sections, each containing 10 items: Section A used the 6-option format and Section B the 8-option one.

Other factors were also considered, such as word meaning and number of paradigmatic and/or syntagmatic associates, so as to ensure that the words used in Section A and B would be essentially comparable.

The distractor type and distribution of associates (e.g., 3-1, 2-2, 1-3) were also manipulated during the construction of the test in order to analyze their effectiveness. The three distractor types investigated included:

- *No relationship:* Distractors do not have any relationship with the target word and/or associates.
- *Form:* Distractors share orthographic similarities with the target word and/or associates. For example, distractors such as *special* and *obvious* were used due to their formal similarity with the target word – *spurious*.
- *Meaning:* Distractors are somewhat related to the target word and/or the associates. For example, the distractors *bland* and *boring* were used due to their sharing some semantic links with the target word *literal*. When possible, distractors that could potentially pair up with one another and/or associates (e.g. *boring idea, boring interpretation*) while still making sense were used.

The draft WAF was piloted on 10 native speakers and necessary revisions were made. Samples of the final instrument are given in Appendix 3.

Procedure. The procedure followed the same basic outline as Study 1. Prior to testing, WAF items and answering requirements were explained in detail and several practice items were provided. After completing the paper test, the students were given an additional 14 WAF items (seven in the 6-option format and seven in the 8-option one). These items were chosen so that they contained a mixture of distractor types. After participants finished each of these, they were asked to share their thought-processes regarding how they arrived at their answers. Here, the retrospection method where participants related their thoughts within 10 seconds or less (Cohen, 1984) was preferred to think-aloud introspection. This was due to the former being less 'reactive' (Dörnyei, 2007, p. 149) compared to the latter as the procedure does not influence the thought processes targeted by the researcher.

After the retrospective interview, participants were interviewed on their knowledge of all 20 target words tested in the WAF. While this is similar to Read (1998) and Study 1, several modifications were made to the interview procedure and scoring so as to ensure a closer fit with the vocabulary knowledge tested by the WAF. Read's interviews had probed only paradigmatic (meaning) knowledge of the words tested, so Study 1 added questions about collocations in order to cover both areas of word knowledge tested on the WAF. However, these previous interviews elicited only productive knowledge. In view of Read's (1998) observation that the productive/receptive distinction between the interview and WAF could play a part in the varying performances of learners, a receptive dimension was incorporated this time.

Participants were first shown the target word and asked whether they could provide its meaning(s). They were allowed to give synonyms, definitions, examples, sentences, and so on, as in Study 1. If they were unable to provide the meaning(s), they were then shown a sheet of paper with six words including synonyms of the target word (different synonyms from those on the WAF were used when possible) and distractors. If correct synonyms were identified, subjects were required to justify the choices to establish that they were not merely guessing. The same procedure was followed for the syntagmatic associates. Only when productive knowledge was exhausted were the additional six words shown. Likewise, participants were required to verify their receptive knowledge if they chose correctly. Participants were only given credit if they could demonstrate productive and/or receptive knowledge of the exact meanings and collocations that were being assessed in the WAF in order to ensure that the criterion measure of the interview would cohere as closely as possible with the test. Three degrees of knowledge were distinguished:

- *No knowledge:* The participant does not know the target word as evident from his/her not knowing any of the associates in the interview.
- *Partial knowledge:* The participant has some knowledge of the target word. S/he demonstrates knowledge of at least one associate during the interview.
- *Full knowledge:* The participant possesses full knowledge of the target word and is able to demonstrate knowledge of all associates during the interview.

Results and discussion

WAF scores vs. interview scores. The results generally mirror those of Study 1 (Tables 3 and 4). Examinees who selected all correct WAF options (four in the 8-option and three in the 6-option tests) typically had full knowledge of the target words, at least in 70% or more of the cases. Likewise, in cases where the WAF score was 0 or 1, the examinees usually had little knowledge of the words. In the 8-option test, in 79% of the cases, examinees had no knowledge as judged in the interview, and only partial knowledge in 21%. In the 6-option version, the figures are 82% no knowledge, 14% partial knowledge, but also with a handful of cases where students demonstrated full knowledge (4%). This is additional evidence that the WAF seems to work fairly well at the extremes of the scoring scale.

