CrossMark

Applied Psycholinguistics **35** (2014), 193–221
doi:10.1017/S014271641200029X

Fixed-choice word-association tasks as second-language lexical tests: What native-speaker performance reveals about their potential weaknesses

VEDRAN DRONJIC and RENA HELMS-PARK
University of Toronto

Received: February 26, 2011 Accepted for publication: October 23, 2011

ADDRESS FOR CORRESPONDENCE
Vedran Dronjic, Centre for Educational Research on Languages and Literacies, OISE/University of Toronto, 252 Bloor Street West, Toronto, ON M5S 1V6, Canada. E-mail: vedran.dronjic@utoronto.ca

ABSTRACT
Qian and Schedl's Depth of Vocabulary Knowledge Test was administered to 31 native-speaker undergraduates under an "unconstrained" condition, in which the number of responses to headwords was unfixed, whereas a corresponding group ($n = 36$) completed the test under the original "constrained" condition. Results revealed lower accuracy in the unconstrained condition and in paradigmatic versus syntagmatic responses. Native speakers failed to reach the 90% criterion on most unconstrained and many constrained items. Although certain modifications could improve such a test (e.g., eliminating psycholinguistically anomalous headwords, such as adjectives, or presenting responses to headwords discontinuously), two intransigent problems impede test validity. First, collocates in the mental lexicon differ in tightness and vary across dialects, sociolects, and age groups. Second, it is more serious that second-language Depth of Vocabulary Knowledge Tests are likely spot checks of metalinguistic knowledge rather than depth tests that reflect what learners would actually produce in spontaneous utterances.

The last two decades have seen a renewal of interest in vocabulary and a surge in the number of studies on second-language (L2) lexical acquisition. Despite this progress, L2 lexical studies frequently remain uninformed by current psycholinguistic research. Nowhere is this more evident than in the case of lexical testing and the linguistic and psycholinguistic frameworks within which lexical testing is conducted. Standardized, valid, and reliable vocabulary measures are still scarce, particularly so when one considers languages spoken outside the few countries able to sustain lexical testing research. As would be expected, the situation is the best in the case of English, for which there are now quite a few lexical tests

© Cambridge University Press 2012 0142-7164/12 $15.00

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

available, many of which have undergone numerous cycles of empirical testing and improvement. Nonetheless, the existence of disparate views of what it means to "know" a word as well as the lack of attention to the development and use of L2 lexical knowledge has led to depth of knowledge tests that are incompatible with the findings of psycholinguistic research in noticeable ways and/or that have not been validated by native-speaker (NS) performance on these tests. This paper reports our findings on the performance of first-language (L1) English undergraduates on the Qian and Schedl (2004) test, administered to arrive at baseline data for the purposes of comparing the performance of speakers of Romanian and Vietnamese on lexical size and depth tests.[1] Data were collected from NSs under two conditions: instructions that specified the number of correct answers per item (in terms of both correct hits and correct misses) and, in order to minimize "guessability," instructions that had no such specifications.

DEPTH AND BREADTH OF VOCABULARY KNOWLEDGE

In keeping with the view that vocabulary knowledge is not a unidimensional construct, one distinction often made is between receptive and productive vocabulary (Nation, 2001). Another useful distinction is between the number of words known, or breadth of knowledge, and how well these words are known, or depth of knowledge (Nation, 2001; Read, 2004). This distinction has been widely recognized in vocabulary research and testing, but not with complete terminological clarity, particularly in the case of depth. Due to the different operationalizations, Read (2004), along the lines of Cronbach (1942), suggests that it might be time to stop treating depth as a single trait and to break it down into smaller (albeit overlapping) components: (a) *precision of meaning* (i.e., whether the knowledge of a word's meaning is vague or specific); (b) *the comprehensiveness of word knowledge* (e.g., polysemy, word syntax, likely collocates); and (c) *network knowledge* (i.e., how the meaning of a word connects with the meanings of related words). Read admits that this breakdown itself is not without problems. For instance, even in their L1 people do not know all words with the same level of precision. Consider, for instance, the average English speaker's knowledge of classical concepts such as *triangle* versus probabilistic ones such as *furniture*. (For other discussions of the construct of word knowledge, see Chapelle, 1998; Henriksen, 2008; Jiang, 2000; Nation, 1990, 2001; and Qian, 2002.)

Meara (1996) hypothesized that breadth might be the only relevant trait of lexical knowledge for small vocabularies and that dimensions related to depth may become important only once a certain size threshold has been achieved. Meara (1996) and Zareva and colleagues (Zareva, 2007; Zareva, Schwanenflugel, & Nikolova, 2005) set this threshold at approximately 5,000–6,000 and 9,000 words, respectively. The notion that breadth is a more basic dimension of lexical proficiency than depth is in line with the predictions of a class of models of L2 vocabulary acquisition based on Kroll and Stewart's (1994) revised hierarchical model (see also Heredia & Brown, 2006; Jiang, 2000; but, for a differing perspective, see also Kroll & Dijkstra, 2002). These models have in common the claim that L2 vocabulary acquisition starts with the establishment of a direct

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

link between a newly acquired word form and what is perceived as its translation equivalent in the L1 and then gradually progresses toward deeper lexical knowledge.

The existence of a correlation between breadth and depth of vocabulary knowledge in intermediate to advanced second-language users has been demonstrated empirically (Greidanus, Bogaards, van der Linden, Nienhuis, & de Wolf, 2004; Wesche & Paribakht, 1996; see also Qian, 1999, 2002; Qian & Schedl, 2004). However, despite this correlation, there are studies which suggest that these two dimensions still make separate contributions at intermediate to advanced levels to performance on measures of proficiency such as Test of English as a Foreign Language (TOEFL) reading scores (Qian, 1999, 2002; Qian & Schedl, 2004) and cloze test scores (Schoonen & Verhallen, 2008). Furthermore, Wesche and Paribakht (1996) found only a moderate positive correlation ($r = .53$) between scores on their test of lexical depth (Vocabulary Knowledge Scale) and the Eurocentres Vocabulary Size Test, a breadth test (Meara & Jones, 1990).

BREADTH AND DEPTH TESTS AND THE DEPTH OF VOCABULARY KNOWLEDGE (DVK) MEASURE

Depth of lexical knowledge, no matter how it is operationalized, is notoriously difficult to test in a feasible way. As Read (2000) pointed out, the testing of depth inevitably involves a trade-off between the number of targeted items and the number of dimensions of depth that can be assessed in a practicable fashion. It is not particularly surprising, therefore, that most lexical tests to date, L1 or L2, have focused on the measurement of breadth. Some well-known examples include the Peabody Picture Vocabulary Test (Dunn, 1997), used in L1 and L2 research with children; the Vocabulary Levels Test (Nation, 1990; Schmitt, Schmitt, & Clapham, 2001); and the Eurocentres Vocabulary Size Test (Meara & Jones, 1990).

Most available tests of lexical depth adopt either the self-report scale or word-association format. The first type involves responses on a graduated scale, as used by Dale (1965) with English-speaking fifth graders, Zareva (2007) with L1 and L2 English speakers, and Wesche and Paribakht (1996) for L2 speakers. In the last case, for example, participants identify their degree of acquaintance with the test item on a 5-point Vocabulary Knowledge Scale. The first two levels elicit receptive knowledge, and the three highest levels require producing a definition/translation or an example sentence.

Another kind of depth test adopts various forms of the fixed-choice word-association format as exemplified by the Euralex French Tests, developed by Meara (1992). In these tests, test takers are required to identify pairs of words (60 in all) as being connected or not (ratio between items = 2 : 1). Connections can be semantic (synonymy, hyponymy, and antonymy) or syntagmatic (fixed expressions and culture-specific associations).

In its various incarnations, the Word Associates Test (WAT) format is perhaps the most widely used and investigated test of depth of L2 lexical knowledge, developed by Read to test the depth of students' knowledge of the University Word List (Nation, 1990; Xue & Nation, 1984). The original WAT, aiming to address some of the methodological weaknesses of word-association tasks noted in

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

Meara (1983), examined L2 users' ability to identify syntagmatic (collocational), paradigmatic (involving synonymy), and analytic (words used in dictionary definitions) knowledge of L2 lexical items (for a summary of the genesis of the WAT and an account of the various improvements made to the format, see Read, 2000, pp. 180–187). The WAT format has been used by a number of researchers both with L1 and L2 speakers of languages other than English. Variations have included word frequency, number of choices per item, type of distractors, and number of correct responses, whether fixed or varied (Greidanus, Beks, & Wakely, 2006; Greidanus et al., 2004; Henriksen, 2008; and Schoonen & Verhallen, 2008).

