タイトル	Report on a free continuous word association test
著者	MUNBY, Ian
引用	北海学園大学学園論集, 132: 43-78
発行日	2007-06-00

Report on a free continuous word association test

Ian MUNBY

ABSTRACT

It seems logical to suggest that developments in a learner's lexical competence would be mirrored in the number and type of associations that a learner could produce in response to a set of prompt words, such as *comfort*. During the early 1980s, some commentators had assumed that a test could therefore be designed to measure the state of an L2 learner's associational networks which would reflect level of proficiency. However, in 1987, a report of a free continuous word association probe was published by Kruse, Pankhurst and Sharwood-Smith in the journal *Studies in Second Language Acquisition*. This study compared the associations produced by a group of 15 Dutch third-year university students of English with a group of 7 native speakers of English in a test which used a specially designed software to collect up to 12 responses for each of a set of 9 stimulus words. Their responses were measured using three scoring systems: i) weighted stereotypy, ii) non-weighted stereotypy -each based on a norms list- and iii) number of responses entered. In all three measures, no significant difference between the two groups was reported. The experiment was hugely influential since it seemed to prove that the free continuous word association test (WAT) was inadequate as a proficiency measure. This paper will report on a replication of the same study wherein a group of 50 native speakers performed significantly better on average than 45 Japanese learners of English. However, it is suggested that the above conclusion may be unfounded due to the number of serious flaws in the test itself.

INTRODUCTION

Since Richards (1976) and Meara (1980) first commented that the importance of the teaching and learning of vocabulary in foreign/second language learning had been seriously underestimated, a flurry of interest in the field has taken place. Schmitt (1997) isolates several key areas of heightened research interest including the size and growth of lexicons,

the number of words learned through incidental exposure, and the effect of exercise type and strategies on the lexical acquisition process. Even though the quantity of studies in vocabulary acquisition (VA) indicate that the development of lexical competence has moved towards center stage of second language acquisition (SLA) studies, he also notes that no concrete theory of how new words and phrases find their place in the learner's mental lexicon has emerged.

There appear to be several problems with research in the VA field including the following three. First, since lexical development is so enormously multi-faceted or multi-dimensional, simple models cannot be expected, particularly in view of the number of largely uncontrollable variables, such as learner aptitude and motivation, which also complicate research into all fields of SLA research. Second, quantity of research has not been matched by quality, a phenomenon observed by Wolter (2005: 14) with specific reference to research into depth of vocabulary knowledge.

A final problem is that there have been specific areas of comparative neglect in VA studies and one of these is word association (WA). Sinopalnikova (2003) defines WA as follows. "Originally the term 'association' was used in psycholinguistics to refer to the connection or relation between ideas, concepts, or words, which exists in the human mind and manifests in the following way: an appearance of one entity entails the appearance of the other in the mind; thus 'word association' being an association between words" (p. 199). Typically, a productive free word association test (WAT) involves inviting participants to supply single or multiple responses to a set of stimuli, or cue words.

This neglect of WA is surprising for the following two reasons. First, knowing a word's associations is generally accepted to be a crucial component of word knowledge, or actually knowing a word. Richards (1976) included knowing the "network of associations between that word and other words in language" (p. 81) as the sixth of eight word knowledge categories, or assumptions. However, his conclusion on this assumption focusses on the way words are stored in the mind "according to associative bonds" (p. 87), or how words are learned and remembered, rather than whether or not knowing a word's associations actually constitutes word knowledge. This is an important consideration since associational knowledge differs from other forms of word knowledge in that it is not declarative knowledge, such as orthographic or morpholigical knowledge. Meara (1996) also finds the sixth assumption

to be an exception to the set since it is not "driven exclusively by the concerns of descriptive linguistics, rather than by psycholinguistic concerns" (p. 3) yet he considers the set to be the weaker for it since associational knowledge may explain how word knowledge is acquired while most descriptive knowledge does not.

Nation (2001) also includes knowledge of associations as an aspect of both productive and receptive word knowledge, perhaps facilitating or lubricating language use. From this angle it could be concluded that WA is an aspect of VA studies that cannot be ignored. Laufer (2001) has also suggested that tests of "awareness of a word's syntagmatic and paradigmatic relations should not be underestimated" (p. 242). Nevertheless, she also insists that the following three tests: receptive vocabulary size (Vocabulary Levels Test, Nation, 1983), tests of controlled productive ability (Laufer and Nation, 1999) and lexical richness in composition (Lexical Frequency Profile, Laufer and Nation 1995) are the most basic measures of lexical proficiency, implying a more limited role for word association tests.

The second cause for surprise is the fact that word association studies in first language learning can boast a long, rich, and varied history, which other areas of lexical research, such as vocabulary size, cannot. Normative data, in the form of norms lists of associations typically produced by native speakers, has been collected throughout the preceding century. For instance, the Minnesota Word Association Norms (Jenkins, 1952), based on the Kent-Rosanoff lists of target words (1910), and published in a collection of research in the field by Postman and Keppel (1970), was a prime example of research into native speaker associational knowledge. Research into the associations of bilinguals and multilinguals also experimented with a number of different WATs and the findings of this research are often relevant and applicable to studies into L2 learners' associations. Also, after 1970, a small number of studies into L2 learners' associations were conducted until 1987, when a significant decline in interest is observable in the literature, at least in the mainstream. This could be due to the influence of the study by Kruse, Pankhurst and Sharwood-Smith, discussed earlier in the abstract.

However, it may not seem so surprising that attempts to correlate WA knowledge and L2 proficiency have met with little success. While many commentators, such as Harley, regard L2 vocabulary knowledge as "fundamental to the development of L2 proficiency" (1996: 3), there is little conclusive evidence that growth in the L2 learner's lexical store

typically involves a parallel, and readily measurable, growth in the number, variety, and strength of links between words and concepts therein, finally approaching native speaker associative behaviour. Nevertheless, Zareva et al (2005), in a paper which directly addresses the issue of the relationship between lexical competence and proficiency, used a multiple test battery including a word association test to successfully distinguish groups of leaners at Intermediate and Advanced level.

LITERATURE REVIEW

Recent work into WA has been motivated by a variety of research goals using a broad range of data-gathering instruments making it difficult to categorize and identify common trends. Some of this research, testing receptive knowledge, looks into informant ability to make links or identify pairs of associates from sets which include distractors. Examples include the Word Associates Test (Read, 2000) and V_links (Meara and Wolter, 2004). Other research investigates productive WA behaviour by classifying response type, for example, which some commentators (Söderman, 1989; Orita, 2001) have found to shift from syntagmatic to paradigmatic as the learner's proficiency develops. However, recent research by Fitzpatrick (2006) and Nissen and Henriksen (2006) have produced evidence to the contrary. The subject of this inquiry concerns measuring productive word association performance using norming lists, where L2 learners' responses are reduced to raw scores according to their frequency status, or stereotypy ranking. These stereotypy measures are then correlated to an additional proficiency measure with a view to establishing the validity of the productive WAT as a means of assessing global L2 proficiency.

The following is a short summary of some of the findings of research relevant to the third strand. Research by Lambert (1956) showed that the number of responses in WAT was related to proficiency among American students of French, with development towards native speaker response patterns. Later, Riegel and Zivian (1972) investigated both free and restricted, intralingual and interlingual associations of American students of German. They emphasized the need to explore the psycholinguistic processes of second language learning. Crucially, they entertain the notion that the number, variability, and speed of learner responses to target words are dependant on the state of the learner's "repertoire", or the availability of items in the L2 lexicon. In their study, the authors point out that variability, not speed, of response is investigated. 24 subjects were invited to write down as many words as they could think of in English and then in German in two three-minute periods. They

found that, on average, subjects listed 66 English words and 31 German words in each 3-minute session. In other words, items in the L2 lexical store of this particular group of subjects are 47% less accessible than items in the L1 store.

By introducing a 3-minute time limit in the first phase of the tests, a verbal fluency test, they are in fact measuring the accessibility of items in the lexical store. Given that they appreciate the importance of timed conditions, it is therefore surprising that no time conditions were applied to subsequent tasks in the experiment. However, a link between language proficiency and word association is strongly hinted when comparing the French to English (FE) associations between subjects who had some knowledge of French and those who professed none. For example, it is observed that: "the strategy of producing responses identical with the stimuli in their initials becomes less important with language proficiency" (p. 57).

