Zhang and Koda (2017) 2:1
DOI 10.1186/s40862-017-0024-0

Asian-Pacific Journal of Second and Foreign Language Education

REVIEW Open Access

Assessing L2 vocabulary depth with word associates format tests: issues, findings, and suggestions

Dongbo Zhang[1*] and Keiko Koda[2]

* Correspondence: zhangdo6@msu.edu
[1]Department of Teacher Education, Michigan State University, 620 Farm Lane, East Lansing, MI 48824, USA
Full list of author information is available at the end of the article

Abstract

Word Associates Format (WAF) tests are often used to measure second language learners' vocabulary depth with a focus on their network knowledge. Yet, there were often many variations in the specific forms of the tests and the ways they were used, which tended to have an impact on learners' response behaviors and, more importantly, the psychometric properties of the tests. This paper reviews the general practices, key issues, and research findings that pertain to WAF tests in four major areas, including the design features of WAF tests, conditions for test administration, scoring methods, and test-taker characteristics. In each area, a set of variables is identified and described with relevant research findings also presented and discussed. Around eight topics, the General Discussion section provides some suggestions and directions for the development of WAF tests and the use of them as research tools in the future. This paper is hoped to help researchers become better aware that the results generated by a WAF test may vary depending on what specific design the test has, how it is administered and scored, and who the learners are, and consequently, make better decisions in their research that involves a WAF test.

Keywords: Word Association, Word Associates Format, Vocabulary Depth, Assessment, Second Language

Introduction

Vocabulary knowledge is multi-dimensional and entails different aspects of knowledge about knowing a word (Chapelle 1994; Henriksen 1999; Milton and Fitzpatrick 2014; Nagy and Scott 2000; Nation 1990, 2001; Schmitt 2014). Among the various conceptualizations of the dimensions of vocabulary knowledge, the best-known one is perhaps the differentiation between size or breadth and depth, with the former commonly known as referring to how many words one knows and the latter how well one knows those words (Anderson and Freebody 1981; Schmitt 2014; Wesche and Paribakht 1996). Defining vocabulary size in a numeric sense appears to make it "easily" assessable with learners demonstrating knowledge of form-meaning connections for a selected set of words that represent different frequency bands, such as in the case of the Vocabulary Levels Test (Nation 1990; Schmitt et al. 2001). On the other hand, what depth means has been unclear, and the diverse conceptualizations and discussions have posed a challenge to the assessment of this dimension of knowledge (Chapelle 1994; Henriksen 1999;

Nation 2001; *Qian 2002; Read 2000, 2004; Richards 1976; Wesche and Paribakht 1996; see Schmitt 2014 for a recent review).

According to Read (2000), there were two general approaches in the literature to the assessment of second language (L2) vocabulary depth. The "developmental" approach, which reflects the incremental nature of vocabulary acquisition and is represented by the Vocabulary Knowledge Scales (Wesche and Paribakht 1996), describes word mastery as following a continuum from not knowing anything about a word to full mastery characterized by the ability to correctly use the word across contexts. The "dimensional" approach, on the other hand, contends that vocabulary depth encompasses various aspects of knowledge about words, such as form, meaning, and use in both receptive and productive senses and in both spoken and written modalities (Nation 2001; Schmitt 2014). Read (2004), for example, distinguished between three separate but related meanings of depth, including precision of meaning ("the difference between having a limited unclear idea of what a word means and having much more specific knowledge of its meaning), comprehensive word knowledge ("knowing the semantic feature of a word and its orthographic, phonological, morphological, syntactic, collocational and pragmatic characteristics"), and network knowledge ("the incorporation of the word into its related words in the schemata, and the ability to distinguish its meaning and use from related words") (pp. 211–212).

Given the vast scope, there is perhaps no need, and logistically impractical, to assess every aspect of knowledge implied in Read's (2004) three meanings of depth (Schmitt 2014). As Read (2000) argued, including more and more aspects of word knowledge to be assessed means that fewer and fewer words will become the target of assessment, which is certainly not a desirable direction to follow. Focusing on the aspect of network knowledge and based on the concept of word association, *Read (1993; 1998) pioneered in the development of what he called Word Associates Format (WAF) tests for assessing L2 vocabulary depth. As Read (2004) argued,

> "as a learner's vocabulary size increases, newly acquired words need to be accommodated within a network of already known words, and some restructuring of the network may be needed as a result…This means that depth can be understood in terms of learners' developing ability to distinguish semantically related words and, more generally, their knowledge of the various ways in which individual words are linked to each other." (p. 219)

Word association tasks have long been used in the literature to examine developing organizations of words in native-speaking (L1) children's mental lexicon (Aitchison 1994; Entwisle 1966; Nelson 1977). Typically, children are asked to provide a word that has connection with a stimulus word (i.e., free association), and their responses for a set of stimulus words are then categorized according to different types of association relationships, such as paradigmatic (i.e., an associate of the same word class as the stimulus word and performing the same grammatical function in a sentence, such as a synonym), syntagmatic (i.e., an associate of a different word class from the stimulus word and having a sequential relationship with the stimulus word, such as a collocate), and clan/phonological (i.e., an associate having sound resemblance to a stimulus word). Through comparisons of the proportions of the different types of associates in students'

responses (across time), an inference is usually made on how words are organized and how the organization develops in the students' mental lexicon.

The free association paradigm has also been commonly used in the L2 literature to probe into the mental organization of words and the development of that organization in L2 learners (e.g., Fitzpatrick 2013; *Henriksen 2008; Jiang 2002; Nissen and Henriksen 2006; Wolter 2001; Zareva 2007). From the perspective of assessing vocabulary depth, however, the free association paradigm has some notable limitations. Meara (1983; see also Meara 2009) found that L2 learners' responses through free association differed systematically from native speakers' and could be very diverse and unstable across test administrations. In addition, it is not always straightforward and easy to score learners' productions and identify possible individual differences, even though a few ways, typically using native speakers' canonical responses as a benchmark for scoring/coding to represent the "nativelikeness" of learner responses, have been proposed and validated in the literature (e.g., Fitzpatrick et al. 2015; *Henriksen 2008; Higginbotham 2010; Schmitt 1998; Vermeer 2001; Zareva 2007; Zareva and Wolter 2012).

The aforementioned concerns about free association for assessment purposes were a major reason that motivated *Read (1993; 1998) to develop the Word Associates Test (WAT) for assessing university English learners' vocabulary learning. Different from a free association task, the WAT takes a controlled, receptive format. In *Read's (1998) WAT, a target word (i.e., an adjective) is followed by eight other words, half of which are semantically associated with the target word (i.e., associates) and the other half are not (i.e., distractors). The associates have two types of relationship with a target word: paradigmatic and syntagmatic. As shown in the example below (*Read 1998, p. 46), the target word *sudden* is an adjective followed by two boxes of four words. The four words on the left are all adjectives with associates being synonymic to *sudden* (i.e., *quick* and *surprising*), and the four words on the right are all nouns with associates being collocates of *sudden* (i.e., *change* and *noise*). The other four words (e.g., *thirsty* and *school*) are semantically unrelated. Asking learners to choose the associates of *sudden* from a given set of choices, rather than the associates being elicited through a free association paradigm, thus allows researchers to have good control over the responses to be given by the learners. It also makes scoring much easier than in the case of free association, although as will be seen in the review below, how a WAT item is to be best scored is still an issue under debate.

sudden

beautiful ***quick*** ***surprising*** thirsty	***change*** doctor ***noise*** school

Ever since *Read (1993; 1998) developed the WAT, various other WAF tests following the prototype have been developed and validated in English as well as other languages, depending on the specific designing features reviewed later in this paper (see also Additional file 1: Appendix B). Those tests have also been used widely in the L2 literature where learners' level of vocabulary depth needs to be assessed, and individual differences need to be obtained. Despite WAF tests' generally strong psychometric

properties (e.g., reliability and concurrent validity) and usefulness as research tools (e.g., indexing vocabulary depth and predicting the development of language proficiency, such as reading comprehension), there are also a number of questions unanswered about them (Beglar and Nation 2014; Schmitt 2014), given the big variations in what specifically the tests are and how they are used with what learners.

It is thus the interest of this paper to review current practices, key issues, and research findings about WAF tests and their applications in L2 research, and provide some suggestions and directions for the development of WAF tests and the use of them as research tools in the future. Specifically, this review is guided by the following four questions:

1. What design features of WAF tests have received attention from L2 researchers? Are test responses or performance, and more importantly, the psychometric properties of WAF tests influenced by those design features?
2. Does the way WAF tests are administered have an influence on test responses and the psychometric properties of the tests?
3. How are responses to WAF tests scored? How, if at all, do different scoring methods have an influence on the psychometric properties of WAF tests?
4. Do test-taker characteristics have an influence on WAF tests, or do WAF tests create testing bias among different types of test-takers?

To prepare for this review, we referred to two electronic research databases (i.e., MLA International Bibliography and PsycINFO), supplemented by Google Scholar, to located research outputs using a set of keywords, such as *word association/word associates, vocabulary knowledge, depth/deep knowledge, lexical network/semantic network,* and *second language/foreign language*. In addition, the reference lists of some existing publications, particularly recent reviews that involved vocabulary depth, such as Read (2014) and Schmitt (2014), were also consulted. To meet the inclusion criteria for this review, the publications would need to be published, appear in English, and report empirical research, either on the development and validation of a WAF test or using a WAF test primarily as a research tool. To be considered as a WAF test, the focal task/test would need to be based on word association and assess learners' lexical network knowledge in a controlled, receptive format with a target word followed by a number of choices. Consequently, a total of 29 papers with 31 studies met those criteria and were included in this review (see Additional file 1: Appendix A).

The rest of this paper is divided into two major sections, including a review section and a general discussion section. The review section consists of four parts that cover the general practices in the literature, and where applicable research findings as well, about the design features of WAF tests, test administration, scoring, and test-taker characteristics, in correspondence to the four aforementioned questions that guided this review. Additional file 1: Appendix A provides a list of all the studies that involved a WAF test(s) with information about what the test was like and how it was administered and scored. Additional file 1: Appendix B provides a list of the key issues/variables related to the four review areas (i.e., test itself, administration, scoring, and test-takers) and the studies that directly addressed one or more of those issues. Around eight major

issues, the General Discussion section provides some suggestions and directions for the development of WAF tests and the use of them as research tools in the future.

