タイトル	Report on a free continuous word association test (part 7): Qualitative analysis of WAT20 behavior based on post-task
著者	Ian, MUNBY
引用	北海学園大学学園論集(188): 125-142
発行日	2022-07-25

Report on a free continuous word association test (part 7): Qualitative analysis of WAT20 behavior based on post-task questionnaires and interviews

Ian MUNBY

INTRODUCTION

In the studies reported in Munby (2007, 2008, 2018, 2019a, 2019b, 2019c) non-native subjects at higher levels of proficiency generally achieve higher scores than their lower-level peers on the WAT (on both the number of response and stereotypy measures). Further, some non-native subjects perform relatively well on the WAT, and others poorly, in comparison with their performance on standard proficiency or vocabulary knowledge measures. For example, in Table 2 (Munby, 2019b), Pearson correlations between the WAT20 stereotypy measure and the translation test stood at $r = .791$ ($p < 0.01$). Following Cohen et al. (2000, p.202) correlations at this level, while accurate enough for making predictions about the performance of groups on two different tests, are not high enough to predict individual performance. One could take this as evidence that both abilities and gains in different aspects of L2 learner proficiency and lexical knowledge are uneven. Alternatively put, learners may exhibit strengths in some aspects of L2 ability, but weakness in others, possibly as part of a developmental process rather than any permanent feature of a learner's ability. In this way, a low WAT score may testify to poor lexical processing ability, perhaps stemming from weakness in lexical fluency (in the "number of responses" measure), and lack of productive knowledge of native-like associations (in the stereotypy measure), or a combination of the two.

Several commentators have put forward models of lexical development that can account for variability in performance on L2 word association tasks. To take an example, Kroll & Stewart, in their Revised Hierarchical Model (1994, see Figure 1), suggest a model of language interconnection in which second language lexical items are linked to first language words alongside links to concepts. In this model lies one potential explanation for the varied performance of L2 learners in an L2 word association task or test. This is that the rate of L2 learner response production in timed

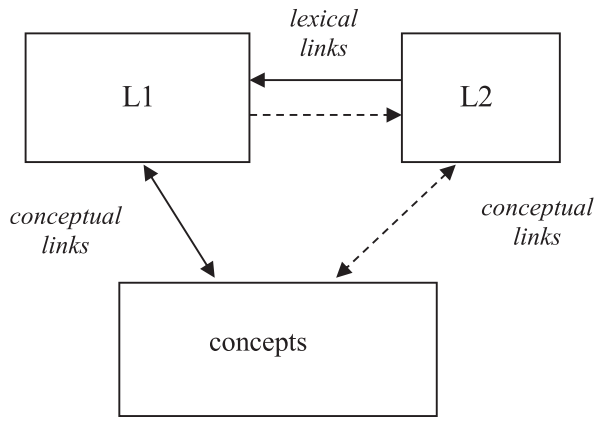

Figure 1. The revised hierarchical model of lexical and conceptual representation in bilingual memory. (Kroll & Stewart, 1994, p.158).

conditions could be slowed down by these L1-dependent lexical links. According to Sakui & Gaies, (1999), in a study of young Japanese adult learners, self-reported L1 dependence is stronger in lower-level learners than in their higher-level peers. With WAT20, this phenomenon may contribute to the achievement of higher scores among higher level, less L1 dependent learners. I shall return to this issue later.

Wolter (2001), in a study of learner response type in L2 word association, suggests another model that could account for differing rates of learner response production in L2 word association tasks. This model is termed the DIWK (Depth of Individual Word Knowledge) model, wherein individual word knowledge can be described on a continuum from well-known, fairly well-known, moderately well-known, and slightly known, to unknown. He adds: "The real interest in a DIWK model, of course, lies not in the patterns of the responses themselves but rather in the subconscious connections they reveal between the words that form the whole of the mental lexicon" (Wolter, 2001, p.48). The underlying theory is that learners who have high-quality subconscious connections between lexical items, and speedy access to them, are likely to perform better than those who do not on a WAT such as the one I have been developing in this series of studies. Wolter also comments on the difficulty of the "very subconscious process of producing a single response to a single prompt word with a smaller mental lexicon" (2001, p.65). However, despite the attractions of this model, in the course of informal post-WAT discussions with both native and non-native subjects in the studies reported in Munby (2007, 2008, 2018, 2019a, 2019b, 2019c), I had heard numerous claims that the responses they provided were often not subconscious, but edited or mediated. This could be due to the way responses are collected in