Meara (1983) noted that the study of word associations over time has the potential to track the absorption or integration process of new words in the lexicon until they find their proper place. Reporting on an earlier study (1978) to investigate this phenomenon, a group of English students studying French at A-level were given a list of 40 French words that they were unlikely to know. Unsurprisingly, a low number of associations was elicited including few native like responses. Half of these words were then taught to the students and a subsequent WAT and WA retest evoked "an increase in the proportion of native-like responses". While Meara observed that L2 learner responses are in "a state of flux" and may not represent any permanent feature of learner lexical knowledge, the hint of a link to proficiency, as evidenced by an increase in nativeness, is still implied.

In a longitudinal study, Randall (1980) aimed to "investigate the changes taking place in the learners' associations over time in order to try to see how the word store of a learner was different from the word store of a native speaker, and to see if the word store in fact became more like that of the native speaker as exposure to the language increased" (p. B9). To this end, he investigated the development of word associations of a group of 26 upper intermediate level students on intensive courses (Cambridge First Certificate preparation courses) at five different languages schools in the UK on two occasions with a nine-week interval. Multiple responses to 50 of the 100 words on the Kent-Rosanoff list were elicited and measured according to the Minnesota norms lists. These 50 cue words, drawn at random, were divided

into three groups (type 1-3) depending on their norm characteristics, with type 1 words exhibiting high primary response proportions. Stimuli were delivered orally to the subjects, who were tested as a group, with a thirty-second time limit per stimuli to write up to 5 responses per word. Scores from each of these tests were correlated with a pair of simultaneous language competence tests (the Nelson Language test battery) which ranked students on a scale of 10 levels from beginner to advanced. The first of these tests corresponded to the entrance level of an FCE preparation course, and the second corresponded to actual FCE level. The four sections of the tests primarily measured linguistic competence through focus on grammar and lexical knowledge using a multiple-choice gap-filling format and cloze. In addition to totaling the number of responses entered, informant associations were measured for stereotypy according to the Minnesota norms-lists.

Randall concludes that between the two WATs, on average, "total number of responses increases with increasing proficiency in the language" (p. C6). However, when the subjects are ranked individually according to FCE grades and the Nelson tests, there is only weak support for the claim that proficiency can be correlated with number of responses given. Regarding stereotypy scores, the mean group scores did increase between initial and final test.

The critical feature of Randall's work is the complex system of weighting that he used, wherein a maximum score of 25 (5×5 points) was apparently awarded for a primary response which is also listed as a primary response on the norms list. This formed the basis of a suspect weighted stereotypy scoring mechanism applied in the following study.

Kruse, Pankhurst, and Sharwood-Smith (1987) report on the viability of a word association test as a means of comparing native speaker and L2 learner lexical knowledge, and by extension, as a means of measuring L2 learner proficiency. In preliminary theoretical remarks, the authors comment that learner-produced sets of associations to a prompt word must originate from shared knowledge of the world, including social-cultural knowledge, and knowledge of the L2 lexicon. If L1 and L2 lexicons are organized in the same way, L2 learner responses will signal the state of development of the L2 learner mental lexicon. The problem is that the tests may also, or perhaps exclusively, be measuring socio-cultural knowledge. These are just some of the crucial concerns which shall be discussed later in connection with the 2006 replication of the original 1987 study.

In the 1987 experiment, a multiple free association test computer program displayed and collected a maximum of 12 responses to 10 stimuli: *man, high, sickness, short, fruit, mutton, priest, eating, comfort, and anger*. Responses to *man* were dropped from the final calculations because of a mistake. No restrictions were put on response type and informants were instructed to type in single English words which came to mind in response to the cue appearing on the screen. Each response was removed from the screen immediately after it had been typed in and recorded on text file.

The subjects were 15 third year Dutch English majors, University of Utrecht, and the control group comprised 7 native speakers. The subjects were given 30 seconds for each stimulus word not including the time taken to type in the responses. The timer on the program deactivated while participants typed in their responses so as not to penalize slow typing speed. Test reliability was established by test-retest method, and validity was assessed through correlation with two language proficiency tests. The first was a 50-gap cloze test using a text where every sixth or seventh word had been deleted, in addition to a grammatical error monitoring test. All subjects (including native speakers) also took these tests. The subjects were given the word association test on two separate occasions about two weeks apart, but the control group did the test only once. All subjects took the WAT individually and results were compared with a control group of native speakers. As reported earlier the experiment was based on a test method developed by Randall (1980), who used two methods to analyze results: the number of responses and the degree of stereotypy. Comparison of responses was done on the basis of the Minnesota norms lists (Jenkins) in Postman and Keppel (1970).

The selection of the stimuli was based on the Kent-Rosanoff list (1910). Each of the stimulus words were chosen at random, one each from 10 categories of stimuli of different strengths devised by Den Dulk (1985) according to the Postman-Keppel norms list, a supposed improvement on the three frequency groupings suggested by Randall. The word MAN has a high frequency of primary response-311 out of 1,000. The primary response for ANGER is only 36. The word associations were scored in two different ways: quantatively (number of responses, Tables 1, 2, and 3 row A) and qualitatively (degree of stereotypy, rows B and C). Qualitative scoring measured stereotypy in two different ways

1) Non-weighted stereotypy score. Varying order-related scores were not given. Based on Randall (1980), scores were given according to the number of responses which matched the

norm responses for each cue.

2) Weighted stereotypy score. Each subject's total score was obtained by summing the scores on responses to all nine stimulus words. This was based on specific scores for stereotypic responses according to whether they were native speaker primary, secondary, or subsequent responses. This takes into account the order in which the responses occur in the norms lists producing a weighted stereotypy score. For an example of a scoring grid see Appendix 1, and for an example of one subject's scored responses, (2006) see Appendix 2.

RESULTS

There is only a small difference between native and non-native speaker scores in the WA test (see Table 1). Although it was expected that native speakers would perform better than non-native speakers, the latter group achieved higher scores in Test 2 (re-test) than the former in measure A & C (response scores and weighted stereotypy), but not for measure B (weighted stereotypy).

Table 1
Mean scores, standard deviations, and theoretical maximum for all scoring methods (Kruse, Pankhurst, and Sharwood-Smith, 1987).

	Non-native speakers		Native speakers	
	Test 1 Mean (SD)	Test 2 Mean (SD)	Mean (SD)	Theoretical maximum
A	76.8 (17.9)	82.8 (19.1)	79.9 (14.2)	108
B	23.4 (7.3)	22.9 (5.7)	25.7 (7.2)	108
C	1475 (377)	1542 (337)	1509 (414)	15,552

A=number of responses, B=non-weighted stereotypy, C=weighted stereotypy

Correlations between test and re-test in Table 2 are not high. The highest correlation is achieved on number of response scores (0.759 Test A).

Table 2
Test-retest reliability correlations (Kruse, Pankhurst, and Sharwood-Smith, 1987).

Test A: Number of responses	r=.76**
Test B: Non-weighted stereotypy	r=.66**
Test C: Weighted stereotypy	r=.55*

** $p<.01$, * $p<.05$

Cloze and grammar monitoring were correlated with WA scores for responses, non-weighted stereotypy, and weighted stereotypy (see Table 3). Both exact and acceptable answers on the cloze were counted in a score out a possible 50. The mean scores for both WA tests were used. All scores were converted to standard scores. Correlations between both stereotypy measures and cloze are significant but very weak, although stronger than correlations between "number of responses" scores and cloze.

Table 3
Correlations between the association scores and the proficiency measures (Kruse, Pankhurst, and Sharwood-Smith, 1987).

	Cloze test	Grammar monitoring test
Test A: Number of responses	r=.44*	r=.58*
Test B: Non-weighted stereotypy	r=.55*	r=.30 (n.s.)
Test C: Weighted stereotypy	r=.54*	r=.15 (n.s.)

p<.05*

Strongest correlations emerged between grammar monitoring scores and number of responses. However, there was no correlation between grammar monitoring scores and either of the two stereotypy measures. Mean weighted stereotypy scores for each of the nine stimulus words for the control group (native speakers) and non-native subjects (students) in both Test 1 and re-test (Test 2) are detailed in Table 4.