A series of modifications of the WAT by Qian (1998, 1999, 2002) and Qian and Schedl (2004), conducted as part of the TOEFL 2000 project, yielded a version of this format called the DVK measure. The test consists of 40 items, each an adjective selected from a pool of pre-1995 TOEFL vocabulary items. For each item, eight potential answers are provided. They are grouped in two four-item boxes, the one on the left containing potential paradigmatic associates (synonyms or near synonyms) and the one on the right containing potential syntagmatic associates (nouns that might be modified by the target adjective). The number of correct choices in each of the four-item boxes can be one, two, three, or four, so that there are always four correct answers and four distractors per item. This is explicitly stated in the test instructions.

Qian and Schedl (2004) report high reliability for the overall test ($\alpha = 0.91$) and for its components: the paradigmatic ($\alpha = 0.83$) and syntagmatic ($\alpha = 0.86$) parts of the test. DVK scores were highly correlated ($r = .84$) with TOEFL vocabulary scores. Moreover, both sets of scores were highly correlated with TOEFL reading scores ($r = .74$ for the DVK, $r = .75$ for the TOEFL vocabulary test, with the difference between the two correlations not reaching statistical significance).

The majority of the participants in Qian and Schedl's (2004) study reported having little difficulty with the test format and instructions. The authors reported that all the interviewed participants but one agreed that "it was rather difficult to guess" the correct answers (p. 44) because the number of correct answers in each box varied from item to item (e.g., a headword could have three paradigmatic responses and one syntagmatic one, whereas another headword could have two paradigmatic and two syntagmatic responses). Qian and Schedl conclude that "the improved DVK format has considerably reduced the potential for successful guessing" (p. 44). Note that their conclusion runs counter to Read's report on the guessability of items in the WAT (Read, 1993, 2000).

Our review of the literature reveals that the existing word-association tests for L2 speakers of English have mostly bypassed several important issues without which these psychometric instruments can be neither understood nor optimized (see also Meara, 2009). One such issue centers on the organization of the mental lexicon and the links therein (e.g., tight and loose semantic and collocational ties within networks) in both stable L1 lexicons and developing L2 ones. A second issue is one frequently invoked in psycholinguistic research but, thus far at least, rarely investigated in the second-language research on testing: the nature of lexical activation processes such as the role of items competing for selection and the inhibition of irrelevant candidates. In light of the realities of lexical activation, some of the features of the WAT/DVK may actually make it harder for test takers

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

to access appropriate answers to test items in semantic memory, for example, selecting nouns for adjectival headwords instead of vice versa (Schriefers, 1993), placing orthographically and/or phonologically similar headwords continguously (e.g., *perceptual* and *perpetual*) or using polysemous headwords that trip up the learners by activating nontargeted meanings (e.g., a core nominal meaning, as in *mountain peak*, rather than *peak* as an adjective, as in *peak time*).

In addition, it remains unclear in what ways performance might vary when the constraint on the number of correct responses is lifted (as in the depth tests of French in Greidanus et al., 2006; and Greidanus & Nienhuis, 2001). Finally, how NSs might fare on this test remains unknown. Although those working on non-English depth tests have used NS data for comparative purposes (e.g., Greidanus et al., 2006; and Henriksen, 2008), the DVK has not been previously validated with NSs of English.

When data are analyzed at the level of group, NSs usually provide relatively consistent responses on word-association tasks (Fitzpatrick, 2007; Meara, 1983; but see the Discussion for a crucial caveat). We strongly believe that L2 tests where NSs (whether monolingual or bilingual) struggle with various items or end up with less than 90% consistency in responses should not be used to evaluate L2 speakers. (However, this does not exclude the possibility of using such tests to simply compare and contrast NS and L2 speakers for exploratory, nonprescriptive, or nonevaluative purposes.) With these issues in mind, we set out to answer the following research questions regarding the performance of native English speakers on Qian and Schedl's (2004) version of the WAT/DVK:

1. Do NSs perform differently when there is no constraint on the number of correct responses in the instructions and guessing is minimized?
2. What is the relationship between NSs' performance on the paradigmatic and syntagmatic segments of the test?
3. Do NSs score at least 90% on the DVK and its subsections?

As noted above, the study reported here is part of a larger research project investigating bilinguals' and monolinguals' performance on various L2 tests of English lexical breadth and depth. Some of these tests were newly designed by our team, whereas others, such as the one under scrutiny here, had been published by other researchers

METHOD

Participants

The participants in the present study (see Table 1) were undergraduate students whose L1 was English in terms of both chronology and language dominance, studying on the campus of a major university in Canada. In order to reflect the range of disciplines on the campus, participants were recruited through notices in the study areas of the main library. The screening criteria required them to be speakers of English and to be enrolled in an undergraduate program at the university.

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

Table 1. *Participant data*

	Constrained ($n = 36$)		Unconstrained ($n = 31$)	
	Frequency	Percentage	Frequency	Percentage
Sex				
Male	17	47.22	18	58.07
Female	19	52.77	13	41.94
Age range				
18–29	36	100	30	96.77
50+	0	0	1	3.23
Major/specialist				
Life sciences	9	25	6	19.35
Psychology	8	22.22	4	12.90
Other sciences	3	8.33	0	0
Management/economics	5	13.89	6	19.35
Social sciences	2	5.56	8	25.81
Humanities	5	13.89	7	22.58
Program year				
1	7	19.44	11	35.48
2	4	11.11	3	9.68
3	9	25	9	29.03
4	13	36.11	7	22.58
5	3	8.33	1	3.23
L2				
None	24	66.67	20	64.52
French	4	11.11	3	9.68
Spanish	2	5.56	0	0
Italian	1	2.78	0	0
Modern Greek	0	0	1	3.23
Bengali	1	2.78	0	0
Urdu	0	0	1	3.23
Sinhalese	0	0	1	3.23
Tamil	2	5.56	2	6.45
Cantonese	2	5.56	0	0
Japanese	0	0	1	3.23
Arabic	0	0	1	3.23
Hebrew	0	0	1	3.23

	R/W/L/S[a]	L2 Status	L2	R/W/L/S[a]	L2 Status
L2					
French	4/4/4/3.5	Official	French	3/3/4/2	Official
French	2/1/2/1	Official	French	3/3/4/2	Official
French	4/3/4/4	Official	French	1/1/1/1	Official
French	2/2/3/2	Official	Greek	4/4/4/4	Heritage
Spanish	3/3/3/2	Foreign	Urdu	1/1/3/1	Heritage
Spanish	3/3/3/3	Foreign	Sinhalese	1/1/5/3	Heritage
Italian	3/2/4/3	Heritage	Tamil	1/1/4/4	Heritage
Bengali	0/0/2/1	Heritage	Tamil	1/1/4/2	Heritage
Tamil	1/1/1/1	Heritage	Japanese	1/2/3/3	Foreign
Tamil	2/1/4/4	Heritage	Arabic	3/3/4/3	Heritage
Cantonese	2/1/4/4	Heritage	Hebrew	3/3/4/3	?
Cantonese	0/0/2/2	Heritage	—	—	—

Note: Official is French as a second language (L2), learned in the Canadian school system.
[a]Proficiency self-ratings for reading (R), writing (W), listening (L), and speaking (S).

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

The group that completed the test with the constrained number of responses (the "constrained group") consisted of 36 NSs (17 males, 19 females). Independent of the constrained group, the "unconstrained group" consisted of 31 NSs (18 males, 13 females). All of the students belonged to the 18–29 age range with the exception of one student in the 50-plus range in the unconstrained group. Careful examination of this individual's test data revealed that she performed very close to the average for her group, and was, therefore, behaviorally, a typical member of the unconstrained group. In the constrained and unconstrained groups, there were 12 and 11, respectively, who reported speaking a second language. The most frequently claimed second language was French, as is to be expected in Ontario, with a smattering of heritage languages such as Tamil, Sinhalese, Urdu, and Arabic, reflecting the demographic profile of the larger community in which this campus is embedded. We suspect that there might have been others with a minimal knowledge of a language such as French who did not report this in the questionnaire.

In a city such as Toronto, a large urban center that is the destination of immigrants from every corner of the globe and where French is routinely taught in the public school system, "pure monolinguals" of the kind that might have existed in previous decades are a conspicuous minority and generally belong to a very narrow social stratum and age group. Therefore, for the purposes of this study, it was decided that the NS groups would include those who might have a second language but not those whose second language had ended up being stronger than their first (here English). In reality, there were no participants who had to be excluded for this reason.

Instruments

Questionnaire. The questions (see Appendix A) were used for the purposes of conducting a guided interview with participants. The interviews were run by three research assistants who were in their final year of a psycholinguistics specialist program at the same university and who had a strong background in linguistics, language acquisition, and bilingualism. Participants who had undergone the initial screening were asked to provide details about their language background, discipline, program year, and age range. If participants had a second language, they needed to specify their level of proficiency in each of the four skills (reading, writing, listening, and speaking) on a 5-point Likert scale. The research assistants were instructed to ask additional questions about a participant's language background if they felt unsure of a participant's language status. Conducting the screening in this fashion helped to confirm that English was a speaker's dominant language (e.g., based on an informal gauging of his or her communication skills).