WAT20, in written form to printed stimuli, as opposed to spoken stimuli and responses used by Wolter (2001). In this way, WAT20 may not provide an efficient "window into the mental lexicon", or sample of the network of subconscious links between words.

If participants do not always provide responses that occur to them subconsciously, this represents a challenge to the validity of WAT20, particularly concerning the pre-task instructions wherein subjects are invited to provide the responses that the stimulus words make them think of, with the possible result that performance could be affected. This is related to a further issue regarding the face validity of WAT20. It is by no means certain that non-native subjects perceive WAT20 as an activity that is a useful and challenging measure of L2 vocabulary knowledge as it is intended to be.

On a similar note, there is another factor that might affect test performance; one could attribute variation in WAT performance to the subject attitude to the test. According to Daller et al. (2007): "The first threat to the validity of the test arises from the testees' attitudes towards the test, their willingness to participate, or not, due to negative experiences with previous tests and their familiarity with the test format" (p.17). Further, in considering some fundamental issues in assessing vocabulary knowledge through tests, Nation (2007) states: "Of all the factors looked at in this paper, the one that troubles me the most is the one of learner attitude because this is the one where the researcher has the least control" (p.43).

To throw light on (i) the influence of the models of bilingual lexical processing stated above, (ii) subject perception of their L2 lexical processing ability in the context of WAT20, and (iii) the "attitude threat" mentioned by Daller, all 111 non-native subjects who participated in the study in Munby (2019b) were invited to complete a WAT attitude and awareness questionnaire immediately following WAT20, and before completing the two proficiency countermeasures. Note that of these 111, 50 subjects later participated in the longitudinal study of WAT20 reported in Munby (2019c). To probe more deeply into the associational behavior of this group of subjects, I refer to post-task (unstructured) interviews with four of the non-native subjects and three of the native subjects concerning their associational behavior. The purpose here was also to ask questions about particular responses to gain further introspective data. The following research questions were designed to address the issues outlined above. With the concerns expressed by Daller et al. (2007) in mind, the first research question is:

RQ1 Does non-native subject attitude to WAT20 affect performance?

To determine whether or not non-native associative behavior is influenced by L1 dependent processes, as suggested by Kroll & Stewart (1994), I ask:

RQ2 Are L2 learner responses mediated by L1?

Following the comments on the DIWK model by Wolter (2001), this WAT may measure both the quality or depth of L2 individual word knowledge and the number of L2 words known. To see if there is a match between what theory suggests the WAT is measuring, and non-native subject perception of their performance in the WAT, I ask:

RQ3 Do the subjects perceive the WAT as a useful measure of their L2 lexical processing ability?

With RQ4, I also investigate the claim underlying one of the cue selection criteria in Munby (2008, 2018), namely that all the cue words should be known to the subjects taking the test. As mentioned in Munby (2018), if a subject does not know the meaning of a cue word, it is almost impossible to provide a native-like association.

RQ4 Do the non-native subjects know the meaning of all the cue words?

The final research question, derived from the concerns stated previously in this section, relates to both the questionnaire and comments provided in recorded interviews. If associations are consciously chosen according to subject response preference, WAT20 may not qualify as a "window into the mental lexicon". Of particular interest is the phenomenon of the chained response, or a response connected with the previous response rather than the cue word. This phenomenon was reported in the replication study in Munby (2007). To throw light on these issues, I ask:

RQ5 Does the associative behavior of both native and non-native subjects reflect subconscious links between the cue words and responses?