But what about the problematic split scores? Previous research (including Study 1) has revealed the WAF's vulnerability to guessing, but the severity of this problem has yet to be fully examined. In the 8-option test, split scores of 2 or 3 correct lead to the following breakdown of interview scores: 43% partial knowledge, 40% no knowledge, and 17% full knowledge. In the 6-option test, the breakdown for a WAF score of 2 is 35% partial knowledge, 41% no knowledge, and 24% full knowledge. Just as in Study 1, these results show that split scores do not lead to any reliable interpretation.

Table 3. WAF item scores vs. interview scores (8-option)

WAF item score	Interview score			*Total number of words tested*
	Full knowledge	Partial knowledge	No knowledge	
4	139	29	30	198
3	37	78	53	168
2	6	32	48	86
1	–	5	16	21
0	–	–	3	3
Total	182	144	150	476

Table 4. WAF item scores vs. interview scores (6-option)

WAF item score	Interview score			*Total number of words tested*
	Full knowledge	Partial knowledge	No knowledge	
3	195	22	41	258
2	35	52	60	147
1	3	10	45	58
0	–	–	13	13
Total	233	84	159	476

We can also do an 'accuracy' analysis similar to Study 1. For this, we can count WAF scores of 4 as full knowledge, 2 and 3 as partial knowledge, and 0 and 1 as no knowledge on the 8-option test version. For the 6-option version, it will be 3 = full knowledge, 2 = partial, and 0 and 1 = no knowledge. Using these conventions, the WAF and interview scores matched in 56% of the cases, the WAF scores exceeded the interview scores in 34% of the cases (overestimation), and were less than the interview scores in 10% (underestimation) on the 8-option test. For the 6-option version, it was 64% matching, 26% overestimation, and 10% underestimation. Compared to Study 1, there is a higher percentage of matching between the WAF and interview scores, and this is probably due to the wider interview score bands in Study 2. However, given the closer correspondence of the interview elicitation with the WAF item format (receptive knowledge also accepted) the Study 2 results are probably a better indication of the relationship between WAF scores and examinee knowledge. This means that between one-half to two-thirds of the time (depending on which WAF version is used), the WAF provides a good measure of knowledge. But the Study 2 results also confirm that the WAF overestimates examinee knowledge between about one-quarter (6-option) to one-third (8-option) of the time. There is also about a 10% chance of underestimation on both test versions.

The tendency of the WAF towards overestimation is worrying and is related to the WAF's vulnerability to guessing. For instance, it raises the question of whether test-takers are only able to guess successfully if they have partial knowledge of the word or

Table 5. Strategies used on the WAF when examinees had no or partial knowledge of the target words

	Strategies	No knowledge	Partial knowledge
A	Student appears to be making inferences from partial semantic knowledge of the target word and associates	–	1
B	Student appears to be making inferences amongst associates and distractors without knowing the stimulus word	1	2
C	Student appears to be making inferences from the orthographic form of the target word and the options given	2	5
D	Students appear to be inferring the likely distribution of associates for each item (e.g., 3-1, 2-2, 1-3)	3	4
E	Students appear to be making inferences based on very loose or incorrect semantic associations	4	3
F	Unsystematic guessing	5	6

1 = most frequent, 6 = least frequent

whether they are successful even with no knowledge of the word at all. These results indicate a high possibility that participants are able to get 2 or more associates despite having no knowledge of the target word (87% 8-option; 64% 6-option). Even the lower-proficiency learners in Study 1 were able to guess 2 or more associates with no knowledge of the target words 72% of the time. Conversely, full knowledge nearly always leads to high WAF scores. Thus the WAF appears to be better at indicating what learners do know rather than what they do not know.