Constrained test. This test was identical in content to Qian and Schedl's (2004) test, with minor differences in format (circling answers rather than blackening corresponding letters from A to H; for Qian and Schedl's entire test, see pp. 50–52 of their paper). The 40 headwords in this test are bare adjectives, selected from vocabulary in a pre-1995 TOEFL test. Participants were instructed to choose four correct answers out of a total of eight choices for each headword. The four

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

correct items could consist of between one and four synonyms (in the column on the left) and between one and four collocates (in the column on the right). For example, the adjectival headword "celebrated" has three synonyms ("renowned," "well known," and "famous") and only one collocate ("play"), amounting to a total of four correct answers. The remaining four are distractors.

Celebrated
(A) renowned (B) festive (C) well known (D) famous
(E) persuasion (F) recognition (G) understanding (H) play

See Appendix B for the instructions given to participants in the constrained group.

Unconstrained test. The only difference between the constrained and unconstrained versions of the test was in the instructions, which in the case of the latter stated that participants could pick one to eight correct answers per item. Note that, because the test was identical to the constrained one, in theory only four hits (correct answers) per headword were supposed to be correct. (See Appendix C for the instructions given to participants in the unconstrained group.)

Procedure

Following the initial screening, participants filled out the questionnaire. Those confirmed as NSs were asked to complete a vocabulary size task (for validating a lexical test that was part of the larger research project in which the current study is embedded) and then randomly assigned the constrained or unconstrained version of the DVK test.

Although the tasks were unpaced, most participants completed the constrained and unconstrained versions of the DVK in around 20 and 25 min, respectively. The DVK data were inputted manually into a spreadsheet, with each of the eight possible answers to a headword prompt in its own column. Each participant's responses were coded as one of the following: (a) hit, (b) correct rejection, (c) false alarm, or (d) miss. For example, in the case of "celebrated," circling "famous" as a synonym would be a hit and *not* circling "festive" would be a correct rejection; in contrast, circling "festive" would be a false positive (incorrect hit) and *not* circling "famous" would be a miss.

RESULTS

The first research question asked how NS judgments differed when there was no constraint on the number of correct responses versus a constraint of four correct answers per headword. The second question asked if NSs' performance differed on the paradigmatic and syntagmatic segments of the test.

To answer these questions, each participant's score was calculated out of a maximum of 160 because there were 40 headwords and each of the headwords had four correct associates and four distractors. To ensure comparability with previous WAT/DVK research, we report overall scores here. (The problems with reporting

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

Table 2. *Ranges, minimum, and maximum scores on the Depth of Vocabulary Knowledge measure for the constrained and unconstrained groups*

	Overall Score		Paradigmatic		Syntagmatic	
	Range	Min.–Max.	Range	Min.–Max.	Range	Min.–Max.
Original constr.	43	113–156	24	55–79	43	37–80
Original unconstr.	98	50–148	49	27–76	67	10–77
Revised constr.	88	224–312	43	113–156	51	106–157
Revised unconstr.	116	180–296	68	85–152	63	85–148

Note: Min.–Max., minimum–maximum score.

overall DVK scores will be discussed later.) The maximum possible score on the paradigmatic subtest was 79, whereas the maximum on the syntagmatic subtest was 81. As mentioned above, participants in the unconstrained group were told that there were between 1 and 8 possible correct answers per item. This meant that, with a scoring method where only (correct) hits were awarded a point, a participant could, at least in theory, circle all answers and obtain a maximum score. To safeguard against this potential test-taking strategy, we computed an additional set of scores wherein each hit and each correct rejection was awarded a point. Note that this diverges from Qian and Schedl's scoring method. Thus, a participant could achieve a maximum overall score of 320, with the maximum possible paradigmatic score being 160 (79 hits and 81 correct rejections) and the maximum syntagmatic score being 160 (81 hits and 79 correct rejections). Where necessary, we will refer to the latter set of scores as "revised scores," whereas we will reserve the term "original scores" for the former. The ranges as well as minimum and maximum raw original scores and revised scores achieved on the test by the constrained and unconstrained groups are summarized in Table 2.

The unconstrained condition elicited more variability in the overall scores as well as on each of the subtests. In addition, the syntagmatic subtest yielded more variability in the scores compared to the paradigmatic subtest. In other words, participants' performance was more homogeneous when they knew the number of possible correct answers per headword and when they had to match words based on meaning rather than judge collocability of the headword.

Table 3 summarizes the means and standard deviations of the overall original and revised scores and the original and revised scores achieved on the two subtests. The means are reported as both raw scores and percentages.[2] (See Table 4 and Table 5.)

As can be seen from Table 3, participants in the constrained condition achieved consistently higher scores than those in the unconstrained condition. An independent-samples t test confirmed the difference in the overall original scores as significant, t (41.64) = 3.75, p = .001. The effect size d of 0.94 was large. Likewise, the difference in overall revised scores was also significant between the conditions, t (65) = 3.30, p = .002, also with a large effect size (d = 0.83). In other

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

Table 3. *Means and standard deviations of the original and revised Depth of Vocabulary Knowledge measure scores for the constrained and unconstrained groups*

	Overall Score		Paradigmatic		Syntagmatic	
	M (*SD*)	%	*M* (*SD*)	%	*M* (*SD*)	%
			Original			
Constr.	139.50 (12.01)	87.19	71.25 (6.04)	90.19	68.25 (9.94)	84.26
Unconstr.	121.03 (25.07)	75.58	62.81 (10.50)	79.51	58.23 (16.69)	71.89
			Revised			
Constr.	280.36 (22.16)	87.61	143.31 (10.53)	89.57	137.06 (12.58)	85.66
Unconstr.	259.81 (28.82)	81.19*	135.61 (15.47)	84.76*	124.19 (15.19)	77.62**

Note: The means are reported as raw scores and percentages for three comparisons. The α level is corrected to 0.017 (Bonferroni).
*Significantly different from the corresponding original score at $p = .001$.
**Significantly different from the corresponding original score at $p = .014$.

Table 4. *Ranges, minimum, and maximum scores on the Depth of Vocabulary Knowledge measure for the constrained (n = 24) and unconstrained (n = 20) groups*

	Overall Score		Paradigmatic		Syntagmatic	
	Range	Min.–Max.	Range	Min.–Max.	Range	Min.–Max.
Original constr.	38	113–151	18	61–79	41	37–78
Original unconstr.	64	84–148	26	50–76	43	34–77
Revised constr.	79	224–303	41	113–154	44	106–150
Revised unconstr.	108	188–296	68	84–152	44	104–148

Note: The scores are for only those participants who did not report knowledge of a second language. Min.–Max., minimum–maximum score.

words, when the participants knew how many correct answers there were, they achieved higher scores on the DVK. Furthermore, the scores on both subtests were higher in the constrained condition than in the unconstrained one. In addition, in each of the two groups, the scores on the paradigmatic subtest were higher than those on the syntagmatic subtest.

It can be seen from Table 3 that this overall pattern of results also holds for the revised scores. Percentage-wise, the original scores and revised scores are almost identical in the constrained condition. However, when it comes to the unconstrained version of the test, the difference between original scores and revised

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

Table 5. *Means and standard deviations of the original and revised Depth of Vocabulary Knowledge measure scores for the constrained (n = 24) and unconstrained (n = 20) groups*

	Overall Score		Paradigmatic		Syntagmatic	
	M (*SD*)	%	*M* (*SD*)	%	*M* (*SD*)	%
			Original			
Constr.	137.38 (12.17)	85.86	70.83 (5.70)	89.66	66.54 (11.20)	82.15
Unconstr.	128.10 (17.17)	80.06	66.10 (6.74)	83.67	62.00 (12.53)	76.54
			Revised			
Constr.	275.58 (22.75)	86.12	141.00 (11.16)	88.13	134.58 (12.87)	84.11
Unconstr.	266.25 (25.55)	83.20	137.90 (15.29)	86.19	128.35 (11.96)	80.22

Note: The means are reported as raw scores and percentages and they are for only those participants who did not report knowledge of a second language.

scores appears quite pronounced and is statistically significant for the overall scores, t (30) = 3.77, p = .001, with a medium effect size, d = 0.68; for the paradigmatic scores, t (30) = 3.75, p = .001, with a medium effect size, d = 0.67; and for the syntagmatic scores, t (30) = 2.60, p = .014, with a medium effect size, d = 0.47. In each case, the participants scored better when the revised scoring system was used. The Bonferroni-corrected alpha level for these three comparisons was 0.017.