Review

Design features of WAF tests

There are many issues to be considered for developing a WAF test, such as what words to be used, what association relationships to be addressed, how many options should be included, whether the number of associates should be fixed across items or it can vary, how associates are to be distributed among choices, what kind of distractors to be used. As the review below shows, these issues are all important, but not all have received (an equal amount of) attention in the literature; and the research that directly addressed these issues to inform WAT test development and use is also very limited.

Word Frequency

Developing a WAF test first requires selecting a certain number of words as target words (and also choices). An immediate consideration closely related to the content and construct validity of a WAF test is whether learners should have at least partial knowledge of a target word (and its choices) or it is acceptable to include low frequency words that may not be known to the learners. Presumably, all words should be known to the target group(s) of learners so that any variance in performance on the test would represent learners' individual differences in network knowledge rather than how many words (in the test) they actually know, which would be the focus of a vocabulary size test.

Given the aforementioned concern, there is perhaps no surprise that high frequency words or words considered the most useful for target group(s) of leaners were often sampled for developing a WAT test. In developing the first version of his WAT for university ESL learners, *Read (1993), for example, randomly selected the 50 stimulus words from the University Word List (Xue and Nation 1984), which is a list of 836 words commonly appearing in academic texts. In addition, the choice words were ensured to have similar or higher frequency than the stimulus words. Later, for the revised WAT, *Read (1998) seemed to give more careful consideration for high frequency words as the target words were mostly sampled from Bernard's Second and Third Thousand Word Lists. "Because the purpose of the test was to measure depth of knowledge, the emphasis was on words with which most of the test-takers were likely to have at least some familiarity" (*Read 1998, p. 45). *Schoonen and Verhallen (2008), in their report on the development of a Dutch WAT for young L2 learners, also made it clear that "the starting point was that familiarity with all the words could be taken for granted for nine-year-olds in Grade 3 of primary school (the youngest target group)" (pp. 219–220).

Despite careful consideration for word frequency, learners' familiarity with the words developers chose is obviously "an assumption, not a certainty" (*Schoonen and Verhallen 2008, p. 219). A consequence of this lack of certainty is that the validity of the test might be threatened. While some learners may choose not to respond for target words that they do not know and are unwilling to guess, other may make guesses in a similar situation (*Read 1993; 1998; Schmitt 2014). *Read (1993), for example, found from learners' verbal

reports that some higher proficiency learners tended to guess, often with success, for target words that were unknown or partially known (e.g., *denominator* and *diffuse*).

Although selecting high frequency words presumably known to all learners appeared to be a general principle, a few studies attempted to examine how words of distinct frequency bands may indeed have an impact on WAF tests. In their development of a WAF test for Dutch-speaking university learners of French, *Greidanus and Nienhus (2001) selected the 50 target words and their 300 choices from five distinct frequency bands based on a frequency list of about 5000 words (i.e., 1000, 2000, 3000, 4000, and 5000). There was a general pattern that learners scored significantly better on more frequent items than less frequent ones, which was true for learners with different years of studying French in their respective university. Yet, no significant difference was observed between the two lowest frequency levels (i.e., 1000 and 2000) for the more proficient group of learners. A subsequent study (*Greidanus et al. 2004), however, produced mixed findings. For the initial version of a Deep Word Knowledge (DWK) test, very similar word frequency effects like in *Greidanus and Nienhus (2001) were observed. An improved version of the DWK, however, overall, showed significant frequency effects only among less proficient learners but not more proficient learners (and native French speakers). *Greidanus et al. (2005) further addressed the effect of word frequency on test performance by selecting the target (and choice) words from the frequency range of 5000–10,000. Different from the findings of *Greidanus and Nienhus (2001), no significant difference was found across all the frequency levels (i.e., 6000, 7000, 8000, 9000, 10,000) for all groups of participants. In *Horiba's (2012) Japanese WAF test, all the words were selected from the two higher or more difficult levels of the four levels of words for the Japanese Language Proficiency Test. It was found that Korean-speaking learners' performance was significantly better on more frequent target words than on less frequent ones; yet, such a word frequency effect did not surface for Chinese-speaking learners.

The lack of significant frequency effects in *Greidanus and Nienhus (2001) for more advanced learners, *Horiba (2012) for Chinese learners, and *Greidanus et al. (2005) seemed to suggest that there may be a "threshold" word frequency for a frequency effect to occur or not occur among learners at a certain proficiency level. On the one hand, if test words are highly frequent, such as in the case of those selected from 1000 and 2000 frequency bands in *Greidanus and Nienhus (2001), advanced learners might have developed similarly strong network knowledge for all words within the frequency band(s) disregarding the words' actual frequency; on the other hand, if words tend to be low in frequency, such as in the case of *Horiba (2012) and *Greidanus et al. (2005), it might make the WAF test very similar to a test of vocabulary breath. In *Horiba's (2012) study, Korean-speaking learners' WAF test scores showed a very high correlation (r about .90) with a measure that assessed their vocabulary breadth, which perhaps explains why over and above breath, depth did not explain any significant amount of additional variance in reading comprehension.

In summary, while it seems desirable to ensure that the words of a WAF test are high in frequency, as this would help reduce the kind of guessing effect reported in *Read (1993; 1998) and thus the possibility of threatening the validity of the test, having the frequency too high might end up with the test not being able to discriminate the depth knowledge of advanced learners. On the other hand, having all words too challenging

(or very low in frequency) might threaten the validity of the test as a depth measure, too, as there is a risk that the test may be essentially addressing learners' vocabulary breadth, in addition to a risk of guessing. The intricate issue of word frequency warrants more attention in future research, preferably with psychometric evidence over and beyond test score comparisons.

Word class

In addition to frequency, another consideration for word selection is about word class. While all studies used content words (i.e., nouns, adjectives, and verbs) (see Additional file 1: Appendix A), they differed in which one or more of the word classes were included. *Greidanus and Nienhus (2001), for example, included nouns, adjectives, as well as verbs (but mostly nouns) in their Dutch WAF test. *Henriksen's (2008) word connection task included an equal number of nouns and adjectives. *Schoonen and Verhallen (2008) included words of all three form classes, but their proportional distributions were unknown. In *Read's (1993) initial version of the English WAT, target words also included adjectives, nouns, and verbs. However, the heterogeneity in structure across items led him to suggest that "it will be necessary to develop tests that focus on more homogeneous subsets of vocabulary items so that greater consistency can be achieved in the semantic relationships among the words and in the pattern of responses elicited" (p. 369). Consequently, *Read (1998) chose to focus only on adjectives in his revised WAT. Such an approach was also adopted by *Qian and Schedl (2004) when they developed a Depth of Vocabulary Knowledge (DVK) measure for possible inclusion in the TOEFL.

The narrower focus on adjectives only, which did not seem to result in weak psychometric properties of a WAF test (*Read 1998; *Qian and Schedl 2004), is not free from concerns. For example, when the test is used as a research tool for measuring learners' vocabulary depth (e.g., *Akbarian 2010; *Guo and Roehrig 2011; *Qian 1999; 2002; *Qian and Schedl 2004; *Zhang 2012), the scope of the competence measured would be necessarily narrow and would not be able to capture the full repertoire of the network knowledge that learners have; as a result, the predictive effect of vocabulary depth (on reading comprehension) might have been underestimated. In addition, as *Dronjic and Helms-Park (2014) argued, noun phrases, such as in the case of adjective-noun collocations like *sudden change*, should have the noun as the center. In natural speech production of adjectival phrases, nouns govern the choice of their modifiers, rather than the other way around, whether the modifier is prenominal or post-nominal. To this end, a WAF test with adjectives as target words and nouns as choices seems to test a process opposed to that in natural speech production in that the test requires learners to begin with a modifier and then search for its potential heads.

The above concerns seem to suggest that it would be desirable to include words from different form classes and it might not even be a good choice to have adjectives included. A conclusion like this, however, would be too hasty, in view of the heterogeneity of learner responses reported in *Read (1993). A deeper understanding of this issue would require comparisons of learners' response patterns across words of different form classes as well as the psychometric properties of subsets of a WAF test that includes different word classes. So far, there has been little research in this direction. In

*Read (1993) where item-wise heterogeneity was discussed, how the heterogeneity might be specifically attributed to diverse word classes was unknown because the author did not conduct any direct comparison by either drawing upon learners' verbal reports or examining the psychometric properties of the test separately for verbs, adjectives, and nouns. *Greidanus and Nienhus (2001) and *Schoonen and Verhallen (2008), while both including words of different form classes as mentioned earlier, did not attempt to separately analyze them and make any comparison, either. Using a free association task (as opposed to a WAF test), Nissen and Henriksen (2006) found word class tended to moderate the distribution of associates belonging to different types of association relationships among Danish-speaking learners of English as a Foreign Language (EFL). The authors suggested that words of different form classes may be organized differently in learners' mental lexicon. Thus, it should be a strong interest in future research to examine whether verbal, nominal, and adjectival target words might involve different thought processes among learners, and more importantly, whether or not learners' responses to subsets for those word classes would be unidimensional in assessing their network knowledge.

Association relationships

Another decision to be made for a WAF test is the types of association relationships to be tested for the target words selected. As shown in Additional file 1: Appendix A, there were variations in what association relationships were addressed in previous WAF tests, but most included paradigmatic and syntagmatic association (e.g., *Qian and Schedl 2004; *Read 1998), with some others considering analytic association as well (e.g., *Greidanus and Nienhus 2001; *Greidanus et al. 2005; *Horiba 2012; *Read 1993).