Strategic behavior of examinees taking the WAF. The above analyses suggest that overestimation of participants' knowledge due to successful guessing is a pervasive problem for the WAF. In order to explore this, we need to take a closer look at the test-taking behavior of the examinees, particularly the strategies used by participants when they had no or only partial knowledge of the target words. These are illustrated in Table 5, which shows the various strategies employed by participants during the retrospective interview, and ranked according to their frequency of use. (The identification of these strategies was based in part on Study 1's findings.)

The success and/or failure of a strategy were very much dependent on its user, and there were cases where participants used more than one strategy and/or varied their choice of strateg(ies) depending on the WAF item concerned. However, on the whole, a meaning-based approach was preferred by the majority of participants. Strategy A was the most frequently used strategy by participants with partial knowledge of the target words. It reflects positive (though partial) knowledge of participants, and represents the type of test-taking behavior we would like to see. Essentially the same meaning-based strategy (B) was also most frequent for students with no knowledge of the target words. Read (1993) also observed that his participants used the semantic connections between associates to infer the answers. We see this in Extract 1, where WenJing (all names are pseudonyms) makes the connection between *claims* and *grounds*, and then *false*.

Extract 1. WAF test-taking strategies

Spurious

Special Obvious **False** Spurned	**Claims** Spur **Argument** **Grounds**

(WenJing's answers in **bold**)

Interviewer: How did you know get all the options correct without knowing the meaning of *spurious*?

WenJing: … guessing … as in I am quite sure that *spurned* and *spur* is out …

Interviewer: Why are these options out?

WenJing: Don't know eh… Usually these words you put there to confuse us one … so I look across the eight words, I would think that *spur* and *spurned* are trick-words … because it is not this two words … I am left to select among the other six … and *claims* and *grounds* are related ma … since *obvious*, *special* and *spurned* have different meanings … I know that there the left box only has one correct option … *false* fits the best with *claims* and *grounds* … then since I have to choose one more option from the right box, so *argument*.

Strategy C occurs when participants base their inferences on the orthographic form of the target word and options. From participants' verbal reports, we found this strategy can be realized in three ways. First, participants may use familiar word parts (i.e. affixes) and/or similarities of the target word(s) with other words they know to guide their guesses. Second, participants can use formal similarities between the target word and options given. For example, one participant (erroneously) chose the collocation *liability* based on its rhyming with the target word *reliable*. Third, participants can sometimes eliminate formal distractors as 'trick words', and it is likely that Wenjing spotted *spur* and *spurned* as distractors because of their formal similarity. Like WenJing, most participants were able to eliminate *spur* and *spurned* from their choices; however, it should be noted that some participants chose these distractors because of their formal similarities with *spurious*.

Having gone through some WAF-items prior to the test, WenJing's awareness that there were a fixed number of associates, with at least one associate occurring in both the left and right boxes, led to her inference of the likely distribution of associates in the item (Strategy D). Finally, Strategy E refers to students making incorrect semantic associations, which often leads to unsuccessful guesses and reveals their lack of knowledge regarding the target word (Extract 2).

Extract 2. Incorrect semantic associations

Potential

Possible **Good** **Rewarding**	Customers **Benefit** Pride

(Aziz's answer in **bold**)

Aziz: I think *potential* has a good meaning so *good*… and *rewarding* is a good thing…

So what does examinee strategic behavior mean for the WAF? Their use of Strategy A is reflective of their partial knowledge, and so it is appropriate that they be rewarded for this. Given that Strategy E often results in incorrect guesses, it probably does not pose a serious threat. Strategy B depends on each examinee's idiosyncratic knowledge and so is difficult to predict or control for. However, items should ideally be written so as to prevent test-takers from being able to make easy inferences amongst associates and distractors without knowing the target word. There is also a need to consider Strategies C and D when creating WAF items. Strategy C poses an interesting dilemma: orthographicallysimilar distracters can be successful as evident from their attractiveness to some participants, yet they can also be ineffective given that they are easily identified by test-wise participants as 'trick-options'. Given that the majority of participants took a meaning-based approach to the WAF and used formal similarity as a means to discount distractors, it is our opinion that formal distracters should be less frequently used. Strategy D can be pre-empted to a large degree by purposefully varying the distribution of associates so as to prevent guessing.