In terms of the original scores on the two subtests, a repeated-measure analysis of variance (ANOVA) revealed a significant main effect of group, F (1, 65) = 15.54, p < .001. The η_p^2 value of 0.19 indicated that this was a medium-sized effect, meaning that group accounted for 19% of the variance in scores on the two subtests. In addition, there was a significant main effect of subtest (i.e., type of association, paradigmatic vs. syntagmatic): $F(1, 65) = 14.59, p < .001$. This effect was also medium-sized ($\eta_p^2 = 0.18$); subtest accounted for 18% of the variance in the scores. The interaction between group and subtest was not significant (see Figure 1).

In terms of the revised scores on the subtests, a repeated-measure ANOVA revealed a significant main effect of group (constrained vs. unconstrained), F (1, 65) = 10.86, p = .002. The η_p^2 value of 0.14 indicated that this was a medium-sized effect, meaning that group accounted for 14% of the variance in scores on the two subtests. In addition, there was a significant main effect of subtest (i.e., type of association, paradigmatic vs. syntagmatic): $F(1, 65) = 68.31, p < .001$. The effect was large ($\eta_p^2 = 0.51$); subtest accounted for 51% of the variance in scores. The interaction between group and subtest was significant, F (1, 65) = 5.85, p < .02, but with a small effect size ($\eta_p^2 = 0.08$; see Figure 2). This interaction indicates

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

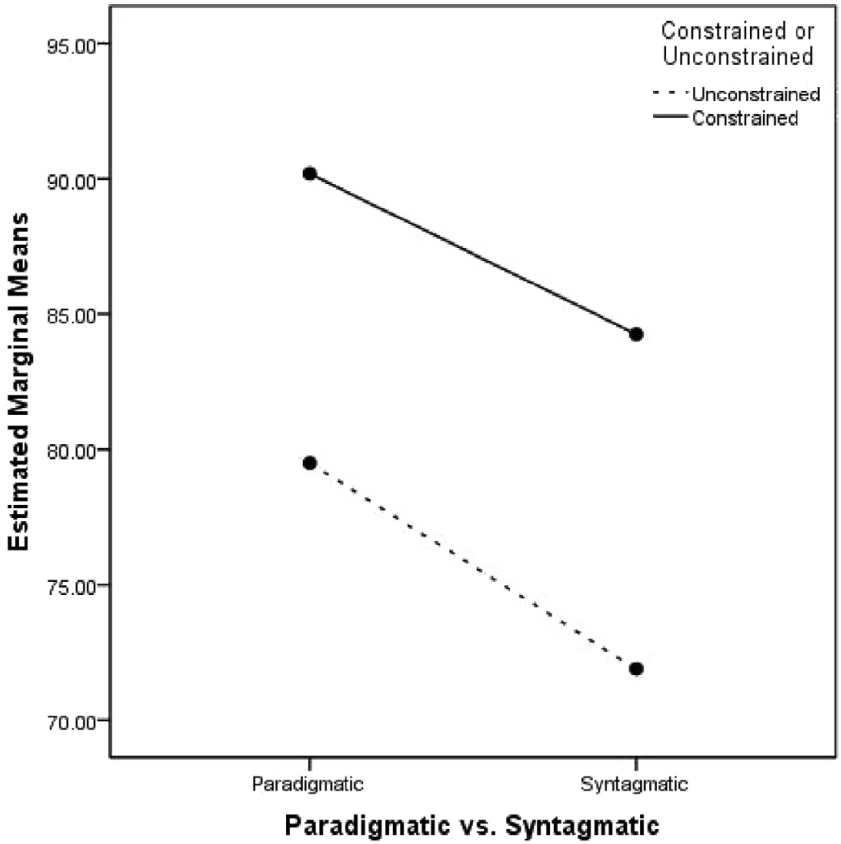


Figure 1. An analysis of variance plot for the original scores.

that the paradigmatic and syntagmatic revised scores were not equally affected by the introduction of the unconstrained condition, with syntagmatic scores being more negatively affected than their paradigmatic counterparts.

In light of the usual expectation in L2 testing that NSs should be able to achieve at or above the 90% mark (see, e.g., Klein-Braley, 1997), the results demonstrated that this was mostly not the case. Under the original scoring method, the 90% criterion would translate into the following mean scores: an overall score of 144, a paradigmatic score of 71.1, and a syntagmatic score of 72.9. With respect to the revised scoring scheme, the corresponding scores would be 288 (overall), 144 (paradigmatic), and 144 (syntagmatic). All of the mean scores except the original scores on the paradigmatic subtest for the constrained group were below the 90% mark. A series of one-sample *t* tests (the Bonferroni-corrected alpha level was 0.01 for the constrained group and 0.0083 for the unconstrained group, with 30 and 35 degrees of freedom, and five and six comparisons, respectively) showed that

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

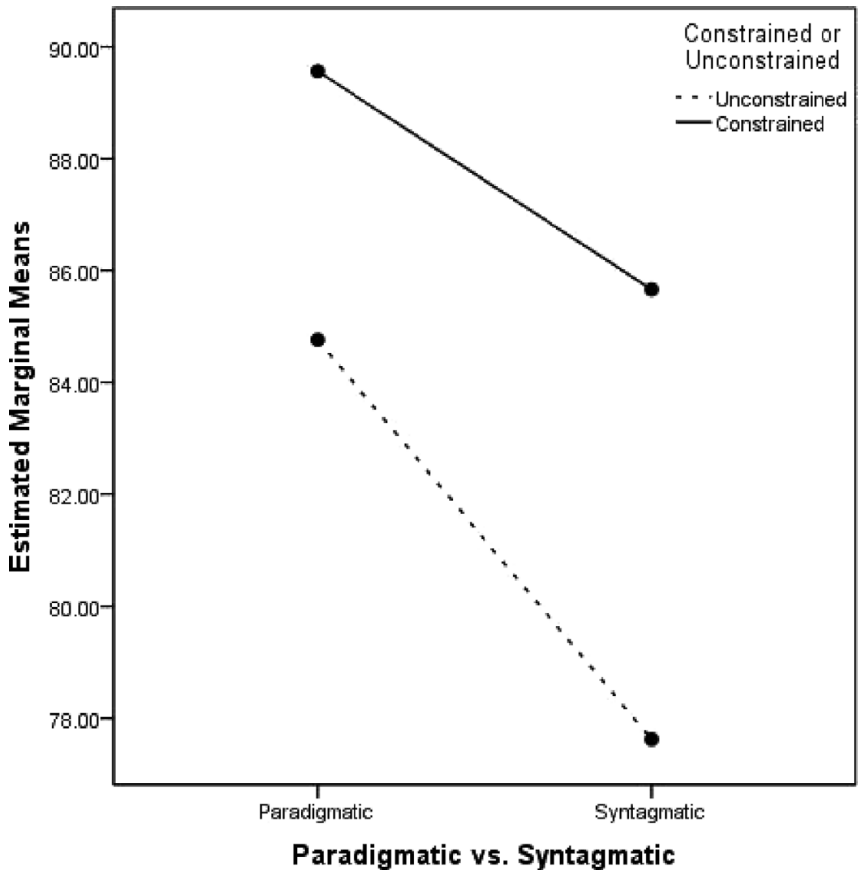


Figure 2. An analysis of variance plot for the revised scores.

four of the five original scores that were under the 90% mark were significantly different from the expected NS performance, overall unconstrained, $t\,(30) = 5.10$, $p < .001$, $d = 0.92$; paradigmatic unconstrained, $t\,(30) = 4.40$, $p < .001$, $d = 0.79$; syntagmatic constrained, $t\,(35) = 2.81$, $p = .008$, $d = 0.47$; and syntagmatic unconstrained, $t\,(30) = 4.90$, $p < .001$, $d = 0.88$. Although the mean score was below the 90% criterion, due to the (conservative) Bonferroni correction of the alpha level, the overall constrained scores were not significantly different from the expected minimum NS performance, $t\,(35) = 2.25$, $p = .03$, but the difference did still approach significance. In other words, in terms of original scores, participants met the 90% criterion only when they knew how many correct answers there were per item and only with the overall and paradigmatic scores.

The revised scores closely mirrored the above pattern. Four of the five revised scores under the 90% mark were significantly different (the Bonferroni-corrected

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

Table 6. *Ranges, minimum, and maximum d' scores for the constrained and unconstrained groups*

	Overall d'		Paradigmatic d'		Syntagmatic d'	
	Range	Min.–Max.	Range	Min.–Max.	Range	Min.–Max.
Constrained	7.47	−4.79 to 2.68	4.60	−2.85 to 1.75	6.42	−3.91 to 2.51
Unconstrained	6.40	−4.52 to 1.88	7.05	−5.42 to 1.63	4.60	−2.85 to 1.75

Note: Min.–Max., minimum–maximum score.

alpha level was 0.01 for the constrained group and 0.0083 for the unconstrained group, with 30 and 35 degrees of freedom, and five and six comparisons, respectively) from the expected NS performance, overall unconstrained, $t\ (30) = 5.45$, $p < .001$, $d = 0.98$; paradigmatic unconstrained, $t\ (30) = 3.02$, $p < .005$, $d = 0.54$; syntagmatic constrained, $t\ (35) = 3.31$, $p < .002$, $d = 0.55$; and syntagmatic unconstrained, $t\ (30) = 7.26$, $p < .001$, $d = 1.30$). Although the mean score was below 90%, and the difference approached significance, the overall constrained revised score was not significantly different from the expected minimum NS performance, $t\ (35) = 2.07$, $p = .046$ under the imposed Bonferroni adjustment. In other words, in terms of revised scores, participants were unable to meet the 90% criterion when they did not know how many correct answers there were per item.[3] In addition, they were not able to perform at 90% on the syntagmatic subtest regardless of the instructions they were given.