Table 4
Mean weighted stereotypy scores for each word. (Kruse, Pankhurst, and Sharwood-Smith, 1987. p. 151)

Stimulus	Students Test 1	Students Test 2	Native Speakers
high	181.40	191.14	193.86
sickness	242.33	214.26	284.29
short	152.53	170.60	192.71
fruit	254.13	228.53	188.00
mutton	257.46	311.93	222.71
priest	152.58	198.26	239.14
eating	116.80	113.00	159.71
comfort	143.40	87.73	179.71
anger	55.13	37.86	106.14

Scores tended to be higher for native speakers for each of the nine words. Native speakers scored lower than the learners with two stimuli: *mutton* and *fruit*, which cannot be accounted for by low-stereotype status. This is further evidence that the WA test cannot

distinguish between the two groups. However, where stereotypy is low (eg. *comfort, anger*) there is greater difference between native/non-native speaker scores. The authors describe the results as "disappointing", apart from the test-retest reliability measure for the number of responses of 0.76 ($p<.01$) (Table 2, test A), but they suggest that extending both the number of items and the number of allowed responses may increase this reliability coefficient to 0.9 Reliability measures of WA scores against cloze and grammar-monitoring scores show that this WA experiment cannot measure proficiency reliably. Only the number of responses can measure proficiency to a small degree. This suggests that factors other than language proficiency affect performance on the WA test, perhaps the effects of cultural background knowledge and intelligence.

PRELIMINARY COMMENTS ON THE EXPERIMENT.

This was essentially a one-off study and it is possible that the above conclusions are undermined by a severely flawed approach to measuring both associational knowledge and proficiency. These issues are discussed in connection with the findings of the replication experiment in the discussion section, where a brief description of research findings following 1987 is also included. However, there is no mistaking the influence of the Kruse, Pankhurst, and M. Sharwood-Smith experiment, published in the highly influential journal *Studies in Second Language Acquisition*. This influence manifests itself in two ways. First, the paper has been quoted more often than not in work published in this field (Sanford, 1994; Meara, 1996; Schmitt and Meara, 1997; Schmitt, 1998; Orita, 1999; 2002a, 2002b.; Wolter, 2001, 2002; Zareeva, 2005, Fitzpatrick, 2006). Second, it is possible that some researchers abandoned the field either completely (Randall) or temporarily (Meara) as a result of their findings.

The following is a summary of what was established, or at least what can be more or less safely assumed in the field up until 1987. First, the number of responses entered to stimuli in free WAT increases with proficiency. Second, responses provided by subjects vary in type and number depending on (i) the type and frequency of stimuli, (ii) the language of stimuli and response, for example L1 to L2 English to German or French and vice-versa (Riegel and Zivian, 1972) (iii) the degree of subject familiarity with the stimuli, and (iv) the level of proficiency of the subject. Third, native and non-native speaker associative behaviour differ significantly (Meara, 1983; Söderman,1992).

However, up until 1987, and perhaps even today, a certain amount was and is still not

known. This includes the optimum choice and number of stimuli and how this affects associative behaviour. The optimum number of responses to elicit per cue word is also unclear, as is the notion of assigning weighted and unweighted stereotypy scores based on the Minnesota norms lists, or any norms list. Uncertainty also surrounds the question of what is the best measure of proficiency to link WAT scores to, or whether in fact a measure of lexical competence, such as vocabulary size, would not be more appropriate.

AIMS OF THE REPLICATION STUDY

The main aim of this replication is to attempt to find a link between this kind of free multiple response WAT and proficiency. There follows a discussion of the methodology and conclusions of the original probe with particular emphasis on the appropriacy of the following:

i) Number and level of subjects
ii) Number of responses elicited
iii) Scoring system
iv) Selection of stimuli
v) Norms lists
vi) Test procedure
vii) Proficiency tests

Further questions include:

a) Is the continuous free WAT a valid test of lexical competence, language skill, or proficiency?
b) Should a clear link between knowledge of associations and proficiency be expected?
c) If so, what dimension, or dimensions, or aspects of lexical competence does the continuous free WAT measure?
d) Are the L1 and L2 lexicon structured in a similar way?
e) How could the test be improved?

THE REPLICATION. STUDY METHOD.

It has to be said that exact replication was not possible in view of the following three circumstances. To begin with, the original proficiency tests were unavailable. Efforts to track them down began with contact with one of the original authors, Michael Sharwood-Smith, which led in turn to a contact with another, James Pankhurst. It transpired that, at

the time at least, Utrecht University destroyed all copies of master's theses after 5 years' shelf life. The published report was based on an unpublished masters dissertation by Heleen Kruse, which disappeared along with important details of the scoring system. The second circumstance was the fact that the original software was also unavailable, although it is interesting to note that James Pankhurst very kindly cooperated in reproducing a version of the software similar to, and, according to him, better than the one he had created nearly 20 years earlier. The third key circumstance affecting replicability of the experiment was the unavailability of a significant number of advanced level subjects, at least not as high as the Dutch subjects in the original study.

The situation was remedied in the following manner. First, a 50 item multiple choice grammatical error recognition test from part V of the old (2005) TOEIC test was used, with examples taken from published materials. This section has been dropped from the updated, revised version of the TOEIC test (2006). A self-designed cloze test, with 50 gaps and with every sixth or seventh word deleted (as in the original) was used instead of any published cloze test because they were much too difficult for my subjects and appeared likely to seriously compromise their discriminatory power. See Appendix 3. Finally, the word association test software was designed according to specifications which matched the original as closely as possible and created by Paul Meara to whom the author remains eternally indebted.

Subjects

The 45 non-native subjects (15 male, and 30 female) were all Japanese, and nearly all of them were English majors at Hokkai Gakuen University, Sapporo, Japan. They included 18 first year students, 6 second years, and 18 third and fourth years. 3 professionals (2 ELT publishers' representatives and one high school teacher) also took the test. With the exception of the first year students, who took the test in class time, all students had volunteered to take the tests and were not paid for their services. Of the 45, 9 subjects took the tests individually, and the remainder took the tests as groups in a computer laboratory, but in all cases there was a two-week interval between test and re-test and none of the subjects knew they would be taking the test again.

The native speaker control group comprised 50 unpaid volunteer subjects (35 male and 15 female), with an average age of about 45. By nationality, the breakdown was: USA (25),

Canada (12), UK (9), Australia (4) and all subjects, with the exception of one, were, or had been teachers of English living in Japan.

Test instructions and procedure.

All subjects were invited to type in a maximum of 12 responses to a set of 10 stimuli, as in the original 1987 experiment. Subjects did not receive any training in the different response types available, but were simply told to type in any words which the stimuli made them think of. Subjects were also advised to (i) type in as many words as they could think of, up to 12, within the time allowed of thirty seconds, (ii) refrain from dictionary use, (iii) type in single, English words, (iv) to avoid proper nouns if possible, and (v) to notice that the timer would deactivate while they were typing, allowing them 30 seconds pure thinking time. They were also informed that that there were no right or wrong anwers. The software allowed two practice items, *lake* and *alien*, before beginning the real test. All native subjects in the control group (native speakers) took the WAT once only, as in the original test, but only 12 took the proficiency tests. The testing sequence was a twin session format of WAT1 followed immediately by the cloze test, then WAT2 (re-test after a two week interval) followed immediately by the grammar monitoring test. 30 minutes were allowed for each of the proficiency tests. None of the non-native subjects knew that they would take the same WAT again.

Scoring

Although the scores for the first stimulus *man* were dropped from the scoring, as in the original, all subjects entered responses for this cue, together with responses for two pre-test practice items: *alien* and *lake*. Scores on the WAT appeared in text file but were transferred to Excel files where each response were scored by hand, as in the original, using the norms lists for non-weighted stereotypy, and a scoring grid for each of the stimuli for weighted stereotypy (see Appendix 1), for an example. Spelling was not penalized and misspelled items which were recognizable as items on the norms lists were accepted. Stimulus words entered as responses for the same prompt were discounted, as were responses entered more than once for the same stimulus. Plural forms of items on the norms lists were also accepted, but responses were not lemmatized so as to reflect the organization of the original lists. In other words, *irritability* appears on the norms list as a response to *anger*, but *irritated* does not, and was therefore not counted as scoring in the stereotypy measure. All scores were summed by Excel auto-sum. The cloze tests were marked by the author using

a bank of possible answers which were either supplied by non-native test-takers and judged acceptable by myself or generated from the 12 native speakers who took the test.

THE REPLICATION. STUDY RESULTS

The following Tables 1-4 correspond to Tables 1-4 in the original experiment.

Table 1 shows that the native speakers outperform the non-native subjects in all three measures, but performance by native speakers varied considerably

Table 1 (2006)
Mean scores, standard deviations, and theoretical maximum for all scoring methods of the word association test.