Underestimation is less of a problem for the WAF, but given that even 'native speakers sometimes have difficulty finding a third or fourth associate for a particular item' (Read, 2000, p. 186), its occurrence is not really surprising. Two reasons emerged from the verbal reports for underestimation. First, participants may be too fixated on a particular meaning sense of a word and neglect other associates. This situation usually occurs when both a paradigmatic and syntagmatic associate taps into a similar meaning sense, as the item in Extract 3.

Extract 3. Fixation on a single meaning sense

Classic

Typical Predictable **Elegant** **Timeless**	Conventions Laws Norm **Example**

(Answers in **bold**)

Mei Ling preferred to choose options that can be rationalized to fit the 'typical' meaning sense of *classic*. This is despite her knowing of the other meanings of *classic*:

Mei Ling's answers: *Typical, Predictable, Timeless, Example*
Interviewer: Why did you choose *predictable*?
Mei Ling: Why *predictable*? … well … I am thinking in terms of science…sometimes you do physics they tell you it is a classic example … means that you can predict … cause it is always going to go like that …
Interviewer: Would you consider *elegant*?
Mei Ling: *Elegant* … oh … it's like the lady is very classy or classic design is it? Oh, yeah it is possible … but wrong context altogether eh … think I was too concentrated on one context then forgot … never think of other possibilities … **I think if you put eight words then people will try to link the correct answers** … that is why I chose *predictable* over *elegant* … maybe if you put design in the right box then I will choose *elegant* …

Here, underestimation was due to Mei Ling attempting to fit her other choices with the meaning sense of ‘typical’ which was similarly signaled by the syntagmatic associate of *example.* Given that a ‘classic example’ is one which is commonly used to exemplify a concept in physics and can be used to ‘predict’ whether other ‘typical’ or similar phenomenon fall into the same category, there seems to have been a misapprehension that answers should form a unified whole or share linkages with one another (see **bold**).

The second reason for underestimation is due to participants not being able to discriminate the best answers from possible options. This issue was especially salient when participants had to select from the various syntagmatic options which they deemed to be equally valid. This is shown in Extract 4.

Extract 4. Inability to differentiate collocations

Capable

Able Creative **Skilled** Inspiring	**Hands** Brains Thoughts **People**

Huifang: Actually to me, *capable hands*, *capable brains* and *capable people* are possible … but since I chose two options from the left box … I have to eliminate one …

Interviewer: What made you choose *capable brains*?

Huifang: I don’t know … actually I heard people say it before …

Here, despite knowing what the collocations *capable hands* and *capable people* meant, there were a considerable number of participants who still chose *brains* or *thoughts* as their answers. This coincides well with various findings that nonnatives find collocations difficult, especially less frequent ones (Durrant and Schmitt, 2009; Granger and Meunier, 2008). In this sense, the WAF seems to be working well as a depth test, as it is tapping into learners’ uncertainty about collocational combinations.

Scoring the WAF. Most researchers (e.g. Nassaji, 2006; Qian, 1999, 2000; Read, 1993, 1998; Suteja, 2009) have chosen to award one point per correct associate selected (hereafter, the One-Point scoring method). Using the WAF item in Figure 1 as an illustration, the minimum score would be 0 and the maximum 4, with interpretations of ‘No knowledge’ and ‘Full knowledge’ respectively. However, we have seen that interpreting ‘split scores’ (1, 2, or 3 correct) is problematic, as it is very possible to get some of the associates simply through guessing. To counteract guessing on the test, a few researchers have used alternative scoring systems:

- *Correct-Wrong:* Test-takers gain a point for selecting each associate and avoiding each distractor. The maximum and minimum score for a WAF item with four associates is 8 and 0 respectively (Greidanus et al., 2004; Greidanus et al., 2005).
- *All-or-Nothing:* The examinee gets a point per WAF item only if s/he is able to select all of the associates. The maximum and minimum score for an item would be 1 and 0 respectively (Schoonen and Verhallen, 2008).