Section 2: METHOD

The purpose of the post-task WAT20 attitude and awareness questionnaire was to address the first four research questions (RQs1–4) stated in the previous section. The post-task interviews provide insights regarding RQ5. In the following sub-sections, I shall describe the rationale underlying the design of the questionnaire, the interview procedure, and the limitations of the methodology.

2.1 Design of the WAT attitude and awareness questionnaire: rationale and limitations

The 12-item questionnaire uses a 5-point Likert scale to rate attitude and awareness on a scale of 5 (strongly agree), 4 (agree), 3 (Neither agree nor disagree), 2 (Disagree), or 1 (Strongly disagree). The questionnaire was given in Japanese. The English version appears in the appendix.

Some subjects also provided written comments in the space provided at the end of the questionnaire.

Addressing RQ1 (Does non-native subject attitude to WAT20 affect performance?), I included Q1, or Item 1: "I like this kind of activity". This is because WAT20 is a non-standard activity involving the typing of a large number of single words, one after another, for a period of up to thirty minutes, with no clear purpose, reward, feedback, or apparent benefit for the participant. In addition, the absence of clear right or wrong answers may also negatively affect the subject attitude toward the test. As such, there is a possibility that performance may be related to the extent that the subject likes or dislikes the activity. To assess the degree of liking for the activity from a different angle, Q2 (I prefer writing English sentences) seeks to compare subject attitude to writing a series of single words with standard sentence writing activities. Q3 (I tried as hard as I could) aims to compare WAT20 performance with the degree of effort expended. The inclusion of this item was motivated by an observation from Wolter (2002). In his multiple response WAT, Wolter opted for eliciting a maximum of three responses instead of 12 because "producing such a large number of responses to a single prompt word is a task which requires a good deal of effort" (p.5).

The problem is that positive answers to Q1 and Q3 could simply be a reflection of the subjects' desire to respond in the way the researcher hopes they would. Dörnyei (2010) observes that "sometimes respondents deviate from the truth intentionally" (p.8). I will return to the issue of the limitations of the Likert scale questionnaire in discussing the results. Although the subjects were advised that there were no right or wrong answers, Q4 (I sometimes didn't write a word in case it was the wrong response) aims to find if some subjects were less confident in providing responses than others, with the "number of response" measure thus becoming a measure of confidence rather than lexical fluency. A related aim was to discover if any of the subjects were not entering some of the responses that occurred to them. Note that the WAT aimed to collect and assess responses that occurred to the subjects spontaneously. One problem is that subjects were asked to avoid proper nouns or responses of more than one word, in which case the respondents could be signaling that these types of responses had occurred to them, but they had remembered the rules. In other words, the instructions themselves may require a degree of response editing. Q7 (I couldn't think of enough words associated with the cue, so sometimes I just wrote down any English words I could think of) aims to assess subject task fulfillment strategies. It also asks if the subjects were providing responses that have links to the cue word. The purpose of Q11 (Sometimes my mind went blank. I got stuck) is to find if any of the subjects experienced difficulty

producing responses for affective rather than linguistic reasons. At least two subjects had reported, informally, suffering "mental blocks" and "getting stuck" in previous studies. However, one problem is that mental blocks may also be a product of poor ability to produce multiple associations. It may also result from having the feeling of knowing a word, but not being able to access it in timed conditions, for example. Indeed, negative attitudes to this WAT may result from poor L2 lexical processing ability rather than motivational issues.

Addressing RQ2: Are L2 learner responses mediated by L1?

Q8 "When I try to think of a response, I translate it into Japanese in my head" seeks to determine whether or not the subjects were using L1 dependent strategies. Agreement with Q9 "Sometimes I couldn't write a response because I didn't know the English word" would also suggest that the subject is resorting to L1 to provide associations since limited L2 vocabulary size may result in an inability to provide L2 responses.

Addressing RQ3: Do the subjects perceive the WAT as a useful measure of their L2 lexical processing ability?

The following items target the subjects' perception of their lexical processing ability:

Q5 "If I knew more words, I'd be better at the activity".