Apart from including both original and revised DVK scores in our analysis, we also computed three d' scores for each participant, one for the overall test, one for the paradigmatic subtest, and one for the syntagmatic subtest. In signal detection theory (see, e.g., Macmillan & Creelman, 1991), the d' score, the difference between a person's standardized hit and false alarm rates, is a measure of an individual's ability to discriminate between signal and noise, or, in terms of the present test, to accurately identify a large number of correct answers while keeping low the proportion of distractors wrongly selected as answers. A d' score of 0 (or near 0) would indicate chance-level performance, whereas a high absolute d' value is an indicator of sensitivity to the difference between signal and noise. Table 6 summarizes the ranges and minimum and maximum d' scores for the entire test as well as for the two subtests. Table 7 summarizes the means and standard deviations of the d' scores.

As can be seen from Tables 6 and 7, both the constrained and unconstrained groups included participants that performed at chance level as well as those with a good ability to discriminate between the correct answers and distractors on the DVK. In both groups, all six mean d' scores were not significantly different from 0. Due to the presence of a number of outliers (six in each group, two per each of the group d' scores), the d' score distributions were Winsorized at the 5th and 95th percentile levels. An independent samples t test and a repeated-measures ANOVA performed on the Winsorized data indicated that there was still no difference between the constrained and unconstrained groups in terms of d' scores. (A

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

Table 7. *Means and standard deviations of the d′ scores for the constrained and unconstrained groups*

	Overall *d′*		Paradigmatic *d′*		Syntagmatic *d′*	
	M	*SD*	*M*	*SD*	*M*	*SD*
Constrained	$0 < M < 0.0001$	1.88	$0 < M < 0.0001$	1.30	$0 < M < 0.0001$	1.53
Unconstrained	$0 < M < 0.0001$	1.50	$0 < M < 0.0001$	1.57	$0 < M < 0.0001$	1.08

parallel analysis of the non-Winsorized data implied the same, which means that the outliers did not significantly impact the *d′* scores.) In other words, participants were equally likely to employ random guessing as a test-taking strategy regardless of the instructions they were given. Therefore, random guessing appears not to have been the source of the observed differences between the groups and subtests.

To investigate whether the scores on the paradigmatic and syntagmatic tests were related, we computed Spearman correlations between these scores for both groups (the normal distribution assumption for Pearson correlations was not met). In the case of the original DVK scores, for the constrained group, there was no significant correlation between the paradigmatic and syntagmatic subtests. A closer inspection of the data revealed that two participants who had achieved very high paradigmatic scores (among the highest in their group) simultaneously obtained the two lowest syntagmatic scores (at or below 40%). When these two influential data points were disregarded, there emerged a moderate correlation between the syntagmatic and paradigmatic subtests for the constrained group ($r = .383$, $p = .025$). This means that less than 15% of the variance in syntagmatic DVK scores is explained by paradigmatic scores and vice versa. It is important to note, however, that the biographical data reveal no obvious reason why the two participants' scores should be excluded from the analysis. For the unconstrained group, there was a moderate correlation between the paradigmatic and syntagmatic subscores ($r = .46$, $p = .01$), indicating that around 21% of the variance in one subscore could be predicted by the other. In contrast, the revised paradigmatic and syntagmatic scores were highly correlated for both the constrained ($r = .88$, $p < .001$) and unconstrained groups ($r = .76$, $p < .001$), indicating that 77% and 57% of one score could be predicted by the other.

Table 8 presents the alpha reliability scores for the overall DVK as well as for the two subtests in the constrained and unconstrained conditions in the present study together with reliability data from Qian and Schedl's study (2004). (Note that the latter study only featured a constrained condition.) As can be seen from the table, the DVK consistently appears to yield high or acceptable alpha reliability scores, particularly in the unconstrained modality.

A breakdown of the percentages of correct NS responses in each of the test conditions appears in Table 9 (eight per headword when both correct hits and misses are scored; in other words, the revised scoring system was used for the purposes of item analysis). Only 60% of all items were answered correctly by

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

Table 8. *Alpha reliability scores for the Depth of Vocabulary Knowledge measure*

	Qian & Schedl for L2		Present Study for L1	
	Constrained	Unconstrained	Constrained	Unconstrained
Overall	0.91	—	0.94	0.95
Paradigmatic	0.83	—	0.89	0.93
Syntagmatic	0.86	—	0.90	0.95

Note: The scores are according to Qian and Schedl (2004) and the present study. L2, second language; L1, first language.

Table 9. *Item analysis: Numbers and percentages of items with correct native-speaker responses*

	Constrained		Unconstrained	
	No. of Items	%	No. of Items	%
Above 90%	192	60.0	129	40.3
Above 80%	259	80.8	206	64.4
Below 80%	61	20.2	114	35.6
Below 50%	7	2.2	14	4.2

Note: There were 320 items (choices) in the test, four correct and four incorrect per headword.

more than 90% of participants in the constrained condition compared to 40.3% in the unconstrained one.

DISCUSSION

The results of this study bring to light a few pitfalls of a WAT-type test, of which some are easily remedied whereas others remain intractable.

One of the main findings of the present study was the substantial inconsistency of NS responses on various items in both versions of the test, but especially so in the unconstrained one. This remained true even when we used a revised scoring system, which led to enhanced scores. If a criterion of 90% of correct NS answers is adopted as a measure of minimal acceptable discriminative validity for an item in an L2 proficiency test, it can be concluded that a large proportion of the items in this test fall short of this criterion. Whereas 60% of the items in the constrained condition were answered with above 90% consistency, only 40.3% of the items in the unconstrained version met the 90% criterion. Although a smattering of idiosyncratic responses can be expected from NSs even on robust test items, a certain minimum criterion of consistency of NS responses must be imposed if any

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

test is to be a valid indicator of L2 proficiency. This obviously implies that any test of lexical depth that adopts the WAT format must include items that are checked for frequency in an appropriate corpus, for example, the British National Corpus (BNC) or the Corpus of Contemporary American English, especially in the case of multiword items such as collocates. In addition, the test needs to undergo extensive piloting with NSs of a range of varieties of the target language before its use with L2 speakers is even considered. For instance, take Headword 7 (*insufficient*), for which one of the intended correct collocates was *need*. Although *insufficient need* is a possible adjective–noun combination, a search of three corpora, the BNC, Corpus of Contemporary American English, and Collins Wordbanks Online, yields zero hits for this combination, suggesting that *insufficient need* does not qualify as a collocation. Thus, it is not surprising that only 61.11% of the participants in the constrained group and as few as 12.9% in the unconstrained group chose this as a "normal" combination (for instructions that ask participants to pick "normal combinations," see Qian and Schedl, 2004, p. 50). Likewise, in the case of Headword 34 (*terse*), for which one of the intended meanings associates is *concise*, 50% of the participants selected *concise* as an associate of *terse* in the constrained condition versus only 29.03% in the unconstrained condition, possibly because many of the participants did not know the meaning of *terse*. In both cases, we might conjecture that the higher scores in the constrained condition were simply the result of educated guessing (by elimination) because participants knew that they needed to come up with the four least problematic answers per headword.

Although it was long an accepted view that NSs generally display very consistent word-association behavior (Postman & Keppel, 1970), which was also the basis and rationale for the creation of the numerous available word association thesauri (e.g. Kiss, Armstrong, Milroy, & Piper, 1973; Nelson, McEvoy, & Schreiber, 1998; Postman & Keppel, 1970), recent studies have found that this is the case only when L1 responses are analyzed at the level of group (Fitzpatrick, 2007). Our findings, as well as those of Wilks and Meara (2007), convincingly demonstrate that NSs exhibit a considerable degree of variation on receptive word-association tasks, which is probably a function of individual response styles. Our findings, furthermore, suggest that syntagmatic associations may be particularly inconsistent within groups of seemingly comparable speakers. This inconsistency across speakers is also found on productive tasks (Fitzpatrick, 2007). Individuals are quite consistent in their word-association behavior across administrations of the same task and, for bilinguals (particularly more advanced ones), even across languages (Fitzpatrick, 2007, 2009).