Possibly because of the predominance of paradigmatic associates in L2 learners' free association responses over other types of association, including syntagmatic association (e.g., Jiang 2002; Wolter 2001), there was an interest in the WAF literature to compare learners' performance for different types of association relationships. *Greidanus and Nienhus (2001), for example, found Dutch-speaking learners of French, disregarding the number of years of university learning of French, consistently showed a better performance for paradigmatic (and analytic) association than for syntagmatic association. *Greidanus et al. (2005) largely replicated that finding, with a similar test that had lower frequency target words, among more diverse groups of Dutch-speaking French learners (as well as native speakers). A similar finding was also observed in *Horiba (2012) among Chinese-speaking learners of Japanese, but not Korean-speaking learners, for whom there was no significant difference between paradigmatic and syntagmatic association.

*Dronjic and Helms-Park (2014) administered *Qian and Schedl's (2004) DVK to two compatible groups of native English speakers. Largely corroborating the aforementioned findings, paradigmatic scores were found to be significantly higher than syntagmatic scores, disregarding whether the participants knew the number of associates to be selected and how the test was scored. In addition, the participants' responses for syntagmatic association were far more heterogeneous than for paradigmatic association, which seemed to resonate a concern that *Read (1998) voiced about his differentiation between the two types of association relationships in his revised WAT. Specifically, there is a

question as to "whether the two types of associates represent the same kind of knowledge of the target words, or whether, say, the 'semantic' knowledge expressed in the paradigmatic associates is distinct from the 'collocational' knowledge tapped by the syntagmatic ones" (p. 57).

An initial understanding about whether paradigmatic and syntagmatic (and other types of) association may tap the same kind of network knowledge can be obtained from the correlational relationships reported in some studies. *Greidanus and Nienhus (2001), for example, in addition to comparing learners' scores, found that the correlations between the three types of relationships themselves (i.e., paradigmatic, syntagmatic, and analytic) and their correlations with learners' general French proficiency were all very small and non-significant. *Horiba (2012), in contrast, found strong correlations between the same three types of association relationships for Korean-speaking learners of Japanese (*r*s about .79-.85). In addition, all three types of association also showed strong correlations with their vocabulary breadth (*r*s about .77-.91) with paradigmatic association demonstrating the highest correlation. Yet, the correlations between the three types of association were all very small, albeit significant, among Chinese-speaking learners (*r*s about .29-.36). And only paradigmatic and syntagmatic association showed significant correlations with vocabulary breadth (*r*s about .34-.60). Notable variations were also observed in the correlations with reading comprehension for the three types of association (and between the two groups of learners). *Qian and Schedl (2004) found the paradigmatic and syntagmatic sections of the DVK showed a significant correlation of about .80 among university ESL learners. Using the same test and the same scoring method, *Dronjic and Helms-Park (2014), however, found native English speakers' performance on the two subsets showed much smaller, albeit significant, correlations (.383-.460). Only when a new scoring method that gave credit to selection of associates as well as rejection of distractors was used did the two types of association show a strong correlation (.78–.88). The authors thus suggested that "knowledge of collocability and knowledge of semantic relations such as synonymy and polysemy might represent two different dimensions of lexical depth. This finding also reflects the commonsense observation that it is at least theoretically possible to have a speaker who knows a lot about word meanings and little about how these words combine with other words…" (p. 210).

Two studies went beyond simple bivariate correlations to examine the factor structure of association knowledge measured with different types of association. In *Shin's (2015) study on elementary school EFL learners in Korea, two different sets of items were designed to address paradigmatic and syntagmatic association, respectively, which made the test different from *Read's (1998) WAT and many other WAF tests where paradigmatic and syntagmatic association were addressed simultaneously for the same target words. The test as a whole and the two subsets all had moderate and significant correlations with the students' performance on a standardized reading comprehension test. More importantly, confirmatory factor analysis (CFA) revealed that all test items significantly loaded on the factor of their respective association relationship (factor loadings from .35 to .77). In addition, the two factors were also significantly and strongly correlated ($r = .83$).

While *Shin (2015) concluded from the aforementioned CFA result that paradigmatic and syntagmatic relationships tap rather different dimensions of deep word knowledge,

the strong correlation between them seemed to suggest significant overlap between them and the two "factors" might further load on a higher-order factor. In this respect, *Batty's (2012) study provided a more nuanced understanding. Based on *Qian's (2002) DVK administered to Japanese-speaking university EFL learners, *Batty (2012) tested three CFA models: the one factor model hypothesized that all the WAT items loaded on a general vocabulary factor; the two-factor model hypothesized that the syntagmatic and paradigmatic associates formed two separate but correlated factors (i.e., the model tested in *Shin 2015); the bifactor model hypothesized that all WAT items loaded on a single general factor (of vocabulary depth), while the syntagmatic and paradigmatic associates additionally loaded on two separate, smaller factors. The first two models did not show satisfactory model fits, although the two-factor model had slightly better fits than the one-factor model. The bifactor model exhibited the best fits of all three models, and the highest item loadings were overall largely on the general factor of vocabulary knowledge/depth, which was particularly true for the loadings of syntagmatic associates. Despite this finding, it seemed difficult to conclude that paradigmatic and syntagmatic association are unidimensional in accessing vocabulary (depth), because the one-factor model showed poor model fits on the one hand and the raw correlation between the two types of items was low (r = .61) compared to the correlations within each type of items (rs = .79 and .80 for syntagmatic and paradigmatic association, respectively) on the other. In addition, half of the paradigmatic associates loaded actually more highly on the synonym factor than on the general vocabulary factor.

The above studies, which differed in focal languages, learners, and WAF tests, painted a complex picture about whether different types of association tap the same or different aspects of network knowledge. Before more conclusive evidence is obtained in the future, it does not seem as explicit as researchers often did to simply aggregate the scores for different types of association to form a total score to represent learners' vocabulary depth knowledge (see Additional file 1: Appendix A). It is suggested that in future use of WAF tests, separate scores for different types of association and their internal relationships (e.g., correlation) should at least be first reported before a combined score is used to index learners' vocabulary depth. In a case when vocabulary depth is used to predict other variables, such as lexical inferencing (e.g., *Ehsanzadeh 2012; *Nassaji 2004) or reading comprehension (e.g., Akbarian 2010; *Guo and Roehrig 2011; *Horiba 2012; *Qian and Schedl 2004; *Zhang 2012), we suggest that separate analyses be done for different association relationships (with or without analysis using the aggregated score) so that the findings could be more revealing about the nuance of the role of lexical network knowledge in language skills development.

Number of options

Although the options for a target word could be any number, most WAF tests, as shown in Additional file 1: Appendix A, had six (e.g., *Greidanus and Nienhus 2001; *Schoonen and Verhallen 2008) or eight options (e.g., *Read 1993; 1998; *Qian 1999; 2002; *Qian and Schedl 2004), typically with an equal number of associates and distractors. Two notable exceptions are *Horiba's (2012) Japanese WAF test, which had seven options with three associates and four distractors, and *Henriksen's (2008) word connection task, which had ten options with five associates and five distractors.

Despite the variation in number of options, it was unclear whether the choice made for a particular WAF test was more for practical considerations or psychometric advantages. Presumably, having a smaller number of options would give developers more flexibility to find appropriate associates and distractors within a particular pool (or frequency range) of words, and thus reduce the possibility of having to include less frequent words because not enough words at the desirable frequency level could be found. This consideration was represented in Read's (1993) decision to not restrict his choice words to the University Word List. In addition, having fewer options may also make the test less challenging to young learners (*Schoonen and Verhallen 2008).

So far there was little research on how having different numbers of options may influence, if at all, the psychometric properties of WAF tests. *Schmitt et al. (2011) seemed to be the only study that directly compared WAF tests with different numbers of options. In their Study 2, *Schmitt et al. (2011) administered to university ESL learners a WAF test designed following *Read (1998). The test had two versions, one with 6 options and the other 8 options. After taking a paper-and-pencil test for both versions, learners were interviewed to demonstrate their actual knowledge of the words in the test and share the thought processes for their responses. Based on the interview responses, learners' degree of knowledge was coded at three levels, including no knowledge, partial knowledge, and full knowledge. In relation to the focus of this review part on number of options, it appeared that compared to the 6-option version, the 8-option version resulted in a higher proportion of cases of "mismatch" between learners' knowledge assessed through the paper test and the interview. The authors suggested that the 8-option format tended to overestimate learners' actual knowledge more seriously than did the 6-option format and thus may be a less desirable choice for assessing learners' depth knowledge. More validation evidence is needed in the future.

Number of associates

In either situation, 6 or 8 options, there is a need to consider two additional issues with respect to the number of associates. The first one is whether the number of associates should be fixed or it could be allowed to vary across items. As shown in Additional file 1: Appendix A, in most cases, WAF tests had a fixed number of associates with an equal number of associates and distractors. For example, *Read's (1998) WAT and *Qian and Schedl's (2004) DVK items all have four associates (out of eight choices). *Schoonen and Verhallen's (2008) Dutch WAT and *Greidanus and Nienhus (2001) French test items all had three associates (out of six choices). The French tests in *Greidanus et al. (2004; 2005) seemed to be the only ones that had varied numbers of associates (2–4 but with six choices for all).

The second issue is an extension of the first one, that is, should there be an equal number of associates (and distractors) for different types of association relationships? Many WAF tests were unclear on the total number of associates for different types of association relationships. In *Read's (1998) revised WAT, for example, there were three types of distributions for paradigmatic and syntagmatic associates in the left and right boxes, respectively, including 1–3, 2–2, and 3–1. Thus, there is a possibility of associates for the two types of association not being equally represented in the test, which may pose a challenge when there is a need to compare between those types of association, which we

reviewed earlier.[1] This was perhaps a reason that in the few studies where different types of association were compared, each item had one associate for each type of association so that different types of association had an equal number of associates or the same range of scores. For example, in *Greidanus and Nienhus (2001) and *Horiba (2012), an item had three associates (the others were all distractors), one for paradigmatic association, one for syntagmatic association, and the third one for analytic association.