Q6 "If I could think of words more quickly, I'd be better at the activity".

Q10 "It was more difficult to think of a response for some words than for others".

Agreement with Q5 and Q6 would probably indicate that the respondent is aware of what is required to improve her performance on WAT20 concerning both L2 lexical knowledge, or "knowing" words (specifically Q5), and lexical processing ability, or ability to access responses fluently (Q6 and Q10).

Q12 "I knew the meaning of all the cue words" addresses RQ4: Do the non-native subjects know the meaning of all the cue words?

To pick up possible differences in questionnaire response patterns according to performance on WAT20, the subject population was divided into two groups (n = 37) representing the highest third and the lowest third on the WAT A (number of response) measure and the WAT B (stereotypy measure). In this way, there were two different sets of high-low groups, one for each measure. The means of the Likert scale responses for each item in each group were compared. Before presenting results, I point out a further methodological limitation; it is not possible to be sure whether WAT scores are influenced by subject attitude to the task, or that attitude is due to the perception that they have performed well. Nonetheless, we would expect that non-native subjects

who (i) enjoy the activity, (ii) try hard, (iii) express confidence in their responses, (iv) avoid translating cues and responses from English to Japanese, and then back into English, and (v) know the meaning of all the cue words, would perform better than those who do not.

2.2 Post-task interviews

To address RQ4 (Does the associative behavior of both native and non-native subjects reflect subconscious links between the cue words and responses?) individual post-task interviews were set up with four non-native subjects whom I refer to by their approximate level of English as the "advanced male", the "advanced female", the "intermediate", and the "elementary". With native subjects I conducted the interviews following completion of WAT20, or, in the case of non-native subjects, following completion of WAT20, the WAT attitude and awareness questionnaire, the EVST, and the translation task. Although the interviews were unstructured, the purpose was to find out how or why these subjects produced responses both in general and particularly concerning specific responses that were available on their WAT response text files.

Section 3: RESULTS

In Table 1, I list the number of respondents who circled each level of agreement for each item. For example, 37, or exactly one-third of the 111 subjects, indicated that they strongly agreed with the statement Q1. Mean levels of agreement are high in Q1 and Q3, indicating that stated attitude to the WAT20 is unlikely to be a factor affecting results. Similar high mean levels of agreement with Q5 and Q6 indicate that the subjects were aware that WAT20 tests, or challenge their lexical processing skills. In contrast, with Q8, the distribution of Likert scale responses is more evenly spread. Following Kroll & Stewart (1994), it is predicted that differences in response patterns for this item may be related to proficiency.

As mentioned in the previous section, the next step was to sort the 111 subjects into 2 groups of 37: low and high (see Table 2) according to their scores on the following three measures: WAT A (number of responses) and WAT B (stereotypy). The purpose of this analysis was to investigate whether or not scores on these measures could be predicted, or partially accounted for, by the subject attitude to the test and perception of lexical processing abilities. An unpaired t-test was calculated to test for significant differences between the means of the 2 groups. Note that I present the t-values as trend indicators rather than absolutes since the data is skewed towards high (5), and because the scores represent discrete categories rather than a scale.

Table 2 indicates that there are significant differences in response patterns between low and high

Table 1
Distribution of Likert Scale responses to the WAT attitude and awareness questionnaire (n = 111)
5 = Strongly Agree, 4 = Agree, 3 = Neither agree nor disagree, 2 = Disagree, 1 = Strongly disagree.