	Non-native speakers		Native speakers	
	Test 1 Mean (SD)	Test 2 Mean (SD)	Mean (SD)	Theoretical maximum
A	61.8 (25.3)	68.1 (23.6)	94.6 (12.7)	108
B	32.2 (10.2)	34.7 (8.62)	46.3 (9.86)	108
C	1378 (346)	1460 (310)	2141 (496)	5,971

A=number of responses, B=non-weighted stereotypy, C=weighted stereotypy

Non-native subjects perform better on the re-test than in the first test in all three measures. The theoretical maximum for weighted stereotypy is different from the original experiment because the authors miscalculated the maximum number of points that could be scored. It is not exactly clear how Kruse et al arrived at the figure of 15,552. It appears to be the sum of $9\times12\times144$, representing 9 stimuli multiplied by the maximum 12 responses multiplied by a maximum score of 144 (which can only in fact be achieved for the primary response to each stimulus word). The theoretical maximum in this replication is 5,971, achieved by summing the theoretical maximum for each of the 9 stimulus words, which range from 650 to 691. This difference is due to the distribution of responses in the Postman and Keppel norms lists. For example, with reference to Appendix 1, the responses *chair* and *tall* share rank as the fifth most common response to the stimulus *high* on the norms lists meaning that they must be scored equally.

The cloze test, although designed to be simpler than some similar published ones, caused serious problems with higher level learners, while still allowing lower level learners to score some relatively easy points. Scores ranged from 8 (low) to 34 (high) with a mean of 21 (SD 6.76), almost identical to the cloze results cited in the original experiment. A sample of

TOEIC scores of the university students who took the test indicated a range of 300-860.

Table 2 (2006)
Test-retest reliability correlations

Test A: Number of responses	r=.92*
Test B: Non-weighted stereotypy	r=.86*
Test C: Weighted stereotypy	r=.78*

* p<.01.

Test-retest reliability correlations are much stronger than in the original experiment but the same hierarchy emerges, with highest correlations found with the number of responses measure, followed by non-weighted stereotypy, then weighted stereotypy. This variation could be a reflection of the suitability of the scoring measures themselves, with weighted stereotypy tending to produce the least stable re-test scores.

Table 3 (2006)
Correlations between the association scores and the proficiency measures

	Cloze test	Grammar monitoring test
Test A: Number of responses	r=0.45*	r=0.34*
Test B: Non-weighted stereotypy	r=0.49*	r=0.30*
Test C: Weighted stereotypy	r=0.32*	r=0.16 (n.s.)

* p<.05

In Table 3, correlations between the association scores and the proficiency measures are based on a combination, or average, of test and retest scores. In the original experiment, the number of responses measure had the lowest correlation with cloze test scores (r=0.44). In this experiment, the weighted stereotypy measure has the lowest correlation with cloze test scores (r=0.32). In both original and replication, the weakest correlation is between weighted stereotypy and grammar monitoring test scores.

With reference to Table 4, the non-native subjects (students) outscored the native speakers in Test 2 with the weighted stereotypy score for the word *fruit*. In the original probe, non-natives outscored natives for *fruit* and *mutton* in both Test 1 and 2. Scores for *comfort* and *anger* were among the lowest in both original and replication, which makes us question the use of native-speaker based norms lists at all. While Kruse et al suggest that this is an indication of the absence of discriminatory power of the WAT, it could also be

viewed as evidence of testing system weakness.

Table 4 (2006)
Mean weighted stereotypy scores for each stimulus word.

Stimulus	Students Test 1	Students Test 2	Native Speakers
high	213.56	220.84	253.68
sickness	155.36	160.20	247.28
short	158.42	183.07	211.40
fruit	260.67	274.13	268.92
mutton	115.31	147.20	273.18
priest	87.29	117.67	246.68
eating	169.87	163.96	219.66
comfort	99.91	118.31	170.92
anger	77.16	74.422	161.20

Table 5 (2006)
Correlations between non-native subject cloze scores (Test 1, 2006) and WAT performance per stimulus word.

Stimulus	No. responses	Stereotypy measures	
		Non-weighted	Weighted
high	0.32*	0.33*	-0.09 (n.s.)
sickness	0.48**	0.44**	0.39**
short	0.23 (n.s.)	0.23 (n.s.)	-0.02 (n.s.)
fruit	0.47**	0.23*	0.04 (n.s.)
mutton	0.24 (n.s.)	0.03 (n.s.)	0.15 (n.s.)
priest	0.36**	0.39**	0.30*
eating	0.33*	0.29*	-0.06 (n.s.)
comfort	0.35*	0.19 (n.s.)	0.09 (n.s.)
anger	0.48**	0.40**	0.35**

1-sided p-value　* $p=<0.05$　** $p=<0.01$

However, when non-native subject WAT scores for each stimulus word (Test 1) are correlated with their cloze test scores, the most effective of the two stereotypy measures (see Table 5) it is apparent that some cues are substantially more effective than others. Correlations were strongest with the following cues: *sickness* and *anger* in all three measures. This is clear evidence that success in finding a link between proficiency and WAT performance depends to a large extent on the stimulus words chosen. The weighted stereotypy measure was especially weak and there was little apparent difference between the performance of subjects of different levels - low and high intermediate - with some prompt words such as

high, *short*, and *fruit*. This is an important concern which shall be discussed later in connection with the selection of prompt words.

DISCUSSION

It is quite clear that there are simply too many flaws in the 1987 study and lend credibility to the suspicion voiced by Schmitt (1998) that early research into WA has suffered from "unsophisticated methodology" (p. 400). These flaws concern the following seven features of the test: the number and level of the subjects, the number of responses elicited, the scoring system, the selection of stimuli, the norms lists, test procedure, and the proficiency tests. Each one shall be discussed in turn.

i) Number and level of subjects.

Wolter (2003) was correct to point out that the number of subjects, both native (N=7) and non-native (N=15), in the original experiment was simply too small to justify the kind of conclusions they wished to draw. In contrast, the number of native subjects in the replication (N=50) turned out to be higher than necessary. This issue is discussed again later in relation to norms lists (see Table 6). Finally, the discriminatory power of the test in the original experiment is not fully explored because it was done with a group of L2 subjects at a similar level of proficiency. However, it remains plausible that this test does not have the power to discriminate advanced level subjects from native speakers. Since free WAT performance would surely vary across different levels, a fact recently supported by Zareva (2005), it would have been more sensible to test subjects from a variety of levels, from lower to upper intermediate, as in this replication.

ii) Number of responses elicited

This replication also confirms the findings of Riegel and Zivian (1972), Randall (1980) and Kruse et al (1987) that the number of responses entered is the most effective measure of multiple response free WAT performance. However, as Randall points out, it is a crude measure. For example, one non-native subject scored a healthy-looking maximum 12 points for number of responses in response to *fruit*: piano, violin, instruments, play, harmonica, guitar, forest, unicorn, long, stick, beautiful, orchestra. One can assume that this subject had suffered a reading miscue and had entered responses for the word *flute*, and did not make the same mistake in the re-test. But there is an additional element to "number of response crudity", namely the difference in the number of responses entered within the 30-second time

limit is not accurately assessed because no more than 12 responses can be entered. Many of the native subjects entered the full compliment within even half the time allowed. Some of the non-native subjects also supplied 12 responses within the 30-second time limit. Nevertheless, it seems that the native speakers in the Utrecht control group performed very poorly on the WAT, with particular regard to the mean number of responses they produced (79.9) compared with 94.6 in the replication control group. It seems that Kruse et al were able to, in the words of Wolter (2002), "close the door on" the WA-proficiency link based primarily on what looks suspiciously like a poor performance by a small number of English native-speakers who were apparently working as lab technicians at Utrecht University.

iii) Scoring system

There also seems to be a huge problem with the scoring system especially concerning weighted and even non-weighted stereotypy. It seems likely that the norms lists may have been used in a slightly different way because the mean scores of the subjects in the replication seem to be higher than in the original (Table 1). I am not even sure if I was literally on the same page of the Postman & Keppel norms as Kruse et al who mention: "The word *man* … has a very high frequency for the primary response-311- whereas the second and following responses had relatively low frequency. The primary response to *anger* was only 36" (page 145). These figures have been taken from Kenneth Miller's norms lists (pages 41-52 of Postman & Keppel) which compared responses of 400 English students with responses of 200 Australians. These lists only provide the first 5 most frequent responses to the Kent-Rosanoff norms list and could not therefore have been used for scoring lists of up to 12 responses. The norming lists in this replication are based on Jenkins (1952, in Postman & Keppel, Norms of Word Association, 1970, pages 9-37) which Kruse et al surely must have used, and where the primary response for *man* is *woman*, entered by where 767 out of 1000 respondents (p. 25), not 311.