A question to ask about the above issues related to the number of associates is, do they really matter for WAF tests, for example, by affecting their psychometric properties? If *Greidanus et al.'s (2005) argument holds, that is, "if there were always three correct responses, the participants could make their choice by elimination. With a variable number of correct responses they had to determine each time whether the responses word belonged to the network of the stimulus word" (p. 194), having varied numbers of associates across items would mean higher validity of a WAT test. So far, no studies, however, have directly tested this issue.[2] Thus, no evaluation could be made on whether it would be preferred to vary the number of associates across items or have the number fixed. In addition, it was unclear whether there is a need to make sure the number of items is the same for different association relationships, given the possibility of using proportion of correct responses (*Qian and Schedl 2004) and/or scoring methods to create a "balance" between them.

Distribution of associates

In *Read's (1998) revised WAT where the choices for paradigmatic and syntagmatic association were presented in two separate groups/boxes, there were varied distributions of associates, as it was believed that having the same pattern of distribution for all words, such as two associates (and two distractors) in each box, might provide a pattern for learners to follow, and thus might lead to guessing. Consequently, three distributions of associates for paradigmatic and syntagmatic association were adopted to counteract guessing (i.e., 1–3, 2–2, and 3–1). Such a feature of test design appeared to be effective, as it was found in *Qian and Schedl (2004) that learners interviewed all reported that it was difficult to guess, because the number of associates in each box of the DVK, which followed the format of *Read (1998), was not fixed.

On the other hand, *Schmitt et al.'s (2011) Study 2 suggested that the validity of WAF tests could be possibly impacted by associate distributions. Specifically, the study revealed correlations of different strengths for items with different distribution patterns for paradigmatic and syntagmatic associates. For example, those items with the 2–2 distribution showed the strongest correlation between the paper test scores and the scores on an interview, which was believed to better represent learners' actual depth knowledge (r about .871). In contrast, the items with the 1–3 distribution showed the lowest correlation (r about .736), which, as the authors explained, might be attributed to the semantic relatedness of the (three) syntagmatic associates in the choices. In other words, like the learners in *Read (1993), those learners might have used patterns in the choices to guess successfully for a target word that they did not even know.

*Schmitt et al.'s (2011) finding appeared to suggest that compared to the other two distributions, the 1–3 distribution might be most susceptible to guessing and overestimate learners' actual depth knowledge. However, it is noted that the source of guessing

(i.e., semantic relatedness of syntagmatic associates) was perhaps a result of the difficulty that the authors had in finding three collocates with distinct meanings. Thus, it is still questionable whether it was the 1–3 distribution *per se* or a lack of appropriate selection of choice words for test items with that distribution that had resulted in the lowest correlation. More research is certainly needed in the future to further our understanding of how learners' performance, response behaviors, and the psychometric properties of WAF tests may be influenced by how associates are distributed. It might be because of the simpler pattern of associate distribution in the 6-option test (there are only two variations: 1–2 and 2–1) that *Schmitt et al. (2011) did not examine the issue for this format. Nevertheless, given that this format was perhaps the most commonly used one in the literature (see Additional file 1: Appendix A), it should be of interest to explore it as well in future research, and compare the findings with those revealed by *Schmitt et al. (2011) about the 8-option format.

Distractor properties

Another important feature to consider for WAF tests is what distractors to include, notably whether distractors should be semantically related or unrelated to target words. *Read (1993; 1998) argued that distractors should not have semantic links to the stimulus word. This principle was subsequently followed in some other studies, such as Greidanus et al. (2004), but not universally endorsed (e.g., *Henriksen 2008; *Schoonen and Verhallen 2008). *Schoonen and Verhallen (2008), in their WAF test for young Dutch learners, purposefully used semantically related distractors, which had less strong an association with the target words than the associates. The authors believed that generalization and abstraction play an important role in students' word knowledge development, and argued that it is "on the basis of these processes that the attribution of meaning is gradually decontextualized" (p. 157). Thus, WAF tests as a measure of depth knowledge should assess learners' generalized, decontextualized knowledge of a stimulus word (i.e., "words that always belong to the target word;" p. 157), such as *fruit, yellow*, and *peel* for *banana*, rather than knowledge of incidental, content-dependent meanings (e.g., *monkey* for *banana*). In *Henriksen's (2008) word connection task, each target word was followed by 10 words. The five answer words showed the most frequent associations from many native speaker norming lists, whereas the five semantically related distractors were "infrequent responses given by only one native speaker in the norming lists; that is, these words represent potential but clearly more peripheral links in the lexical net" (p. 42).

*Greidanus and Nienhus (2001) explicitly tested how different types of distractors would have an influence on their French WAF test for Dutch-speaking university students. It was found that learners of different proficiency groups consistently showed better scores for the items with semantically related distractors than for the same items with semantically unrelated distractors. This score difference did not appear to be a surprise, given that semantically unrelated distractors were much easier to eliminate than related ones, and that the scoring method valued successful elimination or non-selection of distractors.

Learners' better scores on items with semantically related distractors, however, do not indicate that those items should necessarily be a more preferred test design. *Read (1998), when validating the revised WAT where distractors were semantically

unrelated, found from students' verbal reports that more proficient learners tended to use the relationships, or the lack thereof, among the options, to make guesses on associates for unknown stimulus words. In this respect, a test with semantically related distractors may make guessing harder and engage learners' network knowledge better and thus be a more preferred design. This seemed to be partly confirmed by the better reliability of the items with semantically related distractors (.76, as opposed to .63 for items with semantically unrelated distractors) as well as the correlations *Greidanus and Nienhus (2001) found between the two sets of the test and between the test and learners' general proficiency. Specifically, disregarding learners' proficiency level, the two item types did not show a significant correlation, suggesting that they may tap network knowledge in distinct ways as a result of the variation in distractor properties. In addition, in the lower proficiency group, the scores of neither item type correlated significantly with general proficiency; in the higher proficiency group, however, the items with semantically related distractors, as opposed to those with unrelated distractors, correlated significantly with learners' general proficiency.

*Schmitt et al.'s (2011) validation study suggested that other factors related to WAF tests may also need to be considered for evaluating different types of distractors. In Study 2, *Schmitt et al. (2011) correlated university ESL learners' scores for a paper WAF test where the items had three types of distractors with their scores on an interview which elicited their actual depth knowledge. It was found that the items with distractors having no semantic relationships ($r = .776$) produced a notably weaker correlation with the interview scores than did the items with semantically related distractors ($r = .910$) for the WAF test with six options, whereas in the WAF test with eight options, a reverse pattern was found (rs = .912 and .813, respectively). In both test situations, the correlations between the paper test scores and the interview scores were the least strong for the distractors with orthographic resemblance to stimulus words (rs = .636 and .663, respectively, for the 6-option and the 8-option format). The authors thus concluded that "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," and "formal distracters should be less frequently used" (p. 121).

Whether or not distractors should be semantically related or unrelated to the target word is certainly an issue that deserves more research in the future. On the one hand, it should be helpful to include other types of validation evidence, such as concurrent or predictive validity, to compare the two item designs; on the other hand, variation in item design would also need to be considered in conjunction with other test-related variables, particularly scoring method. *Read (1998), for example, adopted a scoring method that only considered association selection for his English WAT with semantically unrelated distractors; *Schoonen and Verhallen (2008), on the other hand, adopted a method that awarded a point only if the response precisely matches the answer for their Dutch WAF test based on semantically related distractors. Yet, in *Greidanus and Nienhus (2001) where the two item designs were compared, a third method that gives credit for both associates selection and distractor non-selection was used. Thus, it appears that comparing distractor types in isolation from other factors which also have an impact on learners' thought processes or response behaviors would not provide the best evidence for evaluation.

Administering WAF tests

A second set of issues that concern WAF tests pertains to how the tests should be (better) administered. For example, should the test be conducted in print or would it make any difference if all the words are read aloud to learners? Should learners be informed on the number of associates or should they be asked to select as many as they believe to be the associates even though the number of associates is fixed across items? Additionally, given that different scoring methods were often used (see the next section on Scoring WAF Tests), how might learners' knowledge, or the lack thereof, about the method to be used for scoring their responses influence their response behaviors and the psychometric properties of a WAT test?

Written vs. aural modality

As Additional file 1: Appendix A shows, WAF tests were almost exclusively administered in the written form. Learners typically complete a paper-and-pencil test. In rare situations were learners also interviewed for the purpose of validating a paper test (e.g., *Read 1993; 1998; *Schmitt et al. 2011). The issue of modality in assessing vocabulary knowledge is of course not relevant to WAF tests alone. All conceptualizations of what it means to know a word seem to involve word forms, such as sound and orthography (e.g., Nation 2001; Read 2004). This suggests that vocabulary tests all need to assess learners' ability to identify a word in both aural and written modalities. In the L1 literature, however, it is not necessarily the case in that (young) learners typically have acquired a lot of word meanings (and their semantic links) through oral language acquisition yet without being able to recognize all those words in print - they need to learn to decode words to access their meanings (and the network of meanings) in the mental lexicon. Thus, oral vocabulary is often distinguished from written vocabulary. This also explains why young children's (free) word association was often elicited through oral interviews, and why there are oral vocabulary knowledge tests which do not require students to decode print words, such as the Peabody Picture Vocabulary Test.

In the L2 literature, which used to be concerned more about teaching foreign language learners (as opposed to learners of a second language), there tended to be an assumption that "if a word is known, then it is likely to be known in both written and aural forms" as a result of the concurrent focus of classroom instruction and learning on both sound (pronunciation), orthography (spelling), and meaning (Milton 2009, p. 93). Consequently, a view of assessment might have been taken that conceptualizes the ability to recognize words in print as an integral component of vocabulary knowledge and should thus be tested as such. This view could be legitimate if the assessment focus is on form(orthography)-meaning connection, which is the focus of most vocabulary size tests, such as the VLT and Yes/No tests. In other words, variance due to individual differences in the ability to process the orthographic forms of target words (for meaning access) might be considered as "construct-relevant." In the case of WAF tests, however, the legitimacy of this view may be questionable, as the primary assessment focus is on meanings and their links (i.e., network knowledge) (Read 2004). In other words, possible variance induced by orthographic processing should be a significant factor to consider for WAF tests.