	Likert Scale							
Questionnaire items	5	4	3	2	1	Mean	*SD*	N
1. I like this kind of activity.	37	59	12	1	2	4.15	0.79	111
2. I prefer writing English sentences.	14	23	53	15	6	3.22	1.01	111
3. I tried as hard as I could.	58	47	4	1	1	4.44	0.70	111
4. I sometimes didn't write a word in case it was the wrong response.	12	34	12	30	23	2.84	1.35	111
5. If I knew more words I'd be better at the activity.	89	11	7	4	0	4.67	0.75	111
6. If I could think of words more quickly I'd be better at the activity.	82	19	6	1	2	4.58	0.90	110
7. I couldn't think of enough words associat- ed with the cue, so sometimes I just wrote down any English words I could think of.	41	39	13	12	4	3.86	1.23	109
8. When I try to think of a response, I translate into Japanese in my head.	24	34	22	13	18	3.30	1.37	111
9. Sometimes I couldn't write a response because I didn't know the English word.	17	15	17	21	41	2.51	1.48	111
10. It was more difficult to think of a response for some words than for others.	58	41	6	2	4	4.32	0.94	111
11. Sometimes my mind went blank. I got stuck.	20	37	19	17	17	3.21	1.37	110
12. I knew the meaning of all the cue words.	43	32	8	19	9	3.73	1.35	111

when divided according to both WAT A and WAT B scores with only four items Q4, Q8, Q9, and Q12. In contrast, there is no significant difference in the above comparisons of low and high with either of the two measures with items Q1, Q5, Q6, and Q7.

Table 2
A comparison of the means and standard deviations (SD) for Likert scale responses to the WAT attitude and awareness questionnaire for two groups of subjects divided (lowest third and highest third, n = 37 in each).

		LOW	HIGH	*t*-value	*p* value
1. I like this kind of activity.	A	3.97 (0.83)	4.05 (0.85)	0.679	0.415
	B	4.05 (0.91)	4.27 (0.65)	1.174	0.244
2. I prefer writing English sentences.	A	3.00 (0.88)	3.54 (1.04)	2.407*	
	B	3.05 (0.88)	3.49 (1.04)	1.926	0.058
3. I tried as hard as I could.	A	4.38 (0.68)	4.59 (0.76)	1.287	0.202
	B	4.35 (0.72)	4.68 (0.48)	2.298*	
4. I sometimes didn't write a response in case it was the wrong response.	A	3.32 (1.36)	2.46 (1.35)	2.755***	
	B	3.22 (1.36)	2.54 (1.35)	2.151*	

5. If I knew more words I'd be better at the test.	A	4.76 (0.60)	4.57 (0.90)	1.067	0.290
	B	4.78 (0.58)	4.57 (0.84)	1.291	0.201
6. If I could think of words more quickly I'd be better at the activity. [† n = 36]	A	4.70 (0.57)	4.62 (0.86)	0.635	0.487
	B	4.57 (0.90)	4.69 (0.62) †	0.699	0.477
7. I couldn't think of enough words associated with the cue, so sometimes I just wrote down any English words I could think of. [† n = 36]	A	3.81 (1.22)	4.00 (1.05)	0.478	0.713
	B	4.03 (1.13) †	3.87 (1.11)	0.621	0.537
8. When I try to think of a response I translate into Japanese in my head.	A	3.57 (1.19)	2.70 (1.54)	2.697**	
	B	3.70 (1.13)	2.60 (1.42)	3.713***	
9. Sometimes I couldn't write a response because I didn't know the English word.	A	3.14 (1.49)	1.81 (1.27)	4.114***	
	B	3.24 (1.44)	1.78 (1.23)	4.688***	
10. It was more difficult to think of a response for some words than for others.	A	4.46 (0.56)	4.05 (1.22)	0.071	1.834
	B	4.46 (0.73)	4.11 (1.17)	2.830**	
11. Sometimes my mind went blank. I got stuck. [† n = 36]	A	3.54 (1.22)	3.14 (1.49)	1.281	0.205
	B	3.54 (1.17)	2.89 (1.41) †	2.152*	
12. I knew the meaning of all the cue words.	A	3.27 (1.48)	4.38 (0.89)	3.892***	
	B	2.97 (1.44)	4.43 (0.87)	5.273***	

Key to rows: A = WAT A, B = WAT B.
Significant at *p < 0.5, **p < 0.1, p < .0.001***

Section 4: DISCUSSION

In this section, I will address the research questions concerning the results of the analyses in Tables 1 and 2 and in the light of commentary provided by the subjects in the interviews.

RQ1 Does non-native subject attitude to WAT20 affect performance?