It is almost certainly unwise to award a maximum of 144 points for supplying a primary response which matches the most frequent response on the norms lists, while only giving one point for a low stereotypy twelfth response. Randall (1980) questions the practice in a self-critical way which eluded the authors in Utrecht. He rightly points out that the scoring scheme places too much emphasis on order of responses rather than the clusters themselves. Further, even if we do accept that degree of stereotypy must feature in the scoring, the rating of responses on a 12×12 scale fails to reflect the reality of differences in stereotypy in the

norms lists. For example, as evidenced in Appendix 1, incidences on the norms list out of 1000 responses (which appear in brackets after each scoring response) are severely disproportionate to their weighting on the stereotypy scale. For example, with reference to Appendix 1, a subject who, in response to *high*, supplies *low, school*, and *mountain* as first, second and third responses will score 144, 121, and 100 points respectively for each word on the weighted stereotypy scale which does not reflect the response distribution of these items on the lists (675, 49, and 32).

Also, clang responses, such as *priestmentalization*, should probably not have been scored on the "number of response" scale in this replication, but I have no idea how these responses were treated in the original experiment, or even if they occurred at all. In future, I would be inclined to discount them, but allow for lemmatized responses to score on the stereotypy measures wherever possible to avoid situations where subjects narrowly miss being awarded points for responses such as *irritable*, which are not on the norms lists, while *irrititability* is.

iv) Selection of stimuli.

Another serious problem with this probe is that results depend on stimuli chosen and this weakness is not admitted by the authors. While trumpeting the fact that non-natives outscored natives with two stimuli (*mutton* and *fruit*, Table 4) as evidence that the WAT is inadequate as a measure of second language ability, Kruse et al also seem to have ignored words of warning from Randall (1980). He points out that approach to native-like associative behaviour may manifest itself through a divergence from, rather than convergence on, norms. It is easy to see how this happens with responses to words like *fruit*. While simply producing a list of names of fruit seemed to typify lower level subjects, more expert learners and native speakers produced more syntagmatic, or position-based responses such as *fly*. Research by Fitzpatrick (2006) indicates that native subjects prefer syntagmatic responses, which is possibly how native speakers found themselves outscored on average in the weighted stereotypy measure with weighted stereotypy scores for the cue *fruit* (See Table 4). It must also be pointed out that Meara (1983) had gone into some detail about the issue before Kruse et al (1987) had carried out their probe. Since the paper was published in a minor journal- the *Nottingham Linguistics Circular*- it would be safe to assume that the authors had not read it.

Regarding stimuli selection criteria, assuming predictable responses are best avoided,

Meara (1983) warns against lists which contain:

a) very high frequency words which produce what Meara and Fitzpatrick (2000) describe as "dominant primary" responses, such as *man-woman* (p. 22)

b) adjectives and other words which produce their polar opposites (eg. *high*)

c) nouns which are marked for sex which tend to produce the opposite sex in response (eg. *man*)

To this I would add:

d) superordinates involving simple sets such as *fruit, colour*. In this replication, *Fruit* produced sets of responses which included mostly high-scoring names of fruit, especially in lower level subjects.

e) recurrent similar concepts (eg. food concepts: *fruit, mutton, eating*). Some subjects entered the same response for different stimuli, such as *apple* or *mutton* for *eating*.

f) items which can be easily confused with similar words (eg. *mutton/mouton*. See Appendix 2)

g) (For Japanese subjects) words which are used as loan words with different meanings (eg. *mutton-mouton*, which is some kind of winter garment).

h) words unknown to many of the subjects taking the tests (eg. *priest*)

i) words which generate few low frequency responses since these fail to generate scores which differentiate performance by learners of different levels of proficiency.

j) too many concrete nouns (Fitzpatrick, 2006: 128)

Wolter (2002) also found that some items such as *travel*, and all the items used in Lex 30, which were chosen specifically for that reason, elicit a large number of idiosyncratic responses. These are responses which appear only once on the norms lists. Clearly, in future studies, it is crucial to find a solid, criteria-based list of features that could predict the effectiveness of a prompt word in discriminating WA performance across different levels. In this test, only two of the 10 cue words satisfy the selection criteria: *sickness* and *anger*. Clearly, establishing the WAT-proficiency link will depend to a great extent on satisfactory stimulus selection if the weighted stereotypy measure is not abandoned or significantly restructured.

v) Norms lists

One of the greatest shortcomings of this experiment is that the norms lists were hopelessly inadequate. Ideally, norms lists should be based on data drawn from a comparable group taking the same test (Schmitt, 1998). A reliable norms list could be compiled

perhaps from as few as 25 native subjects, or half of the native control group in this study. Table 6, a reconstruction of the data in Table 1 (2006) shows little difference between mean scores of the first 25 subjects and corresponding scores for the whole group (N=50).

Table 6 (Munby)
Mean scores, standard deviations, and theoretical maximum for all scoring methods of the word association test for the first 25 native speaker subjects and for the group taken as a whole.

	Native speakers		
	Mean for the first 25 subjects	Mean (SD) For the whole group of 50	Theoretical maximum
A	91.6	94.6 (12.7)	108
B	45.4	46.3 (9.86)	108
C	2036	2141 (496)	5,971

A=number of responses, B=non-weighted stereotypy, C=weighted stereotypy

The data suggests a certain level of homogeneity in native subject associative behaviour. While Meara (1983) recommends dropping the Minnesota single response-based lists altogether, Kruse et al justify their use on the grounds that single response and multiple response WAT have been shown to produce similar results. This is not logical. For example, the 50 native speakers produced a total of 511 responses for the stimuli *priest* and a large number of negative associations were entered. These included items such as: *abuse, boy, scandal, homosexual*, and *bugger*. However, responses of this type were never primary - indeed they usually only appeared after the first five or six associations had been entered - and suggest some form of remote (or posterior) association.

This is not to say that negative associations for *priest* did not appear on the Minnesota norms list. There are in fact about 4 out of 1000, including *not good*, *jerk*, and *queer*, but the point is that the single response WAT do not have the ability to test more distant links in a subject's semantic network in a way that a multiple response WAT does.

One could argue that priests have not changed, but the general public's opinion or knowledge of them certainly has, ample testimony to the fact that norms lists must be current. In this way, the claim of Kruse et al that the WAT only measures socio-cultural knowledge is another very good reason to support the notion that the Minnesota norms lists must be abandoned (and a good reason that they should not have used them in the first place).

While none of the non-native subjects entered negative associations for *priest*, common responses from both groups to some stimuli included items which were not on the norms lists for the simple reason that they were either little known or did not exist at the time of compilation (1952). Examples include *fruit- kiwi*, *short- skirt*, *mutton- barbecue*. In addition, in both groups, *genghis khan* was a very common response to *mutton* because it happens to be the name of a mutton dish for which the island of Hokkaido is famous. It does not appear in the Minnesota norms lists.

Leaving the socio-cultural mismatch of norms and responses aside, it was disappointing to find a large number of common associations in the data which did not feature on the norms lists. With reference to Appendix 2, *cost* and *level* were common associations, collocations in fact, for the stimuli *high*, but they are non-scoring.

vi) Test procedure and test-retest

In general, subjects perform better on the re-test and Table 1, in both the original and the replication, shows that these higher Test 2 scores occur almost to the same degree. We need to account for this. Many subjects had explained that the test was easier the second time, where they were able to remember and enter the same responses as in the first test, together with a few extra ones that they found time to type in. While Kruse et al (1987) comment that discrepancies in test-retest correlation is another weakness in the test itself, I would instead suggest that this is to be expected for this type of test/re-test situation. However, it's possible that subjects need more practice, even training in free word association (particularly in response type) in order to perform to the best of their ability. Indeed, experiments to measure differences between an association-trained experimental group and an untrained control group would be interesting. In any event, having two practice items before beginning the test is likely insufficient.

I also notice that some subjects supplied responses for prompts like *mutton* and *priest* in the re-test, but not in Test 1. I imagine they had wanted to know the meanings of these items after Test 1 and had checked with friends or dictionaries, even though they did not know that they would be taking the same test again. It is also interesting that native speakers were not invited to take the test again. If they had been, in both the original and the replication, it is possible that the mean scores for all three measures would have been higher. If we therefore discount non-native speaker Test 2 scores from Table 1 (Kruse), the

7 native speakers in the control group could be said to outscore the non-native subjects in all three measures, albeit only slightly.

vii) Proficiency tests

While the authors comment that the correlations in Table 3 are "surprising", closer inspection shows that they have misread their own data. They state (p. 150) that "it was expected that the [number of] response scores would produce a relatively high correlation with cloze scores; they turned out to be the lowest". No reason is given for the prediction, but closer inspection of Table 3 reveals that correlation between weighted stereotypy scores and the grammar test are the lowest, not the correlations between number of response scores and the cloze scores. Indeed the correlations are low but again this may be due to a broader problem with the testing instrument and the small number of subjects.