Table 6. Correlation of WAF and interview scores according to scoring method

Scoring method	r (r^2)	
	6-option format	8-option format
All-or-Nothing	.884* (78.1)	.871* (75.9)
Correct-Wrong	.877* (76.9)	.855* (73.1)
One-Point	.871* (75.9)	.885* (78.3)

*Pearson, $p < .01$ (2-tailed), r^2 value is in %

Table 7. ANOVA analysis according to scoring method

Scoring method	ANOVA[a] effect size (eta^2)	
	6-option format	8-option format
All-or-Nothing	.622	.622
Correct-Wrong	.586	.637
One-Point	.572	.638

[a] All comparisons: F(2, 28), $p < .001$, LSD post-hoc tests show Lower < Middle < Upper

The interview scores give us the means to empirically compare these scoring methods. We first correlated the interview scores with the WAF scores as calculated by the three scoring methods (Table 6). While the All-or-Nothing and One-Point methods have the highest correlations for the 6- and 8-option versions respectively, the various methods are very comparable.

Following the example of previous studies (Greidanus and Nienhus, 2001; Greidanus et al., 2004, Greidanus, Beks, and Wakely, 2005), the validity of the WAF scoring methods was also evaluated in terms of their capability to discriminate between different proficiencies of test-takers. The participants were divided into three groups based on their total interview scores: lower (bottom quartile), middle (middle two quartiles), and upper (top quartile). They were then submitted to an ANOVA and post-hoc LSD analyses. All scoring methods reliably distinguished between the three groups, but the effect sizes differ somewhat (Table 7). Again, All-or-Nothing comes out best for the 6-option version and One-Point for the 8-option version.

Overall, it seems that the various methods produce similar results. The Correct-Wrong method can probably be discounted, as it is more complicated and is little different from the One-Point method. Both essentially yield the same percentage results if the test-takers choose (or guess) the required number of associates per item, and a percentage difference only occurs if they choose more or fewer options than required. This leaves All-or-Nothing and One-Point as the preferred methods, with the former being marginally better for the 6-option format, and the latter for the 8-option format. It is also interesting to note that the ANOVA results strengthen previous research (Greidanus, Beks, and Wakely, 2005; Schoonen and Verhallen, 2008) verifying the

Table 8. Correlating interview and test scores for WAF items composed of different distractor types

Distractor type	r (r^2)		
	6-option format	8-option format	Combined formats
No relationship	.776* (60.2)	.912* (83.2)	.897* (80.4)
Meaning	.910* (82.8)	.813* (66.1)	.899* (80.8)
Form	.637* (40.6)	.661* (43.7)	.693* (48.0)

*Pearson $p < .01$ level (2-tailed), r^2 value is in %

WAF's ability to distinguish between learners with different language proficiencies/ levels of vocabulary knowledge.

WAF item construction: What kind of distractors are the most effective? WAF items should be written so as to minimize the likelihood of successful guessing especially when learners have no knowledge of the target words. To determine how this can best be done, we examine two facets of WAF item construction: type of distractor and distribution of associates within the item.

Previously, Greidanus and Nienhus (2001) explored distractor type (semantically related vs. semantically nonrelated) and association type (paradigmatic, syntagmatic, analytic), and found that semantically related distractors worked better for their advanced learners, who also showed a preference for paradigmatic responses. The items in this study were written to investigate three distractor types: No Relationship, Meaning, and Form. In order to evaluate the influence of these respective distractors on the WAF, the scores of the WAF items containing the respective distractor types and their corresponding interview scores were correlated. Table 8 shows the results for both 6- and 8-option test formats.