Guessing and scoring

Our findings support Read's (1993, 2000) misgivings about guessing being an inherent flaw in the WAT format. Our d' analysis shows that guessing does occur in this type of test. However, although we had expected the unconstrained condition to bring the various items grouped under each headword closer to true independence, thereby reducing potential for guessing, participants were equally likely to employ guessing as a test-taking strategy regardless of the instructions they received. (In

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

other words, those who will guess will guess.) Therefore, the significant difference in the DVK scores that we observed between the constrained and unconstrained conditions cannot be attributable to random guessing. Rather, we believe that it was largely due to the fact that test takers in the unconstrained condition tended to be more conservative when selecting associates. (Recall that numerous intended answers in the DVK test are less than optimal in terms of real-life language usage.) Conversely, the constrained condition was likely to force test takers into choices they would otherwise have been reluctant to make (e.g., *insufficient need*), as they were aware that they needed to find 4 associates per headword.

With regard to the desirability of guessing, although Bogaards (2000) rightly points out that the willingness and ability to infer the correct answer might be reflective of a heightened ability to cope with real-life language comprehension issues, this skill is not what is supposed to be targeted by WAT-type tests. In addition, because guessing is demonstrably an issue with this test format, there is a greater need than ever to create a fine-tuned scoring system that takes not only "hits" into consideration but also correct rejections, false alarms, and misses, because such a system can identify the degree of random guessing that participants are engaging in. It would also seem that the unconstrained format could significantly reduce the skewing of test results brought about by participants' awareness of the desired number of correct answers per headword.

NS syntagmatic and paradigmatic associations

Another important finding of the present study is that NSs identify meaning (paradigmatic) relationships between words on this test more accurately and more consistently than collocational (syntagmatic) ones irrespective of whether the scoring is as per the original test or our revised method. In contrast with Qian's (1998) finding of a strong correlation ($r = .80$, $p = .05$) between the paradigmatic and syntagmatic scores for nonnative speakers, the present study did not find such a correlation for NSs when the original scoring system was used; there was a positive correlation only when the scoring was revised. The implication of this finding is that, at least as far as these NSs of English are concerned, 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. However, one would sooner expect to find such a discrepancy in an L2 speaker than in an L1 speaker (e.g., see Greidanus & Nienhuis, 2001), as L1 speakers largely acquire word meaning and usage naturalistically through exposure to massive amounts of input and ample opportunity for interaction. It is difficult to imagine how L2 speakers, generally having had considerably less opportunity to acquire collocational knowledge, could demonstrate a higher correlation between paradigmatic and syntagmatic lexical knowledge. More important, if there is a correlation with one group of test takers and not with another or one way of scoring and not another, syntagmatic and paradigmatic scores are best not collapsed into a single score.

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

Linguistic structure mapping onto conceptual structure (problem with adjectives as headwords)

Noun phrases are built around nouns. Nouns dictate the choice and/or form of their specifiers (e.g., articles), modifiers (e.g., adjective phrases), and complements (e.g., prepositional phrases) within the noun phrase. For instance, whether an indefinite article is used in an English noun phrase depends on such factors as the countability and number of the head noun. Moreover, in languages with adjective–noun agreement (such as French, Dutch, or Serbo-Croatian), the form of the adjective depends on one or more properties of the noun, such as gender, number, case, or definiteness. Noun phrase production in natural speech requires that the noun lemma be accessed first in order to compute the form of any accompanying specifiers and modifiers (for a study on Dutch, see Schriefers, 1993). Put succinctly, adjectives are about nouns; not the other way around, irrespective of whether the adjective is prenominal or postnominal. However, the WAT format requires the test taker to begin with an adjective and determine whether the supplied nouns are a potential fit, which is the opposite of what probably happens in natural speech production, with the exception of extremely tight collocates that can be treated as chunks, for example, "constitutional rights" or "(no) running water." (For mapping of conceptual structure onto linguistic structure, see Jackendoff, 1990; and Levelt, Roeloefs, & Meyer, 1999.)

Interference and continuous design

Mature mental lexicons are characterized by a dense network of connections within the lexical level of representation as well as between the lexical and conceptual levels (Bock & Levelt, 1994; Collins & Loftus, 1975). For each lexical entry in the mental lexicon, the pertinent phonological, orthographic, morphological, semantic, and syntactic knowledge is interconnected. There are also numerous links of varying strength between a lexical entry and its synonyms, antonyms, hyponyms, hypernyms, cohyponyms, meronyms, phonological and orthographic neighbors, collocates, and so forth (Levelt, 1989; Collins & Quillian, 1969). Within such a network view of the mental lexicon, the testing of lexical depth can be defined as the quantitative and qualitative probing of any number of the above types of connections.

The WAT-DVK is a highly metalinguistic test, and taking it in one short sitting requires repeated evaluation of adjective meanings and judgments of the acceptability of adjective–noun combinations. The fact that it employs a continuous design (i.e., all the potential meaning and collocational associates of a headword are presented together without intervening items) leads to the repeated activation and selection of each headword. In word comprehension, a semantic facilitation effect would be expected whenever semantically related words are presented in quick succession (Meyer & Schvaneveldt, 1971; Neely, 1991). However, note that the process of taking a WAT-type test not only entails comprehension processes but inevitably also involves articulating various word combinations vocally or subvocally (both to test meanings in the paradigmatic column and to determine potential syntagmatic collocates in the other). In short, this process

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

is in many ways akin to speech production. In speech production, when two semantically related lexical items are activated in close temporal proximity, the selection of the second item becomes more difficult if both words belong to the same level of semantic organization, as is the case with synonyms (Schriefers, Meyer, & Levelt, 1990; see also Alario, Segui, & Ferrand, 2000; Costa, Alario, & Sebastián-Gallés, 2009; Vitkovitch, Rutter, & Read, 2001; and Wheeldon & Monsell, 1994).

Although the exact underlying neural architecture of this interference effect is still under dispute (i.e., whether it happens through competition among lexical items during spreading activation or through inhibitory links in the mental lexicon), its behavioral manifestations remain a fact (Levelt et al., 1999; see also Dell & O'Seaghdha, 1991; Ferrand, 1999; and Finkbeiner & Nicol, 2003). Therefore, we can conceive of the possibility that a similar or stronger semantic interference effect might exist between the various senses of a polysemous word, for example, when the test taker is expected to pick separate meanings of a headword such as *obscure* in quick succession: "unclear" as in "an obscure fact" or "unknown" as in "an obscure musician" (DVK Item 9; Qian & Schedl, 2004). In other words, it may become increasingly difficult to access alternate meanings of a word with each new attempt when these attempts occur in close succession. A similar interference effect may come into play with polysemy involving contiguous collocates of a headword prompt (e.g., *fine day* and *fine athlete* in Item 22). Furthermore, contiguous orthographically/phonologically similar headwords such as *perceptible* and *perpetual* (Items 38 and 39) may interfere with each other in receptive lexical access (Rastle, 2009). We believe that the effects outlined above may be one reason why NSs performed worse than expected on the DVK, especially in the unconstrained version of the test.

Another reason why a continuous design might not be the best solution for a word-association test is that the frequent repetition of a stimulus word could perhaps cause this word to temporarily lose some of its meaning for the test taker. This semantic satiation effect has been the subject of psycholinguistic and psychological inquiry for a long time (e.g., Black, 2001; Cohene, Smith, & Klein, 1978; Fillenbaum, 1964; Pynte, 1991). Although the satiation effect is reminiscent of selective adaptation (the process whereby the repetition of a sensory stimulus reduces sensitivity to it), event-related potential research suggests that semantic satiation can be a separate effect occurring in semantic memory, independently of perceptual adaptation (Kounios, Kotz, & Holcomb, 2000).

CONCLUSION

The WAT format has a number of obvious virtues. It is quick to administer, easy to score, and appears to have face validity (as also determined by Qian & Schedl, 2004). Furthermore, WAT-type tests created for L2 learners of English in the last 20 years (e.g., by Read, 1993; and Qian & Schedl, 2004) have been pioneering in the context of L2 lexical testing and have provided much-needed additions to the narrow repertoire of potential depth tests. However, the present

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

study highlights some of the pitfalls of fixed-answer WAT-type tests, some of which can be remedied through a reformulation of test format and content (for L2 speakers of French, as in Greidanus et al., 2004, 2006) as well as analysis (Fitzpatrick, 2009). Yet other problems appear to have no easy solutions and are major enough to cast doubt on the use of WAT-type tests as valid instruments of lexical depth assessment.