Since the overall degree of reported liking of the test (Q1) was high (mean 4.15, *SD* 0.79, see Table 1), it was no surprise that there was no significant difference between low and high groups divided according to WAT A and WAT B in Table 2. In contrast, although the reported degree of effort expended on the test (Q3) was similarly high (mean 4.44, *SD* 0.70), there is a significant difference between the low group and high group when divided according to WAT B but not WAT A. These two observations are tempered by the comments of Dörnyei (2010), who warns that questionnaire responses may not be truthful. Added to this, Cohen et al. (2000) warn that "one respondent's 'agree' may be another's 'strongly agree'" (p.253). Further, the degree of liking and effort may also affect subject performance in the countermeasures upon which concurrent validity is assessed. With this in mind, there is no convincing evidence that non-native performance on WAT20 can be predicted by the degree of reported liking or expended effort. However, results for Q2 show that when the subjects are divided according to WAT A scores, the group which entered the largest number of responses expressed a significantly stronger preference for writing English sentences

than the low group. In other words, preference for writing sentences rather than word associations is not a reason for the low group entering fewer responses than the high group.

Concerning Q4, the high group is significantly more confident in entering responses than the low group (divided according to both WAT A and WAT B). This suggests that WAT performance can be predicted to a certain extent by the degree of confidence in entering responses. In contrast, this lack of confidence may be the result rather than the cause or factor contributing to low-scoring performance. In other words, subjects who score low on WAT20 may be reluctant to enter a larger number of responses due to awareness that their cue-response links are not "right" or tenuous. This tendency could be a function of proficiency rather than an overly cautious attitude to the WAT. However, if some subjects were less reluctant to provide responses than others, this may affect WAT performance. Here, one would expect to find a significant difference between low and high groups in Q7 (I couldn't think of enough words associated with the cue, so sometimes I just wrote down any English words I could think of). Since no significant difference was found between low and high here, evidence that low-scoring subjects were simply cautious is weak.

Finally, as commented earlier in Section 2.1, the interpretation of Q11 (Sometimes my mind went blank. I got stuck) is not straightforward. When sorted according to WAT B, members of the low group report a significantly higher incidence of occurrence of their minds going blank and getting stuck than the high group. However, this difference is not significant with WAT A, the measure that one would predict would be most affected by the occurrence. In other words, if a subject is unable to produce responses due to the feeling of getting stuck, one would expect this to affect the number rather than the quality of responses provided. This suggests that the occurrence of mental blockage interrupting response production may have more to do with the level of proficiency than the ability to produce responses in timed conditions. In sum, in considering the results of this survey regarding RQ1, there is only weak evidence to suggest that learner attitude to the WAT20 is a factor influencing performance on it.

RQ2 Are L2 learner responses mediated by L1?

This question was addressed through two items: Q8 (When I try to think of a response, I translate it into Japanese in my head) and Q9 (Sometimes I couldn't write a response because I didn't know the English word). With both items, the low group reports a significantly higher rate of agreement than the high group when sorted according to both measures: WAT A and WAT B. This finding indicates that learners who are less able to produce responses, in either quantity or native-like

quality than their peers, think of responses in L1 before translating them back into English, before finally typing and entering them. This is backed up by written comments (in L1 Japanese) volunteered at the end of the questionnaire such as: "When I thought of my responses, I thought in Japanese first, then translated into English. Then I just typed the English word. If you do not know the cue words well, it is difficult". This is a complicated, labored process that can be predicted from the model of L2 processing put forward by Kroll & Stewart (1994). This process may also have two key effects on non-native WAT performance. First, the process of producing L1 responses in this way would likely be much slower than for a higher-level subject who is able to think entirely, or mostly, in L2. This finding lends support to the idea that the availability of L2 responses in the learner lexical store may be compromised by L1 dependency at lower levels. This is likely to have a fundamental effect on fluency or access speed in response production. The second point is that this process is liable to affect the type of responses provided by the lower-level learner; these are almost certainly influenced by L1 associative networks to a large extent. Vivas et al. (2019) voiced similar concerns in their study of the English word association responses provided by L1 Spanish speakers. They conclude that L1 influence likely prompted their subjects to produce more language-specific responses which may therefore be less likely to be universal. Further, for those learners who do claim to access links between words in this way, there appears to be a problem again with knowing L2 equivalents for L1 responses that occur to them. Q9 indicates that this problem is more acute with learners of lower levels of ability. This "dead end" effect of the fruitless search for L2 equivalents for intended responses that appear to the subject in L1 would probably absorb limited thinking time and result in lower rates of response production. For this reason, it may be possible to predict that learners with fewer L2 items, or smaller L2 productive vocabulary size, will score lower on WAT20, and account for the finding of a link between WAT20 performance and L2 vocabulary size and proficiency.