From the correlation results, it appears that items with Form distractors are the least robust in measuring participants' knowledge of the target words. On the other hand, while No-relationship and Meaning distractors yield relatively similar correlations overall, it seems that Meaning and No-relationship distractors work better for the 6- and 8-option formats respectively. These findings are reinforced by an ANOVA and post-hoc LSD analysis we carried out, which was similar to that in the scoring method section (Table 9). Items with Form distractors did not distinguish the three proficiency groups at

Table 9. ANOVA analysis according to scoring method

Distractor type	6-option (eta^2)	8-option (eta^2)
No relationship	L < M = U** (.539)	L < M < U** (.696)
Meaning	L < M < U** (.591)	L < M < U** (.513)
Form	Not significant	L = M < U* (.342)

*F(2, 28), $p < .05$
**F(2, 28), $p < .001$
All LSD post-hoc tests $p < .05$
L = Lower proficiency group, M = Middle group, U = Upper group

Table 10. Correlation between test and interview scores for items with different distributions of associates (8-option format)

Distribution of associates	r (r^2)
1–3	.736* (54.2)
2–2	.871* (75.9)
3–1	.820* (67.2)

*Pearson $p < .01$ level (2-tailed), r^2 value is in %

all on the 6-option version, and only distinguished the upper proficiency students from the rest on the 8-option version. Items with Meaning distractors reliably distinguished the three groups on both test versions. The No-relationship items only distinguished the low group from the rest on the 6-option version, but reliably distinguished all three groups (with a large effect size) on the 8-option version. These results support our conclusion in the Strategy section, where we felt that Form distractors should be minimized. They also suggest that the different WAF versions may benefit from different distractor types, with Meaning-based distractors being better for the shorter 6-option version, but with No-relationship distractors being better for the 8-option version.

WAF item construction: distribution of associates. In order to prevent guessing, Read (1998) decided to vary the distribution of associates. Given that the 6-option format only offers two variations (i.e. 1 paradigmatic and 2 syntagmatic associate or vice versa); this section will only discuss the 8-option format. The distributions of associates are as follows:

- 1–3: One paradigmatic and three syntagmatic associates
- 2–2: Two paradigmatic and two syntagmatic associates
- 3–1: Three paradigmatic and one syntagmatic associates

From the correlation results in Table 10, it appears that the 1–3 distribution is the least valid in measuring participants' knowledge, whereas the 2–2 distribution produces the best results. The low correlation result for the 1–3 distribution relative to the others is probably due to the difficulty in finding three collocations whose meanings are distinct from one another. This appeared to allow participants to use Strategy D, where they attempted to make semantic associations between the options given. It is thus likely that the relatedness of the syntagmatic associates in the 1–3 distribution makes such an item more susceptible to successful guessing. On the other hand, the lower correlation of the 3–1 distribution as compared to the 2–2 distribution is probably due to underestimation. Here, participants may be too focused on particular meaning senses such that they neglected other associates.

General Discussion

Most of the results have been discussed in the previous relevant sections. However, a few additional comments are warranted on some of the issues.

Which WAF format to use?

From the previous analysis, it appears that the 8-option format would probably be better for more advanced learners due to it being less susceptible to guessing. The relative vulnerability of the 6-option format is based on two findings. First, the All-or-Nothing scoring method is the most valid for it. Also, it yielded a lower correlation score relative to the 8-option format when the One-point method is used. Second, orthographic distractors are more salient as 'trick-options' on the 6-option format. Nevertheless, in view of Greidanus et al.'s (2004) observation that it might be challenging to always find a fixed number of associates, a 6-option format could be the better choice depending on the target word tested. For example, the 8-option format may be more suited to the testing of higher frequency words given that they are more likely polysemous (Schmitt, 2000). In terms of time, the 8-option items take only slightly longer on average (39.0 seconds) than the 6-option items (32.8) and so this should make little practical difference.

How to score the WAF?

We suggested that the All-or-Nothing method is probably best for the 6-option version, and the One-Point method for the 8-option version. But the best method may also depend on the examiner's agenda. For instance, if one requires a more accurate indication of whether participants know the target words or not, then All-or-Nothing is probably the best option given that it inflates test-takers' scores the least when they have no knowledge of the target words. However, it also has the drawback of not really rewarding participants for partial knowledge. If the WAF is used for pedagogical purposes, a less harsh scoring system (i.e. One-Point or Correct-Wrong) would probably be preferred so as not to demoralize students, and to provide positive feedback on partial knowledge.

Which distractor type to use?