As discussed above, the current design of these tests may place unusual and unnatural strain on the language-processing system because it involves the repeated activation of headwords and their associates in quick succession and without meaningful context. Such a task may be especially burdensome for L2 learners, for example, because switching from core meanings of polysemous words to peripheral meanings can be expected to be particularly difficult when lexical semantic networks are being newly built. The issues related to activation, however, can be resolved through scrambling the various headword meanings and collocations so that any two items linked to a headword are well separated. Furthermore, to remedy the variability in the quality of individual items, the class of the headword can be chosen in accordance with psycholinguistic principles (e.g., verbs rather than their arguments, or nouns rather than their modifying adjectives, as has been done by Greidanus & Nienhuis, 2001). In addition, the test creator can ensure that these word classes are unambiguous and do not have potential responses that could trip learners up. (The headword "celebrated" in the DVK, for example, could easily be interpreted as a verb rather than a deverbal adjective; if so, an incorrect response such as "festive" could become a distracting choice for semantic reasons.) There are also good reasons for keeping syntagmatic and paradigmatic scores separate and avoiding forcing test takers into unnatural choices by leaving the number of correct answers unconstrained (as in Greidanus et al., 2004).

A more serious problem with WAT-type tests lies in the choice of associations accompanying headwords, especially when these are collocational. Ideally, associations need to be developed with reference to large corpora of spoken and written language, such as the BNC or the COBUILD Bank of English corpus, with careful attention to the frequency of headwords and the use of typical rather than marginal collocations. Certainly test developers' intuitions alone will not suffice because NS judgments can exhibit considerable variability (e.g., Dronjic, 2008). Furthermore, these tests need to be piloted with not just L2 speakers but also NSs from a variety of backgrounds (e.g., in terms of level of education, socioeconomic status, or language background), followed by statistical analysis of item difficulty and reliability.

However, the problem that remains is that in any language both meaning and collocation vary from speaker to speaker across various dimensions such as personal cognitive or response style, dialect, age, gender, socioeconomic status, and profession. Consider, for example, the new meaning of *radical* or the dialectal variation in the use of phrases such as *get ill*, *get pregnant*, and *fall ill*, *fall pregnant*. Moreover, collocations tend to vary in terms of fixedness; some can be rather loose and others very tight (Moon, 1997), reflecting the constant tension between two opposing forces that govern lexical patterning: the open-choice principle and the

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

idiom principle (Sinclair, 1991). Thus, it is unlikely that the choice of collocations will ever fit all contexts or will remain stable for more than a few years. Even more serious questions arise when L2 speakers' culture-specific responses are to be evaluated. We feel that it is probably unreasonable in many instances to treat bilinguals' responses as being nontargetlike (i.e., defective in some sense) simply because they do not dovetail with monolingual NS' responses (Meara & Wolter, 2004; Wilks & Meara, 2007). This may be an especially grave concern when a test is receptive and not productive, that is, involving judgments rather than actual spontaneous production.

If only the most robust associates were to be chosen, it is not clear whether the test could make fine-tuned distinctions between learners of varying lexical abilities. As our findings as well as those of Fitzpatrick (2007, 2009) and Wilks and Meara (2007) suggest, it may simply not be justified to evaluate individuals based on whether their personal word-association behavior conforms to that of a majority of speakers in their group or not, as much of this behavior appears not to depend on proficiency but idiosyncratic response preferences. Fitzpatrick's (2009) finding that bilinguals tend to converge on their own L1 word-association style with increasing L2 proficiency seems to suggest that the only potentially valid way of evaluating L2 speakers based on the WAT format would be to also collect data in their L1. Although this may be possible in some contexts (e.g., research or evaluation in the context of a monolingual L2 class), it will be highly impractical in others (e.g., heterogeneous classes with people from numerous L1 backgrounds). At any rate, much more theoretical inquiry into this will be necessary before we start seeing the first valid and reliable testing applications of these interesting findings.

The most vexing issue is whether WAT-type tests, which are predominantly receptive, measure lexical depth. To expect learners to know the meaning of words within an appropriate frequency band or register is reasonable, as in the case of the DVK's paradigmatic column no spontaneous production is involved. In terms of collocational knowledge, the lack of production is even more grave because a learner's ability to distinguish between native and nonnative collocations receptively is barely relevant to using these collocations meaningfully in spontaneous production, and we believe that collocations are meaningful test items only in production (e.g., being able to come up with "torrential rain" in discourse rather than picking "torrential rain" rather than *"torrential snow" receptively).

If lexical depth includes a fine-tuned knowledge of the grammatical behavior of words, their connotations or stylistic nuances, as well as the richness of the associational networks in the mental lexicon, a test that does not entail spontaneous production but instead conducts a spot check of a learner's ability to pick out a limited number of preassigned options that are presumed correct, a WAT-type test is not a test of depth. (Note that the subvocal or vocal testing of collocates discussed earlier is not spontaneous production.) In this sense, Wesche and Paribakht's (1996) Vocabulary Knowledge Scale is superior because the maximum score for an item includes its use in a sentence, displaying not only word syntax but possibly also collocational knowledge. Furthermore, the collocational

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

selections made in a multiple choice type context are made with the help of what Bialystok (1994) would call "analyzed knowledge," as is the case when learners complete grammaticality judgment or cloze tests. In cases where the learner is simply guessing, this might not even be "analyzed knowledge" but "indeterminate knowledge" (Ellis, 1991). As for providing multiple meanings of polysemous words, the task is akin to completing some types of crossword puzzles, where the solution lies in falling upon extended or peripheral meanings of otherwise common words.

In light of these concerns we recommend that word-association tasks are best used for exploratory purposes, that is, to explore the organization and functioning of the L2 mental lexicon, paralleling what has been traditionally done in L1 contexts, or to ascertain what similarities and differences exist between the L1 and L2 lexicons (Fitzpatrick, 2007; Henriksen, 2008; Meara, 1983, 2009). Particularly valuable in this respect are free, timed L2 word-association tasks, which can reflect group-level quantitative differences between L1 and L2 associations (e.g., Zareva, 2007) as well as qualitative differences (e.g., Wolter, 2006), which are to be expected for both interlingual linguistic reasons and cultural ones. In other words, these L2 association tasks are probably much more suitable for descriptive purposes, that is, investigating the structure and functioning of the mental lexicon, than for prescription or evaluation.

APPENDIX A

Questionnaire for participants

1. What year of university are you in? 1st 2nd 3rd 4th other
2. What program(s) are you in (i.e., your major/specialization)?
3. What is the very first language that you learned (i.e., chronologically first)?
4. a. If you speak another language, what language is it? (*Note:* This should be your second strongest language.)
   b. Please rate your proficiency in this language on a scale of 1 (*beginner*) to 5 (*nativelike*):

Reading				
Beginner				Nativelike
1	2	3	4	5
Writing				
Beginner				Nativelike
1	2	3	4	5
Listening				
Beginner				Nativelike
1	2	3	4	5
Speaking				
Beginner				Nativelike
1	2	3	4	5

5. What age group are you in?
   18–29 30–39 40–49 50+
6. Research assistant's notes:

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

APPENDIX B

The following are the instructions for the constrained version of the DVK, which is based on Qian and Schedl (2004, p. 50).
Directions: In this test, there are 40 items. Each item looks like this:

Sound

(A) logical	(B) healthy	(E) snow	(F) temperature
(C) bold	(D) solid	(G) sleep	(H) dance

Please note:

Some of the words here in the left box are similar in meaning to the meaning of *sound.*	Some of the words in the right box are nouns that can be used after *sound* in a phrase or a sentence.
In the left box, **"logical," "healthy," and "solid"** all share the meaning of "sound."	We do not normally say "sound snow," "sound temperature," or "sound dance," but we often say "sound sleep." So **"sleep"** is the correct answer on this side.

There are eight words in the two boxes (A–H) in the 40 items below. Please highlight **only four** correct/appropriate answers. In each case there are **only four** answers that are correct/appropriate.

In the case of the "sound" in the example above, the answers would look like this:

Sound

(A) **logical**	(B) **healthy**	(E) snow	(F) temperature
(C) bold	(D) **solid**	(G) **sleep**	(H) dance

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

APPENDIX C

The following are the instructions for the unconstrained version of the DVK.
Directions: In this test, there are 40 items. Each item looks like this:

Sound

(A) logical	(B) healthy	(E) snow	(F) temperature
(C) bold	(D) solid	(G) sleep	(H) dance

Please note:

Some of the words here in the left box are similar in meaning to the meaning of *sound*	Some of the words in the right box are nouns that can be used after *sound* in a phrase or a sentence
In the left box, **"logical," "healthy," and "solid"** all share the meaning of "sound."	We do not normally say "sound snow," "sound temperature," or "sound dance," but we often say "sound sleep." So **"sleep"** is the correct answer on this side.

There are eight words in the two boxes (A–H) in the 40 items below. Please highlight **any number** of correct/appropriate answers. In each case there could be **one to eight** answers that are correct/appropriate.