RQ3 Do the subjects perceive the WAT as a useful measure of their L2 lexical processing ability? Concerning Q5 (If I knew more words I'd be better at the test) and Q6 (If I could think of words more quickly I'd be better at the activity), Table 1 indicates that there is overwhelming agreement among all participants that performance on the WAT is linked to the number of words known and the ability to produce responses quickly. The following claim is representative of numerous post-task written comments on the questionnaire to the same effect: "My English word power is too poor. I could hardly make associations at all. I thought I should learn more English words". Another wrote: "It was difficult. It took me a long time to think of a response and type it, and I couldn't type enough responses to many of the cue words. I need to learn more words".

This is an important positive pedagogical outcome since it appears that, even for the high-level learner, providing up to twelve responses to high-frequency cue words within a time limit presents a challenge and inspires a positive awareness of his or her potential and need to progress. While with Q5 and Q6 there is no significant difference between high and low groups when divided by all three measures (see Table 2), the means for Q5 (mean 4.67, *SD* 0.95) are higher than for Q6 at 4.58, *SD* 0.90 (see Table 1). A paired t-test was conducted to compare the means of the whole group of 111 subjects to investigate whether or not learner perception of how many words they know (Q5) is more important than how well or quickly they can access them (Q6). The difference was not found to be significant. With Q10 (It was more difficult to think of a response for some words than for others), there is also a high level of overall agreement at 4.32, *SD* 0.94 (see Table 1). This also supports the notion that the WAT presents a challenge to all participants. Further, compared with the high group, the low group (when divided according to WAT B scores) reports a significantly higher degree of difficulty in thinking of responses for some cue words more than others. In contrast, there is no significant difference between low and high when divided according to WAT A for this item. In sum, there is no doubt that the subjects do perceive the WAT as a useful measure of their L2 lexical processing ability.

RQ4 Do the non-native subjects know the meaning of all the cue words?

Concerning Q12 (I knew the meaning of all the cue words) Table 1 indicates that 19 subjects responded at Likert level 2 (disagree) and 9 at level 1 (strongly disagree). Further analysis shows that, of these 28 subjects, 18 had failed to provide any responses for one or more cue words. However, the problem of the "zero response" is not confined to a small number of problematic cues. While *lead*, the cue most often left blank, provoked a complete lack of responses in 8 cases, *cut* was the only cue in the set of twenty that never failed to elicit at least one response from the pool of 111 subjects. On the other hand, ten of the 28 subjects in the Likert 1 or 2 bands for this item did provide responses for all the cue words. While these subjects did not always score points on the stereotypy scale, they appeared to demonstrate some knowledge of the cue word. For example, one of these 10 subjects provides the following non-scoring cue-response clusters: *choice* ＞*or*, *clothes*, *shoe*; *keep*＞*ball*, *earth*; *lead*＞*everyone*, *king*; *marry*＞*glad*, *want*, *girlfriend*, *bouquet*. It could be said that the experience of responding to L2 cues in this format raises awareness of what it means to know a word, a further positive pedagogical outcome. For example, one subject commented: "This kind of activity is hard because even if you know many words, you can't always make associations in your brain".