Table 3 in both the original and the replication experiment indicates that the cloze test is a more reliable measure of WAT performance than the grammar monitoring test. There are two key problems here. First, the approach of using a cloze test and grammar-monitoring test as a proficiency measure is inappropriate to a large degree. Kruse et al comment that reliability measures of WAT scores against cloze and grammar-monitoring scores (Table 3) show that this WA experiment cannot measure proficiency reliably. However, it's not difficult to argue that neither the cloze, nor the grammar-monitoring test, nor even a combination of the two provide reliable measures of the level of proficiency of these learners. While it is generally accepted that the cloze test does measure a number of elements of linguistic competence (Fotos, 1991), it is highly questionable that this represents a reliable overall measure of proficiency. If the cloze test were an accurate measure of proficiency, then of course there would be no need for an additional measure of grammatical knowledge. Further, to my knowledge, there is no evidence in the literature to suggest a clear link between grammatical and lexical knowledge, hence the disastrous correlation figures.

A key problem is that language production and fluency, or communicative competence, are not measured in either of these two proficiency tests. In the context of language fluency, individual scores on the proficiency tests were often at odds with my personal estimation of the level of the subjects taking them. In brief, some third and fourth year students seemed to compensate for poor scores in the cloze and grammar-monitoring by performing well in the

WAT. I'm speaking here of my own students who were able to articulate (in L2) their dissatisfaction with the style of proficiency measurement in this experiment. Some first year students, who can hardly communicate in English, performed relatively well on traditional, non-communicative, error-focused tests.

viii) The software and the timer

Generally the software was suitable for its intended purpose. The main problem is that some non-native subjects appeared to linger between completion of the process of typing a response and pressing "enter". It's possible that they were checking spelling or typing, but the danger is that the subject can in fact "beat the clock" by using this timer-deactivated pause to think of the next response. The solution could be to make alterations to the software so that any pause of more than 3 seconds causes the timer to reactivate. Even here, it is possible to deactivate the timer and buy more thinking time by deliberately rocking the keys. If we accept that accessibility is a dimension of lexical competence, then the timer is crucially important in measuring the speed with which subjects can produce responses. This is a key measure of fluency of language production which is assessed in speaking and writing components of high stakes tests such as IELTS and now in the new computer-based TOEIC. Although ETS, the creators of TOEIC, had always claimed that scores on their original TOEIC test (listening, reading, and grammar) accurately reflect the level of testee proficiency, the new TOEIC test includes a speaking and writing component which requires test-takers to produce language orally and written under timed conditions. The new TOEIC test seems more likely to reflect L2 learner communicative competence than both the earlier version of TOEIC and the proficiency measuring instruments in the Kruse, Pankhurst and Sharwood-Smith study and this replication. Interestingly, testee typing speed will affect perfomance in the new TOEIC tests, but ETS claim that it is a fair measure of "real world" language ability since the test assesses test-takers ability to participate in the international workplace, and the speed with which e-mails can be written (one of the new TOEIC test tasks) is a key factor. Further, slow typists will also now be penalised in the way that slow speakers, in both L1 and L2, have always been in the speaking component of the IELTS test and in the Cambridge suite of English examinations (PET, FCE).

CONCLUSIONS

In sum, it is not possible to say either if there is or if there isn't a clear link between free WAT performance and L2 proficiency because there are too many problems with the testing

instrument. However, it is true to say that socio-cultural knowledge is being measured, and this may interfere with the scoring unless the norms lists are drawn from similar populations.

In response to the research questions listed in "Aims of the study", the following are important issues that concern what remains unknown about word association and the free continuous WAT

a) Is the continuous free WAT a valid test of lexical competence, language skill, or proficiency?

An improved test that took into account all of the issues listed above may reveal a clearer link. This may finally only shed light on the associative behavior of a certain group of learners (eg. Japanese university students of English living in Hokkaido) with a limited set of stimuli. The same improved test would likely produce entirely different results from a group of learners from a different background because socio-cultural knowledge is undeniably a key factor influencing WAT performance, as commented by Kruse, Pankhurst and Sharwood-Smith (1987: 142). For example, as mentioned earlier, non-native subjects would score points for responses such as *genghis khan* for the stimulus *mutton* if measured by a norms list drawn from native speakers who shared the same cultural background or environment (Hokkaido, Japan). Evidence from the original Postman & Keppel lists (1970) comparing responses produced by Australians and British seems to bear this out in the same way. For example, the primary response to *mutton* was *chop* (111 out of 200) in the Australian group, but *chop* does not feature among the top 5 responses to mutton entered by the English group of 400 (Miller, 1970, pp. 39-47). The same socio-culturally dependent variation, or lack of commonality, is evident in discrepancies between response hierarchies on the EAT (Edinburgh Associative Thesaurus, Kiss et al, 1973) and the Postman & Keppel norms. For example, *mad* (353 out of 1000) is the primary response to *anger* in the latter lists but does not feature at all on the EAT list of top 15 responses, probably due to linguistic or cultural variation.

Another problem is that, since a large number of responses (a maximum of 12) is being elicited, chaining will inevitably occur. Chained responses are responses to the preceding response, not the original stimuli. There seemed to be plenty of evidence of this in the replication from both native speakers (eg. the stimulus *high* elicits *pot, party, oldfriends, montreal, ontariostreet, jazzfestival, comedyfestival*) and in non-native speakers (*eating* elicits *drinking, alcohol, wine, beer, whisky, high, expensive, bar, dark, narrow, sing*). However, it

is not obvious how the problem of chaining can be contained or eliminated by repeating the stimuli to the subjects orally (Randall, 1980), or having them displayed on the screen in front of the subject and removing each response as it is entered (this study), or having the prompt word printed repeatedly next to each blank space where responses should be entered (Wolter, 2002). Also, responses which appear to be simply "chained" may in fact be evoked by a combination of the prompt word and the preceding response.

Further, this kind of free WAT may be asking subjects to resist the natural "chained" direction of cue-response stimulation. The native subject with the highest weighted stereotypy score in this replication (3765 out of a possible 5971), a graduate of Princeton with a high IQ and interest in semantics, later reported that he had performed well by (i) disbelieving my assertion that there were no right or wrong answers (ii) avoiding all idiosyncratic responses, and (iii) exercising cognitive control to resist this "outward g-force" or "tide of lexical network activation" which had swept some subjects so far away from the original prompt. Attempting to resist this tide and "returning to base" may in part explain why some subjects became confused and entered the same response twice, even three times, to the same stimulus word. In this sense, there could be a fundamental mismatch between the way words are activated in the lexicon and this method of simulating its structure.

To return to the theory that knowing a word's associations is an important aspect of word knowledge (Nation, 2001: 26-28), it is not clear how safe it is to assume that a subject who cannot provide any associations to a stimulus word has no knowledge of the word. In this probe, there were two puzzling instances of subjects entering scoring responses to a cue in Test 1, but who failed to enter any responses at all to the same cue in Test 2. Further, with reference to the Appendix 2, the subject's stereotypy score for *mutton* indicates a degree of word knowledge whereas close examination of the responses themselves suggests that the subject does not know that mutton is a kind of meat, and has confused it with some woolen garment. There were several other similar instances of confusion with the cue *mutton*.

b) Should a clear link between knowledge of associations and proficiency be expected?

Nation and Meara (2002) suggest a "close relationship between how many words you know and ... how well you perform ... on other formal tests of your English ability" (p. 50). Similarly, Nation (2001) implies that knowledge of word families "will increase as proficiency develops" (p. 47). Admittedly, this could simply apply to morphological or semantic knowl-

edge rather than knowledge of a word's associations. However, Read (2000) cautions against expectations of a strong link between general language level and vocabulary knowledge in commenting: "being proficient in a language is not just a matter of knowing a lot of words ... but being able to exploit that knowledge effectively for various communicative purposes" (p. 3).

c) What dimensions of lexical competence does the continuous free WAT measure?