The results seem to indicate that Form-based distractors should generally be avoided. No-relationship distractors are recommended, especially for the 8-option version, and here test compilers simply need to find distractors with no formal or semantic links to each other or the target word. We carried out a more comprehensive analysis of meaning distractors than we have space to report here, but the key implications for writing them as to minimize guessing are as follows:

1. Avoid distractors that are antonymic to the meanings of the target word and/or associates. For instance, the item below is to be avoided as it was found that most participants could divide the paradigmatic options into two pairs – *Safe-Inoffensive* and *Blunt-Direct*. For the shrewd guesser, s/he will be able to get all four associations if their choice of *safe* and *inoffensive* is followed by the elimination of *threat* and *risk*, as the former logically contradicts the latter.

Innoculous

Inoffensive Blunt **Safe** Direct	**Statement** Threat Risk **Question**

2. Use distractors that are more closely related to the meanings of the target word and/or associates (i.e. sharing the same positive/negative connotations, semantic prosody). For instance in the item below, participants who had a vague knowledge of *potential* having a positive connotation would choose the distractor *good*.

Potential

Possible Good Rewarding	**Customers** **Benefits** Pride

3. Use distractors that can potentially pair up with one another and/or associates while potentially still making sense. In view of participants' prevalent use of Strategy D, using distractors such as those shown in the items (i.e. *Inspiring People*, *Creative Hands, Good Customers*) could also increase the difficulty of items.

Capable

Able Creative **Skilled** Inspiring	**Hands** Brain Thoughts **People**

Limitations

The participants in these studies were all intermediate (Study 1) to relatively advanced (Study 2) L2 learners, and so the findings in this report may not apply to other proficiency levels (e.g. beginning L2 learners). The limited number and range of participants also inevitably constrain the generalizability of the studies.

Conclusion

As in most validation studies, this report has found both strengths and limitations of the WAF format. In its present realizations, it is unlikely to be robust enough to become part of high-stakes standardized tests as Qian and Schedl (2004) suggested. However, despite its limitations, it has proven useful in both vocabulary research and classroom applications. It appears to give good indications of positive vocabulary knowledge, but also suffers from a tendency to overestimate learner knowledge. In addition, split scores remain difficult to interpret. It will take further validation research to resolve these issues, but in the meantime, we leave Bogaards (2000, p. 496) to summarize the WAF's value:

> It is questionable whether these reservations disqualify this test format. If its only purpose is to measure how well the selected target items are known, then the test may not do a very good job. But one could be interested also in more general qualitative knowledge of the lexicon. In that case, it would be interesting to be able to make a difference between learners

> who are successful in identifying two or three associates even without knowing the stimulus word, and those who were not struck by any meaningful relationships between the [seven or] nine words given in each item [depending on the format]. Moreover 'resourcefulness in seeking possible associates' and 'confidence to make guesses' may be seen as negative when one wants to know whether selected relationships are recognized by the learner or not. But in a more general way, such strategies seem also to be helpful in normal language use and learners who exploit those means may be said to have richer vocabulary than those who do not.

References

Bogaards P (2000) Testing L2 vocabulary knowledge at a high level: The case of the Euralex French tests. *Applied Linguistics*, 21(4): 490–516.

Cohen A (1984) On taking tests: What the students report. *Language Testing* 1(1): 70–81.

Coxhead A (2000) A new academic word list. *TESOL Quarterly* 34(2): 213–238.

Dörnyei Z (2007) *Research methods in applied linguistics*. Oxford: Oxford University Press.

Dunn LM and Dunn LM (2009) *Peabody Picture Vocabulary Test-Third Edition (PPVT-III)*. Accessed on 30 January 2009 from http://ags.pearsonassessments.com/Group.asp?nGroupInfoID=a12010#dvd.

Durrant P and Schmitt N (2009) To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics* 47(2): 157–177.

Ellis R (2009) Editorial. *Language Teaching Research* 13(4): 333–335.

Granger S and Meunier F (eds). (2008) *Phraseology: An interdisciplinary perspective.* Amsterdam: John Benjamins.

Greidanus T, Beks B, and Wakely R (2005) Testing the development of French word knowledge by advanced Dutch and English-speaking learners and native speakers. *The Modern Language Journal*, 89(2): 221–233.