Such an issue seems to be particularly salient for learners of non-alphabetic languages who come from an alphabetic background, such as English-speaking learners of languages like Chinese and Japanese where a logographic system is used. Unlike alphabetic languages like English, Dutch, and French, which follow the rule of phoneme-to-letter correspondence and allow for the use of alphabetic principle to decode words, Chinese characters and Japanese kanji are square-shape symbols composed of strokes and stroke patterns. Thus, to access the meaning of a word in print, which is often composed of multiple characters, the component characters need to be recognized without the kind of immediate phonological clues available as in the case of words in alphabetic languages.

Thus, given the primary assessment focus of WAF tests on network knowledge, any demonstrated knowledge through the tests presumably should not be confounded by learners' failure to recognize characters or words in print, which would threaten the validity of the tests. This might be a reason why Jiang (2002) adopted the aural modality when he administered a free association task to Chinese L2 learners, that is, learners listened to target words and orally provided associates. It also seems to explain the consideration that *Horiba (2012) had for including kana syllabaries together with kanji characters in her Japanese WAF test.

The aforementioned issue may be particularly important for the assessment of depth knowledge for those who learn Chinese or Japanese with substantial aural/oral experiences with the language, such as heritage language learners or non-heritage learners who spent a substantial amount of time learning the language in a second language or societal context (as opposed to those who learn the language primarily in a foreign language context through classroom instruction).[3] In other words, without appropriate control (e.g., adding pinyin for Chinese words and kana for Japanese kanji) or consideration for the modality of administration (i.e., administering the test in aural/oral form rather than in print), a WAF test might under-estimate the association knowledge of those learners, or testing bias might occur between foreign language learners and heritage learners or those learners for whom the acquisition occurs in a societal context.

Informed vs. uninformed

Previous studies also varied on whether learners are informed on the number of associates to be selected or whether they are asked to select as many associates as they believe to be appropriate even though there is a fixed number of associates across items. *Read (1993), for example, did not tell learners how many associates they were supposed to select; instead, they were asked to choose as many as possible even though they might not be very sure about an item. This practice was also followed in some studies that used a version of *Read (1993; 1998) (e.g., *Qian 1999; *Zhang 2012). In *Greidanus et al. (2004; 2005), learners were told that the number of associates varied without knowing which item had which number of associates. The authors believed that this would avoid participants making their choices by elimination. Yet, many other studies all included the total number of associates in their test instructions (e.g., *Qian and Schedl 2004; *Schoonen and Verhallen 2008). As argued by *Schoonen and Verhallen (2008) for their Dutch WAT which included semantically related distractors, "it is important that the number of required associations is fixed and specified for the test-takers,

because association is a relative feature. Depending on a person's imagination or creativity, he/she could consider all distractors to be related to the target word. A fixed number is helpful in the instruction because it forces test-takers to consider which relations are the strongest or most decontextualized" (p. 218).

From the perspective of learners' thought processes for working on a WAF test, if they know the number of associates, they might tend to stop when they have had the correct number of associates selected, or they would be pushed to engage further with the other choices when they are sure about only one or two of the associates. On the other hand, when learners are not informed on the number of associates, they might tend to only choose those associates with complete certainty (or alternatively, use wild imagination to select as many as possible, which was the concern of *Schoonen and Verhallen 2008), which could end up with selecting fewer associates than the answer, and avoiding deep engagement with other choices to select the correct number of associates. This presumed effect of the condition under which a WAF test is administered (i.e., informed vs. uninformed) on learners' test-taking behaviors might have an impact on their test performance, and more importantly, the psychometric properties of the test.

The aforementioned issue has received little attention in the literature. *Greidanus et al. (2005) administered two parallel DWK tests to Dutch- and English-speaking learners of French (as well as native speakers). As described earlier, both tests included six choices, with the first included a fixed number of associates (i.e., 3) and with learners informed on this number, whereas in the second one, the number of associates varied, and learners were told that the number varied but did not know which item had which number of associates. Learners' performance did not indicate a significant difference between the two tests. Although the first seemed to be slightly stronger in reliability (α = .91) than the second (α = .88), no other information about the psychometric properties of the two tests was reported.

*Dronjic and Helms-Park (2014) specifically manipulated the administration condition by asking two compatible groups of native English speakers to work on *Qian and Schedl's (2004) DVK. One group took the test under the constrained (informed) condition knowing that all items had four associates, and the other group worked in the unconstrained (uninformed) condition being asked to select as many choices as they deemed appropriate. Among many other findings, disregarding scoring methods, the constrained condition showed significantly better performance than the unconstrained condition. In addition, the participants' scores were more homogeneous in the constrained condition than in the unconstrained condition.

*Dronjic and Helms-Park's (2014) findings seemed to suggest a preference over letting learners know the number of associates. While the more homogeneous test performance may not be conclusive evidence for such a preference, an implication for research practice is obvious in that there is a need for researchers to specify how their WAF test is administered, which was not always the case in the literature (see Additional file 1: Appendix A). Given the findings reviewed above, without knowing how learners were instructed to work on the test, it would be difficult to make reasonable comparisons on learners' test performance across studies on the one hand and evaluate the psychometric properties of different WAF tests on the other.

Scoring WAF tests

At least three methods have been used in the literature for scoring WAF tests. The first method, which was initially adopted by *Read (1993), scores learners' responses based on their associate selection only. In other words, neither non-selection of any distractor would be awarded a point nor the selection of one would be penalized. The second method awards a point for selection of an associate as well as non-selection of a distractor. The third method and perhaps also the strictest of the three, awards a point for a response only if it precisely matches the answer (i.e., selection of all associates but not any distractor).

Named by *Schmitt et al. (2011) as the One-Point, Correct-Wrong, and All-or-Nothing methods, respectively, all three methods have been adopted in previous studies (see Additional file 1: Appendix A). Yet, empirically to what extent one might be preferred over another has received little attention in the literature. *Dronjic and Helms-Park (2014) found from native English speakers' responses to *Qian and Schedl's (2004) DVK that scoring methods moderated how association relationships (paradigmatic vs. syntagmatic) and administration condition (constrained vs. unconstrained) affected learners' test performance. For example, with the One-Point method (i.e., the original scores), ANOVA revealed main effects of both type of association relationships and administration condition, and there was no significant interaction effect; with the Correct-Wrong method (i.e., the revised scores), while the main effects remained significant, a significant interaction effect also emerged, that is, the difference between the two conditions was notably smaller for paradigmatic association and larger for syntagmatic association.

*Schmitt et al. (2011) found that for the WAF test with six options, the All-or-Nothing method produced the largest correlation (.884) between learners' paper test and interview scores. The One-Point and Correct-Wrong methods produced correlations of .871 and .877, respectively. For the WAF test with eight options, the One-Point method produced the strongest correlation (.885) and the correlations were .855 and .871 for the Correct-Wrong and All-or-Nothing methods, respectively. *Schmitt et al. (2011) thus concluded that the Correct-Wrong method could be discounted as it is much more complicated than the other two methods without yielding more encouraging results; and the All-or-Nothing might be better for the 6-option format, and the One-Point for the 8-option format.

Despite the evaluation conclusion that favored the All-or-Nothing method, *Schmitt et al. (2011) noted that an evaluation of scoring methods also needs to consider the purpose of the assessment. If the purpose is to know how a student(s) stands in the level of vocabulary depth knowledge compared to his/her peers in a class, then the All-or-Nothing would give the quickest result; and in a research situation, it would perhaps also be time-saving to use this method for scoring responses from a large number of learners. On the other hand, if the purpose is to diagnose the specific associates which learners have mastered or where they may display poor knowledge, the One-Point or the Correct-Wrong method might be a better choice in that they could give more specific information of diagnosis.

It was indicated at the beginning of the previous section on test administration condition that it is perhaps also important to consider the issue of whether or not learners know how their responses are to be scored. No studies listed in Additional file 1:

Appendix A had explicit information about this issue. In language testing, familiarizing test-takers with the format of a test is critical, and in large scale testing, test-takers are usually well-informed on what and how they will be tested, including how their answers will be scored, notably for performance tasks, such as speaking and writing. In the vocabulary assessment literature, many tests have a single correct answer, such as the VLT and Yes/NO checklists. Thus, the scoring method (as in the form of correct number of choices), even though uninformed, could be immediately clear to learners. On the other hand, in the case of WAF tests where multiple choices are usually expected, learners' knowledge of which method is used to score their responses should be an important issue to consider.

To illustrate how learners' test-taking behaviors (and thus the validity of a WAF test) might be influenced by whether or not they know the scoring method to be used, the "One-Point" method, in an extreme case, may result in everybody selecting all choices to achieve the maximum score if the method is known to learners. On the other hand, if learners know that their responses are to be scored with the Correct-Wrong method, they would presumably discourage themselves from using such a testing strategy. In this regard, the lack of clarity in the literature on this issue could be disconcerting. It is suggested that future studies that use a WAF test should specify, among other things, whether or not their participants are clear about how their scores are to be scored. Research is of course also need in the future to examine how learners' knowledge of the scoring method, or the lack thereof, might influence their thought processes and performance on a test and subsequently the psychometric properties of the test.

Test-takers and WAF tests

The impact of test-taker characteristics on test performance and the psychometric properties of a test has long been acknowledged in language testing (Bachman and Palmer 1996). While some characteristics can be unpredictable and may not be controlled for, there are possibilities to avoid or minimize the effects of others through careful attention to test design and administration. Bachman and Palmer (1996), for example, discussed four major categories of individual characteristics of test-takers, including personal characteristics (e.g., age, sex, and native language), the topical knowledge that test-takers bring to the language testing situation, their affective schemata, and language ability. (O'Sullivan, B: Towards a model of performance in oral language testing, unpublished) classified test-taker characteristics into three major categories, including physical/physiological (e.g., age, sex, and disabilities), psychological (e.g., personality, cognitive still, motivation), and experiential (e.g., education, examination preparedness, and target language country residence). Using these lists as a reference, only a few studies in the WAF literature directly addressed a small number of test-taker characteristics, such as L2 proficiency (e.g., *Greidanus and Nienhus 2001; *Henriksen 2008; *Schmitt et al. 2011) and L1 background (*Horiba 2012).