First, it has to be pointed out that there is little agreement as to what these dimensions are. In a chapter addressing the issue, Meara (1996), suggests that organization is a dimension, related to, but separate from vocabulary size. Since then, Haastrup and Henriksen (2000) name three dimensions: partial-precise, receptive-productive, and depth of knowledge. Read (2000) finds this kind of distinction confusing and unhelpful, although Zareeva (2004) appears satisfied with this theoretical framework. Fitzpatrick (2006) describes vocabulary knowledge as consisting of three dimensions: "size (or breadth), depth, and accessibility, or organization" (p. 121). This is puzzling since Meara (1996) did not mention accessibility. Also, it could be argued that accessibility and organization are different. It seems possible that L2 vocabulary knowledge could be "normally organized" but suffer from limited accessibility if the learner is poor or slow at recognizing written words, or even unable to read them at all, which is often my experience of reading Japanese kanji. This is why accessibility could be viewed as being a dimension separate from organization. However, I suspect I could still establish connections between Japanese words, spoken orally, in a way consistent with my general level of proficiency in the language, and prove that my Japanese mental lexicon is "normally organized", but with a weak accessibility dimension that inhibits general growth in my Japanese vocabulary size. In this free continuous WAT, entering responses under timed conditions could be a valid measure of lexical accessibility.

Also, it is possible that these tests provide only glimpses of so many aspects of lexical knowledge, including collocation, affixation, morphology, synonymy and antonymy, that the resultant focus may be too broad to be of any value. This is especially depressing in view of the fact that each of these aspects of word knowledge could be assessed more effectively by other tests, such as tests of collocational competence.

d) Are the L1 and L2 lexicon structured in a similar way?

Wray (2002) suggests that they are different. This is of course the subject of a lengthy

debate, but, if the conclusion is correct, using native speaker norms lists to score non-native speaker associations may lack a sound theoretical basis. A bulk of this theory seems to be based on the fact that native speaker responses are homogenous, and this replication suggests that they are not. The alternative would therefore be to use norms lists based on non-native associates. The main thrust of research would then have to be redirected towards mapping subject performance against typical non-native associative behaviour at different levels, although lack of homogeneity may also be a problem here.

Another question needing an answer is this. Are associations between words in the L2 lexicon the cause or the result of growth in the word store? If the forming of new links is the cause of growth in the word store, then the WAT may have the potential to shed light on preferred learning styles, thus opening up a new field of WA research before conclusions have been established in the original areas of inquiry. For example, if a learner's associations contain a large number of idiosyncratic responses (eg. stimulus: *anger*, primary response: *Mrs. T*) then this may indicate that vocabulary strategies that seek to personalize word knowledge may be suitable for some learners.

e) How could the test be improved?

Since the original experiment by Kruse et al (1987), Meara and Fitzpatrick (2000) report strong correlations between Lex 30 (a test of productive vocabulary using a free word association test format) and a yes/no vocabulary size test (Meara and Jones, 1990). Lex 30 measures responses to 30 prompt words for frequency alone without a norming list. It is possible that factoring in word frequency as an additional means of measuring non-native subject associative knowledge would improve correlations. A cursory glance at the non-native data suggests that responses entered by higher level learners include more lower frequency items than responses provided by their lower level peers. It would therefore seem logical to factor this into the scoring to "reward" the more advanced learner, as in Lex 30. The possibility also remained that the non-native group as a whole outscored the natives on weighted stereotypy (Test 2) for their responses to the stimulus word *fruit* (see Table 4) because scoring responses on the norming list for *fruit* were high frequency items. However, when checking the BNC for frequency counts of the top ten scoring responses on the norming list for *high, sickness, short*, and *fruit* it was a surprise to find that the scoring responses for *fruit* were on average much less frequent than in the other three sets. In this way, while factoring a frequency count into the scoring system may be a move of dubious benefit, using

stimuli with a number of low frequency responses on the normative data is clearly very important.

However, Wolter (2002), using a different format, finds only moderate support for the notion that proficiency and WAT performance can be linked despite making substantial improvements to the multiple free word association test. These included using a C-test as a proficiency measure and a larger set of stimuli with a smaller maximum number of responses allowed. A team of native judges replaced the Minnesota norms lists, but this seems to be a rather arbitrary scoring system. His experiment also appears to suffer from the absence of a time limit to measure subject response production speed (number of responses) for each prompt word. In this way, accessibility, one of the three dimensions of lexical competence, of which lexical item retrieval speed is a key component, is not assessed as efficiently as in this probe.

Future WATs of this kind should probably include a vocabulary levels test, such as the Eurocentres Vocabulary Size test (Meara and Jones, 1990) used by Meara and Fitzpatrick (2000) and Fitzpatrick (2006) instead of a general measure of proficiency. An improved test would involve addressing all of the problems highlighted earlier, especially including the development of new sets of stimuli, scoring systems, and norms lists. A test with 30 stimuli might be more effective, with a time limit of 15 seconds per stimulus word (half the time allowed in this experiment), but with the same limit of 12 words, or no limit at all.

In sum, while Kruse, Pankhurst, and Sharwood-Smith state that: "contrary to the expectations raised by earlier studies, we find that word association tests do not show much promise for the specific role created for them in L2 research" (1987: 153), a substantially improved test, using the same software, still has the potential to establish a link with an alternative measure of vocabulary size.

References

Fotos, S. (1991). The Cloze Test as an Integrative Measure of EFL Proficiency: A Substitute for Essays on College Entrance Examinations? *Language Learning*. 41: 3. 313-336.

Kent, G. H. and A. J. Rosanoff, (1910) A Study of Association in Insanity. *American Journal of Insanity* 737-96 and 317-390.

Kiss, G. R., Armstrong, C., Milroy, R., and Piper, J. (1973). An associative thesaurus of English and its computer analysis. In Aitken, A. J., Bailey, R. W. and Hamilton-Smith, N. (Eds.), The Computer

and Literary Studies. Edinburgh: University Press.

Kruse, H, J. Pankhurst, and M. Sharwood-Smith, (1987). A Multiple Word Association Probe. *Studies in Second Language Acquisition* 9, 2, pp. 141-154.

Fitzpatrick, T. (2006). Habits and rabbits: word associations and the L2 lexicon. *EUROSLA Yearbook, 6*, 121-145.

Haastruup, K and B. Henriksen. (2000). Vocabulary acquisition: acquiring depth of knowledge through network building. *International Journal of Applied Linguistics*, Vol. 10, NO. 2.

Harley, B. (1996). Introduction: Vocabulary Learning and Teaching in a Second Language. *The Canadian Modern Language Review*, 53. 1. pp. 3-12.

Lambert, W.E. (1956). Developmental aspects of second language acquisition. *Journal of Social Psychology*, 43, 83-104.

Lambert, W. E., & Moore, N. (1966). Word association responses: comparisons of American and French monolinguals with Canadian monolinguals and bilinguals. *Journal of Personality and Social Psychology, 60*, 376-383.

Laufer, B. and P. Nation. (1999). A vocabulary-size test of controlled productive ability. *Language Testing* 16, 33-51.

Laufer, B. (2001). Quantitative evaluation of vocabulary: How can it be done and what is it good for? In *Studies in Language Testing. No. 11* Cambridge: Cambridge University Press

Meara, P. (1978). Learners' word associations in French. *The Interlanguage Studies Bulletin* 3 (2): 192-211

Meara, P. (1980). Vocabulary acquisition: a neglected aspect of language learning. *Language Teaching and Linguistic Abstracts, 13*, 221-246.

Meara, P. (1983). Word associations in a second language. *Nottingham Linguistics Circular* 11, 28-38.

Meara, P. and G. Jones. (1990). *Eurocentres Vocabulary Size Test 10Ka*. Zurich: Eurocentres.

Meara, P. (1996). The dimensions of lexical competence. In Brown, G., K. Malmkjaer and J. Williams (Eds) *Performance and competence in second language acquisition*. Cambridge: Cambridge University Press, 35-53.

Meara, (1996). The Vocabulary Knowledge Framework. Available online at: http://www/swan.ac.uk/cals/calsres.vlibrary/pm96d.htm revised 1999.

Meara, P., & Fitzpatrick, T. (2000). Lex 30: an improved method of assessing productive vocabulary in an L2. *System, 28* (1), 19-30

Meara, PM and B. Wolter (2004). V_Links: Beyond Vocabulary Depth. *Angles on the English Speaking World* 4 85-97.

Nation, ISP. (1983). Testing and Teaching Vocabulary. *Guidelines* 5: 12-25.

Nation, ISP (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Miller, K. M. (1970). Free-association responses of English and Australian students to 100 words from the Kent-Rosanoff word association test. In Postman, L., & Keppel, G. (Eds.). *Norms of Word Association*. New York: Academic Press.