Greidanus T, Bogaards P, Van der Linden E, Neinhuis L, and De Wolf T (2004) The construction and validation of a deep Word knowledge for advanced learners of French. In P Bogaards and B Laufer (eds) *Vocabulary in a second language: Selection, acquisition and testing*. Amsterdam: John Benjamins, 191–208.

Greidanus T and Neinhuis L (2001) Testing the quality of word knowledge in a second language by means of word associations: Types of distractors and types of associations. *The Modern Language Journal* 85(4): 467–477.

Johnstone CJ, Bottsford-Miller NA, and Thompson SJ (2006) Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners (Technical Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Accessed 5 March 2009 from http://education.umn.edu/NCEO/OnlinePubs/Tech44/.

Nassaji H (2006) The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy and use. *The Modern Language Journal* 90(3): 387–401.

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

Paribakht TS and Wesche M (1997) Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J Coady and T Huckin (eds), *Second language vocabulary acquisition*. Cambridge: Cambridge University Press, 174–200.

Paul PV, Stallman AC, and 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.

Qian D (1999) Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review* 56(2): 282–308.

Qian D (2000) *Validating the role of depth of knowledge in assessing reading for basic comprehension tasks in TOEFL 2000*. Princeton, NJ: Educational Testing Service.

Qian D (2002) Investigating the relationship between vocabulary knowledge and academic reading performance. *Language Learning* 52(3): 513–536.

Qian D and Schedl M (2004) Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing* 21(1): 28–52.

Read J (1993) The development of a new measure of L2 vocabulary knowledge. *Language Testing* 10(3): 355–371.

Read J (1995) Refining the word associates format as a measure of depth of vocabulary knowledge. *New Zealand Studies in Applied Linguistics* 1: 1–17.

Read J (1998) Validating a test to measure depth of vocabulary knowledge. In A Kunnan (ed.) *Validation in language assessment*. Mahwah, NJ: Lawrence Erlbaum 41–60.

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

Schmitt N (2000) *Vocabulary in language learning*. Cambridge: Cambridge University Press.

Schmitt N (2008) Instructed second language vocabulary learning. *Language Teaching Research* 12(3): 329–363.

Schmitt N (2010) *Researching vocabulary: A vocabulary research manual.* Basingstoke: Palgrave.

Schmitt N, Schmitt D, and Clapham C (2001) Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing* 18(1): 55–88.

Schoonen R, and Verhallen M (2008) The assessment of deep word knowledge in young first and second language learners. *Language Testing* 25(2): 211–236.

Suteja H (2009) The relationship between vocabulary knowledge and academic achievement. Pelita Harapan University: Unpublished Master's thesis at Pascasarjana Universitas, Atma Jaya, Jakarta. Accessed 2 March 2009 from http://l-pis.com/pdf/Relationship.pdf.

Appendix 1. Examples of WAF Items (Study 1)

approximate

different imprecise rough high	cost age vehicle help

obvious

singular apparent visible evident	time overview mind sign

prior

general advance possible serious	experience approval knowledge appliance

Appendix 2. Word Knowledge Interview Questions (Study 1)

1) Can you give me the meaning of _____________ in Japanese? [L1 link with the word]

2) Please give me the first three English words you think of when you hear or see the word________________? [general association knowledge]

3) Can you tell me the meaning of __________________in English?
[open response: e.g. synonyms, definitions, examples or sentences were acceptable ways to answer]

4) Is there any other meaning of ______________? [polysemy, if applicable]

5) Can you think of any nouns that often follow ________________? [adjective - noun collocations]

Appendix 3. Examples of WAF Items for 6-option and 8-option Formats (Study 2)

Negative

Pessimistic Surprised Bad	Rhythm Attitude Union

Systematic

Transparent Orderly Uncertain	Way Emotions Analysis

Innoculous

Inoffensive Blunt Safe Direct	Statement Threat Risk Question

Positive

Good Optimistic Calm Convinced	Purchase Music Results Camp

Subordinate

External Inferior Dense Active	Class Clause Fiction Role