L2 proficiency

As reviewed earlier, an immediate concern for developing a WAF test is related to target group(s) of learners in that words need to be largely known to the learners so as to make the test assess an aspect of knowledge (i.e., depth) distinct from vocabulary size. Thus, learners' proficiency should indeed be an important consideration; yet, only

a few studies specifically compared learners at different proficiency levels to examine the possible impact on WAF tests.

In *Read (1993), ESL learners' verbal report indicated that when they encountered an unknown stimulus word, their strategies appeared to vary according to proficiency level. While lower proficiency learners were more inclined to skip such an item, learners with higher proficiency were more willing to guess, such as using relational information in the choice words for selecting associates. Later, however, *Read (1998) revealed that some high proficiency learners also showed unwillingness to guess. Thus, there did not seem to be convincing evidence for a direct relationship between learners' proficiency level and guessing on the WAT or the validity of the test.

*Greidanus and Nienhus (2001) compared the performance of Dutch-speaking learners at different proficiency levels on a French WAF test. Among other findings, those who had more years of university learning of French performed significantly better than those who had less years of studying French, thus providing strong evidence for the discriminatory power of the test. While there is consistency between the two groups in word frequency effects as reviewed earlier, difference was observed between them with respect to the extent to which items with semantically related and unrelated distractors were related to learners' general language proficiency.

A few studies that compared learners at different proficiency levels also generated contrasting insights about the critical issue of the distinction between vocabulary depth and size, more specifically, how strongly depth and size are correlated and how they may be similar or distinct among lower and higher proficiency learners. In other words, whether or not a WAF test indeed measures an aspect of vocabulary knowledge (i.e., depth) distinct from vocabulary size may depend on the learners in question (Schmitt 2014). *Henriksen (2008) administered a free association task and a WAF test (i.e., the word connection task) that had the same target words to examine the depth knowledge of adolescent Danish-speaking EFL learners. While the productive association task showed consistently significant correlations with the VLT for students at all three grades or proficiency levels (i.e., Grades 7, 10, and 13), the WAF test only showed a significant correlation with the VLT for Grade 7 students, which suggested that depth and size may be less distinct at the lower proficiency level than at the higher proficiency levels. *Nurweni and Read's (1999) study on Indonesian university learners of English, however, showed a contrasting result in that the strength of correlations between a WAF and a word translation task gradually decreased from higher proficiency learners (.81) to the middle group (.43) and lower proficiency learners (.18).

The findings of the studies above provided some interesting but necessarily limited insights about how WAF tests may show diverse test performance, response behaviors, and psychometric properties among learners of different proficiency levels. In relation to the many issues reviewed earlier on WAF tests in this paper, there are certainly more to examine the possible impact of proficiency on WAF tests, such as the implications for the relationships between paradigmatic and syntagmatic sections of WAF tests and the number of associates and their distributions.

L1 background and L1-L2 distance

There is extant research in the L2 literature that showed transfer between learners' two languages and how L1-L2 linguistic distance may have an impact on the processing and use of a target language among learners from disparate L1 backgrounds (e.g., Jarvis and Pavlenko 2008; Koda 2005; Odlin 1989). Few studies, however, aimed to examine WAF tests in relation to learners' L1 background or L1 transfer in L2 network knowledge.

A case in point is the syntagmatic section of a WAF test. As discussed earlier, compared with the paradigmatic section, the syntagmatic section tended to show more variability or heterogeneous responses from native speakers of English (*Dronjic and Helms-Park 2014). Given this finding, it seems legitimate to be concerned about how the syntagmatic section may address differentially the depth knowledge of L2 learners from different L1 backgrounds. *Horiba (2012) was the only study that targeted such an issue.[4] It was found that Korean- and Chinese-speaking learners demonstrated differential patterns of performance for paradigmatic and syntagmatic associates on a Japanese WAF test. Specifically, for Chinese learners, paradigmatic association showed the best performance, and syntagmatic association showed the weakest performance, whereas among the Korean learners, the performance on syntagmatic association was comparable to that on the other types of association. *Horiba (2012) attributed this to different L1s of the two groups of learners. Specifically, Japanese and Korean are both agglutinative SOV languages and syntactically similar, whereas Chinese is an isolating SVO language. Thus, compared to their Korean counterparts, who should have benefited from the linguistic proximity between their L1 and Japanese for syntagmatic association, Chinese speakers' L1-based syntactic knowledge did not seem to be very useful in processing Japanese collocations. Aside from the differential patterns in test performance, it is probably more important to note that among the Korean learners, all types of association relationships demonstrated very strong correlations with themselves and with vocabulary breath, whereas among the Chinese learners, the correlations were notably weaker, albeit all significant, too, than among the Korean learners. Thus, it was possible that the test also differed in its validity of measuring the depth knowledge for Chinese and Korean learners.

General discussion

To provide a general answer to the four questions that guided this review, WAF tests' design features and the ways they are administered and scored all seemed to have an impact on test-takers' responses or performance, and more importantly, the psychometric properties of the tests. In addition, test responses and psychometric properties also appeared to vary across different types of test-takers. Yet, as is clear from the review, the validation evidence for WAF tests is limited, even though they overall have been found to be reliable and valid in assessing learners' depth of vocabulary knowledge with a focus on their lexical network knowledge. In addition, the huge variations in how WAT tests have been used as research tools (see Additional file 1: Appendix A) seem to suggest that the findings derived from those studies about learners' vocabulary knowledge development or the effect of vocabulary depth on language proficiency development may need more careful interpretations. There are also issues that have received scant attention in the literature. To this end, this section provides a general discussion and highlights some suggestions and directions for the development of WAF tests and the use of them as research tools in the future.

Imbalance

It was obvious that current research gave much more attention to diverse design features of WAF tests themselves than to issues concerning test administration, scoring, and test-taker characteristics. A strong focus on the test themselves is arguably important, of course. However, without adequate attention to other issues, the understanding for the practices that were often followed for developing or using WAF tests would be limited, and in some circumstances, the assumptions could be wrong. As a highlight, a common practice for using WAT tests as research tools is that the scores for paradigmatic and syntagmatic association are simply aggregated to form a total score to index learners' vocabulary depth. However, this review pointed to the complexity of the relationships between the two types of association. A caution was thus made that it might not be a desirable practice to simply aggregate the scores; instead, it would be preferred to report the separate scores for different association relationships. We would add that all future studies that use a WAF test as a research tool, in addition to reporting separate scores, also report the factor structure of the different types of association relationships (at least their basic correlations) before a decision is made to use an aggregated score. If a very good factor structure is established, it would be even better to create a latent variable (or calculate a factor sore) to index learners' vocabulary depth rather than a simple aggregated score.

Interaction

Test performance and behaviors are the result of complex interplays between a set of variables that involve the test itself, administration, scoring, and test-takers (Bachman and Palmer 1996), which is no exception for WAF tests, as this review revealed. Researchers thus need to be aware of those complex interactions and be informed and strategic with their use of WAF tests, particularly when vocabulary depth indexed by the tests is used to predict language skills development (e.g., reading comprehension). Depending on the features of a test, how the test is administered and scored, and who the learners are, the target effect could be either overestimated or underestimated. Overall, the information in the literature about the complex interactions between variables within each level and across levels is very limited, and many issues remain to be explored in the future.

Guessing

As it is revealing in this review, a critical concern about WAF tests is guessing. Thus, without surprise, a lot of effort was taken to examine guessing (how and why it happens) and find ways, albeit not necessarily as effective as intended, to counteract it, such as selecting high frequency words, varying the distributions of associates, using semantically unrelated distractors, not letting learners know the number of associates, and using a "better" scoring method.

As reviewed earlier, *Read (1993) reported an inclination of higher-proficiency learners to guess for items and that of lower proficiency learners to skip items with target words unknown to them. Later, *Read (1998), however, found the pattern was not as straightforward as revealed in the earlier study in that some learners with obviously a high proficiency in English also skipped items showing reluctance to make

guesses. *Qian and Schedl (2004) further reported that almost all learners indicated they were reluctant to guess when encountering unfamiliar words on the DVK. *Dronjic and Helms-Park (2014), in their study on native English speakers' responses to *Qian and Schedl's (2004) DVK, computed for each participant the difference between his/her standardized hit and false alarm rates (i.e., d-scores), with a score of 0 (or near 0) indicating chance-level performance and a high absolute value indicating the sensitivity to the difference between associates and distractors. Comparisons of d-scores indicated that the participants, whether or not they knew the number of associates, were equally likely to randomly guess.

The findings highlighted above painted a more complex picture about guessing than *Read's (1993) original concern from the perspective of learner proficiency. In other words, guessing happened to some learners but not others disregarding their proficiency and test administration factors. It thus appears that guessing might be an inherent nature of WAF tests, and whether or not guessing happens might be primarily a reflection of learners' psychological characteristics, such as personality. In other words, those who are inclined to guess will guess for WAF tests no matter what the test is like and how the test is administered and scored and disregarding other learner characteristics like age and proficiency. As *Read (1998) argued, if the assessment unit of WAF tests is the item as a whole, that is, the lexical network between a target word and its associates rather than the individual words themselves (see also *Schoonen and Verhallen 2008), then any variance induced by "guessing" through drawing upon relational information in the choices may indeed be part of the network knowledge that the WAT assesses or the variance is "construct-relevant" (*Read 1998, p. 56). If these arguments hold, it would appear to be more interesting to examine how learners with what psychological characteristics tend to guess or not to guess than whether or not learners guess.