Orita, M. (1999). Word association patterns of Japanese novice EFL learners: A preliminary study. *The JACET Kyushu-Okinawa Chapter Annual Review of English Learning & Teaching, 4*, 79-94.

Orita, M (2002a). Proficiency, lexical development and the mental lexicon: investigating the response type distribution of word associations of Japaenese EFL learners and native speakers. *Research Reports Yatsushiro National College of Technology* 24, 113-124.

Orita, M. (2002b) Word associations of Japanese EFL Learners and native speakers: shifts in

response type disgtribution and the associative development of individual words. *Annual Review of English Language Education in Japan* 13, 111-120.

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

Nissen, H. and B. Henriksen (2006) Word class influence on word association test results. *International Journal of Applied Linguistics* 16: 3

Postman, L., & Keppel, G. (Eds.). (1970). *Norms of Word Association*. New York: Academic Press.

Randall, M. (1980). Word association behavior in learners of English as a foreign language. *Polyglot, 2* (2).

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

Richards, J. (1976). The role of vocabulary teaching. *TESOL Quarterly*, 10, 77-89.

Sanford, K and I. Svetics. (1994) Word associations-types and language proficiency. *PALM 8, 2* (1994), 63-76.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition, 19*, 17-36.

Schmitt, N. (1998). Quantifying word association responses: What is native-like? *System, 26*, 389-401.

Schmitt, N. (2000). *Vocabulary In Language Teaching*. Cambridge: Cambridge University Press.

Sinopalnikova (2003). Word Association as a Resource for Building WordNet. GWC 2004 Proceedings pp. 199-205

Söderman, T. (1992) Word associations of foreign language learners and native speakers - different response types and their relevance to lexical development. In Hammarberg, B (Ed) *Problems, process and product in Language Learning*. Abo: Stockholm

Wolter, B. (2001). Comparing the L1 and L2 mental lexicon: A depth of individual word knowledge model. *Studies in Second Language Acquisition, 23*, 41-69.

Wolter, B. (2002). Assessing proficiency through word associations: is there still hope? *System*, 30 (2002), 315-329.

Wolter, B. (2005). V_Links: A New Approach to Assessing Depth of Word Knowledge. Unpublished PhD thesis, the University of Wales, Swansea.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Zareva, A, P. (2004). Models of lexical knowledge assessment of second language learners of English at higher levels of language proficiency. *System*, 33, 547-562.

Zareva, A, P. Schwanenflugel and Y. Nikolova. (2005). Relationship between lexical competence and language proficiency: Variable Sensitivity. *Studies in Second Language Acquisition* 27: 567-595. Cambridge University Press.

Mean scores and standard deviations were calculated using tools found online at http://www.physics.csbsju.edu/stats/descriptive2.html

Correlations were calculated with:

Wessa, P. (2006), Free Statistics Software, Office for Research Development and Education, version 1. 1. 20, URL http://www.wessa.net/

Appendix 1　Example of a scoring grid for the weighted stereotypy scoring grid for the stimulus *high*.

2. HIGH	1	2	3	4	5	6	7	8	9	10	11	12
low [675]	144	132	120	108	96	84	72	60	48	36	24	12
school [49]	132	121	110	99	88	77	66	55	44	33	22	11
mountain [32]	120	110	100	90	80	70	60	50	40	30	20	10
up [18]	108	99	90	81	72	63	54	45	36	27	18	9
chair [17]　tall [17]	96	88	80	72	64	56	48	40	32	24	16	8
tower [13]	84	77	70	63	56	49	42	35	28	21	14	7
jump [11]	72	66	60	54	48	42	36	30	24	18	12	6
ladder [10]	60	55	50	45	40	35	30	25	20	15	10	5
building [8]　noon [8]	48	44	40	36	32	28	24	20	16	12	8	4
above [7]　cliff [7]	36	33	30	27	24	21	18	15	12	9	6	3
sky [6]	24	22	20	18	16	14	12	10	8	6	4	2
all other responses	12	11	10	9	8	7	6	5	4	3	2	1

The top row 1-12 indicates the order in which responses were entered.

The left-hand column indicates the scoring responses based on the response hierarchy in the 1952 Minnesota Word Association Norms (Jenkins, in Postman and Keppel, 1970). The number in brackets following each scoring response indicates the number of incidences of the item on the norms list. *Low* was elicited as a primary response by 675 out of 1000 responses. Scores are calculated in the following way. If a subject enters *low* as a primary response, 144 points are awarded. If a subject enters *building* as a primary response, 48 points are awarded (see Appendix 2), or only 4 points as a twelfth response.

"All other responses" in the bottom row are for responses which are on the norms lists but with a total number of incidences of less than 6 on the norms list.

Note. It's possible that Kruse et al did not score all these responses, accounting for lower mean stereotypy scores than in the replication.

Appendix 2 Example of student production on the WAT.

The stimulus word appears in the top row above the responses.

Scoring responses include the weighted stereotypy score in brackets, eg. building [48]

The total score for each measure, for each stimuli, appear below the responses.

A=number of responses, B=non-weighted stereotypy, C=weighted stereotypy.

The total for each measure appears at the end of each of the above rows (summed with Excel auto-sum).

HIGH	SICKNESS	SHORT	FRUIT	MUTTON	PRIEST	EATING	COMFORT	ANGER	
building [48]	sad [12]	time [12]	banana [60]	soft		food [144]	mooton	bad [12]	
tree	ill [121]		apple [132]	light		breakfast	chair [132]	wrong	
cost	hospital [50]		peach [10]	white [12]		dinner [30]	bed [110]	red [70]	
level	doctor [54]		strawberry [9]	comfortable		lunch [9]	wear		
score [8]	medicine [8]		blueberry	coat [8]		fruit	place		
	weak [7]		watermelon			dessert [7]			
			cherry [18]			drink [24]			
						snack			
A 5	6	1	7	5	0	8	5	3	40
B 2	6	1	5	2	0	5	2	2	25
C 56	252	12	229	20	0	214	242	82	1107

Appendix 3.1　Cloze test

Whales.

Whales live in the sea and have fins, but they are not fish. They are huge mammals that [1] _____ learned to live in the water. [2] _____ order to breathe air into [3] _____ lungs, they must come to the [4] _____ for air. On the top of [5] _____ head is a blowhole. This is [6] _____ they use to breathe. Some whales, [7] _____ as the great blue whale, can [8] _____ submerged for more than half an [9] _____ before they need to surface for [10] _____.

The biggest whale is the blue [11] _____, which grows to be about 29 [12] _____ long- the height of a nine- [13] _____ building. Adult blue whales have no [14] _____ except man.

Whales use echolocation sounds [15] _____ locate food. These sounds bounce onto [16] _____ animals and then back to the [17] _____. The sounds are very high-pitched, and [18] _____ cannot hear them. The location of [19] _____ echo helps the whales find food.

[20] _____ whales are toothless and others have [21] _____. Those who do have teeth feed [22] _____ shrimp, fish, crabs, and other varieties [23] _____ sea creatures. Toothless whales, such as [24] _____ humpbacks, have bony plates called baleen [25] _____ their upper jaws. The plates have [26] _____ on them that catch the food [27] _____ the whales eat. Some whales eat [28] _____ much as three or four tons [29] _____ food each day.

Even though whales [30] _____ swim anywhere in the ocean, they [31] _____ to follow the same route year [32] _____ year to feed, mate, and breed [33] _____ the same location. Gray whales make [34] _____ longest seasonal migration of any of [35] _____ whales. They travel about 12,500 miles a [36] _____. Some species of whales like to [37] _____ together in herds, which often number [38] _____ to a thousand.

The skin of [39] _____ whale is tough and spongy. Heavy [40] _____ of fat accumulate under the skin. [41] _____ is called blubber and helps to [42] _____ the whale warm in the cold [43] _____ waters. In the past, whales were [44] _____ for the blubber which was then [45] _____ into candles or oil and sold.

[46] _____ are many species of whales that [47] _____ in very serious danger of becoming [48] _____. Most baleen whales are listed as [49] _____ or protected species. Most other whale [50] _____ are doing well and should survive. It would certainly be a great shame to lose any species of these large, magnificent, and intelligent aquatic mammals.

Appendix 3.2　Grammar monitoring test.　Instructions with example.

Grammar error recognition test.

Directions: In questions 1-50, each sentence has four words or phrases underlined. The four underlined parts of the sentence are marked (A), (B), (C), or (D). You are to identify the one underlined word or phrase that should be corrected or rewritten.

Mark your answer on your answer sheet.

Example:

All employee (A) are required to wear (B) their identification (C) badges while (D) at work