In the case of the "sound" in the example above, the answers would look like this:

Sound

(A) **logical**	(B) **healthy**	(E) snow	(F) temperature
(C) bold	(D) **solid**	(G) **sleep**	(H) dance

ACKNOWLEDGMENTS

This research was funded by SSHRC Standard Reseach Grant 410-2004-2000 (to R. H.-P.). We thank Maria Claudia Petrescu, Claudette D'Souza, Ellen Moore, Anna Nirukshi Arulampalam, and Hyeyoon Cho for their kind assistance with data collection and proofreading. In addition, the authors are indebted to Xi Chen, Denis Burnham, and three anonymous reviewers for useful insights that enriched the analysis of the data and the interpretation of the results. Of course, any remaining shortcomings of the article should be attributed to the authors alone.

NOTES

1. The larger purpose of the study was to investigate cognate facilitation in L2 tests.
2. Recall that our data come from a sample of native English speakers consisting of individuals who had reported no knowledge of a second language (approximately two-thirds of each group) as well as those reporting varying degrees of knowledge of an

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

L2 (approximately one-third; see Table 1 for details), including minimal knowledge. A complicating factor is that some participants in the "pure" monolingual group might have also had minimal knowledge of a language such as French because it is customary for students to take "Core French" (French as a subject) in their middle school grades. Although we have provided a detailed rationale for choosing to work with such a sample and have not sought to contrast monolingual and bilingual NSs, we do recognize that certain readers might be interested in the performance of self-reported pure monolinguals on the DVK. We have therefore provided descriptive statistics for all participants who did not report knowing a second language (see Table 4 and Table 5). It should be noted, nonetheless, that these monolingual constrained and unconstrained samples are small ($n = 24$ and 20, respectively) and therefore cannot be assumed to be representative.

3. The exception was the paradigmatic score. Even here, however, the difference from the 90% criterion approached significance.

REFERENCES

Alario, F. X., Segui, J., & Ferrand, L. (2000). Semantic and associative priming in picture naming. *Quarterly Journal of Experimental Psychology, 53A*, 741–764.

Bialystok, E. (1994). Representation and ways of knowing: Three issues in second-language acquisition. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 549–569). San Diego, CA: Academic Press.

Black, S. R. (2001). Semantic satiation and lexical ambiguity resolution. *American Journal of Psychology, 114*, 493–510.

Bock, K., & Levelt, W. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). San Diego, CA: Academic Press.

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

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interface between second language acquisition and language testing research* (pp. 32–70). Cambridge: Cambridge University Press.

Cohene, L. S., Smith, M. C., & Klein, D. (1978). Semantic satiation revisited with a lexical decision task. *Memory and Cognition, 6*, 131–140.

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review, 82*, 407–428.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8*, 240–247.

Costa, A., Alario, F. X., & Sebastián-Gallés, N. (2009). Cross-linguistic research on language production. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 531–546). Clevedon: Multilingual Matters.

Cronbach, L. J. (1942). Measuring knowledge of precise word meaning. *Journal of Educational Research, 36*, 528–534.

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English, 42*, 895–901.

Dell, G. S., & O'Seaghdha, P. (1991). Mediated and convergent lexical priming in language production: A comment on Levelt et al. (1991). *Psychological Review, 98*, 604–614.

Dronjic, V. (2008). *The influence of L2 English on the L1 Serbo-Croatian of Toronto-based bilinguals: The case of grammaticality judgments at the lexis/syntax interface*. Unpublished master's thesis, University of Toronto.

Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Ellis, R. (1991). Grammaticality judgements and second-language acquisition. *Studies in Second Language Acquisition, 13*, 161–186.

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

Ferrand, L. (1999). Applying Ockham's chainsaw in modeling speech production. *Behavioral and Brain Sciences, 22*, 42–43.
Fillenbaum, S. (1964). Semantic satiation and decision latency. *Journal of Experimental Psychology, 68*, 240–244.
Finkbeiner, M., & Nicol, J. (2003). Semantic category effects in second-language word learning. *Applied Psycholinguistics, 24*, 369–383.
Fitzpatrick, T. (2007). Word association patterns: Unpacking the assumptions. *International Journal of Applied Linguistics, 17*, 319–331.
Fitzpatrick, T. (2009). Word association profiles in a first and second language: Puzzles and problems. In T. Fitzpatrick & A. Barfield (Eds.), *Lexical processing in second-language learners* (pp. 38–52). Toronto: Multilingual Matters.
Greidanus, T., Beks, B., & Wakely, R. (2006) Testing the development of French word knowledge by advanced Dutch- and English-speaking learners and native speakers. *Canadian Modern Language Review, 62*, 509–532.
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 (Eds.), *Vocabulary in a second language* (pp. 191–208). Amsterdam: John Benjamins.
Greidanus, T., & Nienhuis, L. (2001). Testing the quality of word knowledge in L2 by means of word associations: Types of distractors and types of associations. *Modern Language Journal, 85*, 567–77.
Henriksen, B. (2008). Declarative lexical knowledge. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), *Vocabulary and writing in a first and second language: Process and development* (pp. 22–66). New York: Palgrave Macmillan.
Heredia, R. R., & Brown, J. M. (2006). Bilingual Memory. In T. K. Bhatia & W. C. Ritchie (Eds.), *The handbook of bilingualism* (pp. 225–249). Malden, MA: Blackwell.
Jackendoff, R. (1990). *Semantic structures*. Cambridge, MA: MIT Press.
Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics, 21*, 47–77.
Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973) An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literary studies*. Edinburgh: Edinburgh University Press.
Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing, 14*, 47–84.
Kounios, J., Kotz, S. A., & Holcomb, P. J. (2000). On the locus of the semantic satiation effect: Evidence from event-related brain potentials. *Memory and Cognition, 28*, 1366–1377.
Kroll, J. F., & Dijkstra, A. F. J. (2002). The bilingual lexicon. In R. A. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 301–321). Oxford: Oxford University Press.
Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33*, 149–174.
Levelt, W. J. M. (1989). *Speaking*. Cambridge, MA: MIT Press.
Levelt, W. J. M., Roeloefs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1–75.
Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.
Meara, P. (1983). Word associations in a second language. *Nottingham Linguistics Circular, 11*, 28–38.
Meara, P. (1992). *Euralex French tests*. Swansea: Centre for Applied Language Studies.
Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjær, & J. Williams (Eds.), *Performance and competence in second-language acquisition* (pp. 35–53). Cambridge: Cambridge University Press.
Meara, P. (2009). *Connected words: Word associations and second-language vocabulary acquisition*. Amsterdam: John Benjamins.
Meara, P., & Jones, G. (1990). *Eurocentres vocabulary size test, version E1.1/K10*. Zurich: Eurocentres Learning Service.
Meara, P. M., & Wolter, B. (2004). V_Links: Beyond vocabulary depth. *Angles on the English-Speaking World, 4*, 85–97.

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

Meyer, D. E., & Schvaneveldt, R. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*, 227–239.

Moon, R. (1997). Vocabulary connections: Multiword items in English. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 40–64). Cambridge: Cambridge University Press.

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

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

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264–336). Hillsdale, NJ: Erlbaum.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. Retrieved from http://www.usf.edu/FreeAssociation/

Postman, L., & Keppel, G. (1970.) *Norms of word associations.* New York: Academic Press.

Pynte, J. (1991). The locus of semantic satiation in category membership decision and acceptability judgment. *Journal of Psycholinguistic Research, 20*, 315–335.

Qian, D. D. (1998). *Depth of vocabulary knowledge: Assessing its role in adults' comprehension in English as a second language*. PhD dissertation, University of Toronto.

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

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.

Rastle, K. (2009). Visual word recognition. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 73–87). Clevedon: Multilingual Matters.

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

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

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 209–227). Amsterdam: John Benjamins.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*, 55–89.

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

Schriefers, H. (1993). Syntactic processes in the production of noun phrases. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 841–850.

Schriefers, H., Meyer, A. S., & Levelt, W. J. M. (1990). Exploring the time course of lexical access in language production: Picture–word interference studies. *Journal of Memory and Language, 29*, 86–102.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Vitkovitch, M., Rutter, C., & Read, A. (2001). Inhibitory effects during object name retrieval: The effect of interval between prime and target on picture naming responses. *British Journal of Psychology, 92*, 483–506.

Wesche, M., & Paribakht, T. S. (1996). Assessing second-language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review, 53*, 13–40.

Wheeldon, L. R., & Monsell, S. (1994). Inhibition of spoken word production by priming a semantic competitor. *Journal of Memory and Language, 33*, 332–356.

Wilks, C., & Meara, P. (2007). Implementing graph theory approaches to the exploration of density and structure in L1 and L2 word-association networks. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 167–181). Cambridge: Cambridge University Press.

Wolter, B., (2006). Lexical network structures and L2 vocabulary acquisition: The role of L1 lexical/conceptual knowledge. *Applied Linguistics, 27*, 741–747.

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press

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., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition, 27*, 567–595.

https://doi.org/10.1017/S014271641200029X Published online by Cambridge University Press