Computer corpora and lexical databases

Developers of WAF tests all seemed to primarily rely on their own expert knowledge to select target words and find associates and distractors for those words (e.g., *Horiba 2012; *Read 1993; 1998; *Schoonen and Verhallen 2008). However, as Fitzpatrick (2007) reported, (adult) native speakers (of English), despite little variability across occasions on the associates they provided through free association (this suggests that they have a stable lexical network), tended to show huge inter-individual variability. *Dronjic and Helms-Park (2014) further found that native English speakers' responses to syntagmatic association were particularly heterogeneous. It was also noted that some of the heterogeneities might be attributed to the non-canonical collocates in the choices. For example, while the target word *insufficient* in the DVK (*Qian and Schedl 2004) could be a possible modifier of the choice word *needs*, the authors' searches in several large corpora for the intended collocation *insufficient needs* did not yield any hit. Possibly because of this non-canonical combination, while some native speakers "correctly" chose *needs* as an associate for *insufficient*, others did not. This led the authors to suggest that "test developers' intuitions alone will not suffice" and "associations need to be developed with reference to large corpora of spoken and written language…with careful attention to the frequency of headwords and the use of typical rather than marginal collocations" (p. 213).

In addition to computer corpora, computer-based WordNet-like lexical databases, which were developed to simulate human beings' mental organization of words and categorize words in synonymic relationships (i.e., synsets) (Fellbaum 1998), should also be a good reference for WAT test development.[5] The different strengths of semantic connectivity between words (Steyvers and Tenenbaum 2005; for a related discussion in the L2 literature, see Wilks and Meara 2002 and Meara 2007) seem to provide an opportunity to generate paradigmatic associates and distractors for WAF tests. Thus, computer corpora and WordNet-like lexical databases provide a possibility for computer-driven, adaptive testing of L2 learners' lexical network knowledge, with consideration of word frequency, collocational canonicity, as well as semantic connectivity, which could be an interesting direction for WAF test development and research in the future.

Piloting and native speakers

L2 word association research has been heavily influenced by the L1 word association literature. It is reflected not only in the use of wordlists (or their translated equivalents) developed for free association for native speakers, such as the Kent-Rosanoff list (e.g., Jiang 2002; Meara 1978; Namei 2004), but also the use of native-speaking norms to probe into the nativelikeness of L2 learners' association (Fitzpatrick 2013). Thus, it did not seem to be a surprise that a common practice in the WAF test literature is that researchers initially used the expertise of themselves (as native speakers and/or experienced teachers of the language) to develop a test, and then pilot the test on other native speakers before it was used for measuring L2 learners (e.g., *Greidanus and Nienhus 2001; *Hellman 2011; *Read 1993; 1998; *Schoonen and Verhallen 2008). However, how specifically responses and feedback from native speakers (other than the researchers themselves) were used to adjust and refine the test was often unclear.

Based on native speakers' responses to *Qian and Schedl's (2004) DVK developed for university English L2 learners, *Dronjic and Helms-Park (2014) found that only about 60 and 40.3% of the 40 items were correctly answered by more than 90% of the participants in the constrained and unconstrained condition, respectively. Thus, "If a criterion of 90% of correct NS [Native Speaker] answers is adopted as a measure of minimal acceptable discriminative validity for an item … a large proportion of the items in this test fall short of this criterion" (p. 208). The authors thus suggested that a WAF test item should only be included for testing L2 learners if it receives 90% accuracy rate from native speakers. Using 90% accuracy rate as a reference should of course not replace the need for drawing upon computer corpora for constructing an initial version of a WAF test, as discussed earlier. In addition, we would add that the reference to native speaker norms and accuracy rate should also consider the age and conceptual development of L2 learners – if the learners are young children, such as in the case of *Schoonen and Verhallen (2008), it would seem questionable to use mature adult native speakers' responses to determine the accuracy rate for item inclusion or exclusion.

Fluency

There are arguments that fluency or the efficiency of gaining access to one's lexical knowledge should be an important dimension of (deep) word knowledge and be

considered in vocabulary assessment (e.g., Chapelle 1994; *Qian 2002; Schmitt 2014). Rather than regarding fluency as a distinct dimension parallel to other dimensions of depth, it might be better to conceptualize it as an all-encompassing ability tied with all other aspects of knowledge (see Daller et al. 2007), including form-meaning connection (i.e., the focus of vocabulary breadth tests) as well as lexical network knowledge. In line with this argument, assessment of network knowledge would need to address not only the knowledge itself, but also learners' access to this knowledge.

Although, as shown in Additional file 1: Appendix A, some studies specified a time limit for learners to complete a WAT test, "the time control" did not seem to be particularly interested in learners' efficiency of accessing their network knowledge over and beyond the knowledge itself. *Cremer and Schoonen (2013) seemed to be only WAF test study that directly touched on this issue by explicitly incorporating a time-sensitive measure. The authors made a differentiation between availability (the knowledge itself) and accessibility of semantic word knowledge, which tapped students' word association ability in relation to context-independent vs. context-dependent word meanings. The semantic knowledge itself was measured with the Dutch WAT of *Schoonen and Verhallen (2008) as a paper-and-pencil test; the accessibility of that knowledge was measured with a computerized semantic-decision task (C-WAT), in which a stimulus word was followed by two choices, one semantically related and the other contextually-related. Children were asked to indicate the semantically related word as quickly as possible with both reaction time and accuracy level recorded. C-WAT response latency was found to explain a unique and significant, albeit small, proportion of variance in reading comprehension over and above children's Dutch WAT scores (and word decoding fluency), which indicated that accessibility is distinct from availability of semantic word knowledge.

*Cremer and Schoonen's (2013) findings suggested that a lack of attention to time or speed factor (e.g., response latency or time-constrained test administration) would constrain the power of WAF tests in discriminating learners' network knowledge. Such an issue also has strong implications for research that examines the effect of depth knowledge, as measured by a WAF test, on language comprehension and production. For example, focusing on WAF test accuracy without considering the efficiency of accessing word relationships might underestimate the predictive effect of depth knowledge on reading comprehension. The fluency issue, of course, does not pertain to WAF tests or network knowledge alone. As Schmitt (2014) and Beglar and Nation (2014) pointed out, assessing vocabulary knowledge with consideration of fluency has received little attention in the L2 assessment literature, and we still know little about the implications of fluency for vocabulary knowledge development and assessment.

Diverse validation evidence

As discussed in *Read (49), both qualitative and quantitative evidence are important for validating a WAF test. However, as this review revealed (see also Additional file 1: Appendix B), most validation evidence for WAF tests was of a quantitative nature, such as discriminability, reliability (Cronbach's α or Kuder-Richardson coefficients), factor structure, identification of misfitting items and persons through Rasch analysis, as well as concurrent or predictive validity with another vocabulary test, interview scores,

reading comprehension or general proficiency. Only in a few studies was qualitative evidence collected through learners' verbal reports to obtain a nuanced knowledge about their response patterns in relation to test features, test administration, and test-taker characteristics (*Read 1993; 1998; *Schmitt et al. 2011). Qualitative validation evidence that probes into learners' detailed thought processes certainly warrants more attention in future research.

Bilingual lexical network

In the review sections, we discussed from two perspectives why language should be a critical factor for consideration in using WAF tests to measure vocabulary depth, including the modality of administering the tests and L1-L2 linguistic distance. Both perspectives were primarily concerned with between-group variabilities rather than the relationships between learners' L1 and L2 lexical network and their implications for L2 vocabulary depth assessment.

The bilingual mental lexicon literature suggests that L1 and L2 lexicon are closely related with respect to not only form-meaning connections but also word relationships. Some studies based on learners' word productions, such as free association or language use (e.g., writing), suggested that L2 lexical network shows significant influence from learners' native language (e.g., Fitzpatrick and Izura 2011; Verspoor 2008). Others based on comparisons of learners' decision-making on L1-L2 congruent and non-congruent collocations also indicated close links between L1 and L2 mental lexicon (e.g., Wolter and Gyllstad 2011).

Little research, however, has been conducted on possible implications of those research findings about bilingual mental lexicon for assessing learners' L2 network knowledge with WAF tests. One issue of interest might be whether the L2 network knowledge revealed in a WAF test is primarily a reflection of the knowledge in learners' native language (i.e., L1 in nature) or it is distinct from L1. *Henriksen (2008) found the performance of Danish-speaking EFL learners at all three grade levels (i.e., Grades 7, 10, and 13) on an English word connection task was significantly less strong than that on a translated version of the task in Danish (i.e., L1 task used the same format, target words, and choices of the L2 task). More importantly, it was found that the correlations between the English and Danish tasks were all significant but small to medium in size (rs = .623, .491, and .685 for the three grades, respectively). These correlations suggested that there was substantial difference, despite significant overlap, between the two languages; yet it remains unknown how much of the variance indeed pertained to a distinction between L1 and L2 network knowledge or whether the variance was primarily due to other factors (e.g., different language forms, as both tasks were administered in print). It is also unclear whether the variance may change as a function of the properties of the word connection task (e.g., word frequency) and L2 proficiency. Hypothetically, for very frequent words and learners with advanced L2 proficiency, there might be a high convergence between L1 and L2 network.

Conclusion

This paper reviewed the general practices, key issues, and research findings that pertain to WAF tests as a measure of L2 learners' vocabulary depth with a focus on network

knowledge. The review was presented in four major areas that included design features of WAF tests, conditions for test administration, scoring methods, and test-taker characteristics. In each area, a set of variables was identified and described with relevant research findings also presented and discussed. Around eight topics, the General Discussion section provided some suggestions and directions for the development of WAF tests and the use of them as research tools in the future.

It is hoped that this paper has achieved its purpose to help researchers make better decisions when they develop and validate a WAF test and/or use that test or an existing test for different research purposes. In particular, we hope that researchers who use a WAF test as a research tool could be better aware that the results generated by the test, such as how much depth knowledge is functional in language skills development (e.g., reading comprehension), may vary depending what specific design the test has, how it is administered, how it is scored, and who the learners are.

Endnotes

[1]Qian and Schedl (2004) scored the 40-item DVK with Read's (1993, 1998) method, which awarded a point for an associate selected without considering distractor rejection or selection. Consequently, the maximum score was 160 (with four associates to be selected for each of the 40 items), but the DVK-meaning (i.e., paradigmatic association) section had a maximum score of 79, whereas the DVK-collocation (i.e., syntagmatic association) section had a maximum score of 81. The authors addressed this imbalance by using the proportion of associates selected for each association type. Another way to address unequal scores between different association relationships in a WAF test following Read's (1998) format might be to use a scoring method that gives credit for both associate selection and distractor rejection, which was used in Greidanus and Nienhus (2001) as a way to counteract guessing (see Scoring WAT Tests later in this paper). This scoring method would make paradigmatic and syntagmatic relationships in WAF tests like Read's (1998) have an equal range of scores (i.e., 0–4), disregarding how associates (for different association relationships) are distributed.

[2]Greidanus et al. (2005) administered two parallel French DWK tests. In one test, all items had a fixed number of associates (i.e., 3), and learners were informed on the number; in the other test, the number of associates varied, and for this test, learners were informed that the number of associates varied but without knowing which item had which number of associates. The two tests appeared to contrast in whether the number of associates was fixed. However, we consider the design was essentially about whether or not learners actually knew the number of associates (i.e., an issue of administration condition). In other words, there is no fundamental difference between not being informed on the number of associates (in the case of the number being fixed; e.g., Read 1993, 1998) and Greidanus et al.'s (2005) second test for which test-takers knew the number variation but did not know the specific number of associates for each item. This is because in both cases, learners would need to actively engage their network knowledge without being able to assume that selecting a certain number of choices would necessarily lead to the answer. Thus, any performance difference between Greidanus et al.'s (2005) two tests, or the lack thereof, may well reflect the conditions in which the tests were administered rather than whether or not the number of associates was fixed. To this end, Greidanus et al.'s (2005) findings on the two tests were reviewed

later under "Informed vs. Uninformed" in the "Administering WAF Tests" section of this paper.

[3]Heritage learners of Chinese typically have significant oral language experiences in the language (hence, a presumably significant network knowledge developed out of those experiences) but are known to have challenges with respect to Chinese orthography and characters.

[4]Schoonen and Verhallen (2008) and Greidanus et al. (2005), while both including learners with different L1 backgrounds, did not seem to be interested in how L1-L2 linguistic distance might matter to their WAF tests. See Appendix B.

[5]The website of the Global WordNet Organization (http://globalwordnet.org/wordnets-in-the-world/) has a list of all the WordNet-like lexical databases in diverse languages in the world.

Additional file

Additional file 1: Lists of studies on WAF tests and key research issues. (XLSX 51 kb)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
DZ and KK contributed to the conceptualization of the review. DZ searched the literature for studies to be included in the review. DZ and KK drafted the manuscript. Both authors read and approved the final manuscript.

Author details
[1]Department of Teacher Education, Michigan State University, 620 Farm Lane, East Lansing, MI 48824, USA.
[2]Department of Modern Languages, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, USA.

Received: 4 November 2016 Accepted: 22 January 2017
Published online: 09 March 2017

References

*** indicating articles that appear in Appendix A and reported on the development of a WAF test or used a WAF test as a research tool**

Aitchison, J. (1994). *Words in the mind: An introduction to the mental lexicon* (2nd ed.). Oxford: Blackwell.

*Akbarian, I. (2010). The relationship between vocabulary size and depth for ESP/EAP learners. *System, 38*, 391–401.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). Newark: International Reading Association.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

*Batty, A. (2012). Identifying dimensions of vocabulary knowledge in the word associates test. *Vocabulary Learning and Instruction, 1*, 70–77.

Beglar, D., & Nation, P. (2014). Assessing vocabulary. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 172–184). UK: Wiley.

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research, 10*, 157–187.

*Cremer, M., & Schoonen, R. (2013). The role of accessibility of semantic word knowledge in monolingual and bilingual fifth-grade reading. *Applied Psycholinguistics, 34*, 1195–1217.

Daller, H., Milton, J., & Treffers-Daller, J. (Eds.). (2007). *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press.

*Dronjic, V., & Helms-Park, R. (2014). Fixed-choice word-association tasks as second-language lexical tests: What native-speaker performance reveals about their potential weaknesses. *Applied Psycholinguistics, 35*, 193–221.

*Ehsanzadeh, S. J. (2012). Depth versus breadth of lexical repertoire: Assessing their roles in EFL students' incidental vocabulary acquisition. *TESL Canada Journal, 29*(2), 24–41.

Entwisle, D. R. (1966). *Word associations of young children*. Baltimore: John Hopkins Press.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge: MIT Press.

Fitzpatrick, T. (2007). Word association patterns: Unpacking the assumptions. *International Journal of Applied Linguistics, 17*, 319–31.

Fitzpatrick, T. (2013). Word associations. In C. Chapelle (Ed.), *Encyclopedia of applied linguistics* (pp. 6193–6199). Oxford: Wiley-Blackwell.

Fitzpatrick, T., & Izura, C. (2011). Word association in L1 and L2: An exploratory study of response types, response times, and interlingual mediation. *Studies in Second Language Acquisition, 33*, 373–398.

Fitzpatrick, T., Playfoot, D., Wray, A., & Wright, M. J. (2015). Establishing the reliability of word association data for investigating individual and group differences. *Applied Linguistics, 36*, 23–50.
*Greidanus, T., & Nienhus, 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, 84*, 567–577.
*Greidanus, T., Bogaards, P., van der Linden, E., Nienhuis, L., & de Wolf, T. (2004). The construction and validation of a deep word knowledge test for advanced learners of French. In P. Bogaards & B. Laufer-Dvorkin (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 191–208). Amsterdam: John Benjamins.
*Greidanus, T., Beks, B., & 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*, 221–233.
*Guo, Y., & Roehrig, A. D. (2011). Roles of general versus second language (L2) knowledge in L2 reading comprehension. *Reading in a Foreign Language, 23*, 42–64.
*Hellman, A. B. (2011). Vocabulary size and depth of word knowledge in adult-onset second language acquisition. *International Journal of Applied Linguistics, 21*, 162–182.
Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition, 21*, 303–317.
*Henriksen, B. (2008). Declarative lexical knowledge. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), *Vocabulary and writing in a first and second language: Processes and development* (pp. 22–66). New York: Palgrave Macmillan.
Higginbotham, G. M. (2010). Individual learner profiles from word association tests: The effect of word frequency. *System, 38*, 379–90.
*Horiba, Y. (2012). Word knowledge and its relationship to text comprehension: A comparative study of Chinese- and Korean-speaking L2 learners and L1 speakers of Japanese. *The Modern Languages Journal, 96*, 108–121.
Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. New York: Routledge.
Jiang, S. (2002). Chinese word associations for English speaking learners of Chinese as a Second Language. *Journal of the Chinese Language Teachers Association, 37*(3), 55–70.
Koda, K. (2005). *Insights into second language reading*. New York: Cambridge University Press.
Meara, P. (1978). Learners' word associations in French. *Interlanguage Studies Bulletin, 3*, 192–211.
Meara, P. (1983). Word associations in a foreign language. *Nottingham Linguistics Circular, 11*, 29–38.
Meara, P. (2007). Simulating word associations in an L2: Approaches to lexical organisation. *International Journal of English Studies, 7*(2), 1–20.
Meara, P. (2009). *Connected words*. Amsterdam: John Benjamins.
Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.
Milton, J., & Fitzpatrick, T. (Eds.). (2014). *Dimensions of vocabulary knowledge*. Basingstroke: Palgrave Macmillan.
Nagy, W. E., & Scott, J. (2000). Vocabulary processes. In M. Kamil, P. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. III, pp. 269–284). Mahwah: Lawrence Erlbaum Associates.
Namei, S. (2004). Bilingual lexical development: A Persian-Swedish word association study. *International Journal of Applied Linguistics, 14*, 363–88.
*Nassaji, H. (2004). The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy use and success. *The Canadian Modern Language Review, 62*, 107–135.
Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.
Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
Nelson, K. (1977). The syntagmatic-paradigmatic shift revisited: A review of research and theory. *Psychological Bulletin, 84*, 93–116.
Nissen, H. B., & Henriksen, B. (2006). Word class influence on word association test results. *International Journal of Applied Linguistics, 16*, 389–408.
*Nurweni, A., & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes, 18*, 161–175.
Odlin, T. (1989). *Language transfer*. Cambridge: Cambridge University Press.
*Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review, 56*, 282–307.
*Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*, 513–536.
*Qian, D. D., & 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. (1998). Validating a test to measure depth of vocabulary knowledge. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60). Mahwah: Lawrence Erlbaum Associates.
Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.
Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined. In B. Laufer & P. Bogaards (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 209–227). Amsterdam: John Benjamins.
Read, J. (2014). Second language vocabulary assessment. *Language Teaching, 46*, 41–52.
Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10*, 77–89.
Schmitt, N. (1998). Quantifying word association responses: What is nativelike? *System, 26*, 389–401.
Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64*, 913–951.
Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the Vocabulary Levels Test. *Language Testing, 18*, 55–88.
*Schmitt, N., Ng, J. W. C., & Garras, J. (2011). The Word Associates Formats: Validation evidence. *Language Testing, 28*, 105–126.
*Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing, 25*, 211–236.
*Shin, H. W. (2015). Psychometric properties of word association test with regard to adolescent EFL learners. *Vocabulary Learning and Instruction, 4*, 9–15.
Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29*, 41–78.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*, 217–234.
Verspoor, M. H. (2008). What bilingual word associations can tell us. In F. Boers & S. Lindstromberg (Eds.), *Cognitive linguistic approaches to teaching vocabulary and phraseology* (pp. 261–290). Berlin: De Gruyter.
Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth vs. breadth. *Canadian Modern Language Review, 53*, 13–39.
Wilks, C., & Meara, P. (2002). Untangling word webs: Graph theory and the notion of density in second language word association networks. *Second Language Research, 18*, 303–324.
Wolter, B. (2001). Comparing the L1 and L2 mental lexicon. *Studies in Second Language Acquisition, 23*, 41–69.
Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics, 32*, 430–449.
Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3*, 215–229.
Zareva, A. (2007). Structure of the second language mental lexicon: how does it compare to native speakers' lexical organization? *Second Language Research, 23*, 123–153.
Zareva, A., & Wolter, B. (2012). The 'promise' of three methods of word association analysis to L2 lexical research. *Second Language Research, 28*, 41–67.
*Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *Modern Language Journal, 96*, 558–575.