In sum, the "zero response" phenomenon is hardly widespread and only concerns 69 cases out of 2,220 response sets (111 subjects multiplied by 20 cues). It should also be noted that of the 8 subjects in the Likert level 3 band (Neither agree nor disagree) there is not a single "zero response" case. One also has to consider why eleven subjects in the Likert levels 4 and 5 bands did not provide responses to one or more cues. Clearly, the issue of whether the subjects really know the meaning of the cue words or not needs to be resolved in a simple fashion by asking them for an L1 translation of the cue words following the test. In this way, it would be possible to compare the subject perception of knowing what a word means with actual word meaning knowledge. Further, the incidence of zero response to a cue needs to be examined in non-timed conditions since it remains possible that failure to provide even a single response could be a result of several factors such as insufficient thinking time or lack of confidence, for example.

One problem that arose was that two lower-level subjects reported having trouble recognizing the cue words because they were capitalized. For example, one subject volunteered the following comment in Japanese: "Because the cue words were in capital letters, I couldn't read them at first (e.g. CHURCH)". Another commented in L1: "Usually I do not see English words all in capital letters so I had a hard time reading the cue words". I had previously thought that capitalization would facilitate word recognition by making the words stand out. Alternatively, displaying the cue words in lower case may represent an improvement to the format of this WAT. Further, most tests of receptive vocabulary size, such as the EVST, also measure lexical recognition skills, although it has to be said that single words are presented in lower case in this test. Nevertheless, this WAT possibly measures three aspects of lexical competence: word recognition, productive vocabulary size, and the ability to produce associations, an ability which may encompass other abilities such as speed of access to the lexical store, and the density of links between lexical items in the bilingual lexicon.

Concerning RQ5 (Does the associative behavior of both native and non-native subjects reflect subconscious links between the cue words and responses?) I discuss some more issues relating to the validity of the construct of WAT20 with particular regard to the guidelines for testees and how they are interpreted. In the context of post-task interviews, I shall discuss four problems with the way the subjects appear to interpret, follow, or fail to follow some of the guidelines explained to them before the task as detailed in the previous study (Munby, 2019a).

(i) "Participants were told that when you see or hear a word it makes you think of another word, and that I wanted to know what responses a set of cue words made them think of".

I divide this statement into two separate, and possibly mistaken, assumptions about the way responses are activated and provided. The first assumption is founded on the notion that cue words provoke other words in the mind of the test-taker. The second notion is that these words are freely available to the researcher for inspection. Neither notion matches with insights gleaned from interviews with both native and non-native participants. To deal with the first assumption, although clearly thousands of single words are provided as responses, cue words also provoke mental images, situations, stories, experiences, concepts, and items related to current environments. From this perspective, it seems that some of the responses typed in are in fact responses to these images, situations, stories, concepts, and environments rather than the cue word itself. For example, the elementary subject responds to *air* through his image of *airport* in providing the response *road*. Similarly, to the cue word *point* the advanced male subject responded with *block* (non-scoring), as in the American grid-like arrangement of streets, and *map* (scoring) the reason being "I am traveling and I was thinking of a map". Further, the advanced female subject claims to have entered the non-scoring response *annoying* for *gas* because of her experience of participating in a children's TV quiz show where a smoky gas appeared when the contestant got a quiz answer wrong. To *gas*, she also provided the non-scoring, experiential response *coagulate* because "my teacher taught me that when we heat propane it coagulates". If she had simply entered *propane*, the source of the association *coagulate*, she could have scored a point.

This brings us to a discussion of what underlies the second part of the instruction. The researcher wants to know "what words the cue word makes them think of". The advanced male subject's first response to the word *police* was, according to him, the word *gun*, but he did not enter it because, he claimed, he did not like guns, and provided a less violent alternative *button* (*baton*). Similarly, in response to *break*, he enters several particles such as *down*, *up*, *through*, *out* but the first one that occurred to him was *in*, which he declined to type. He states: "I started thinking of some dangerous words but then I thought I should behave myself", and the crime-related *break in* was one example he commented on. Similarly, post WAT interviews with the native subjects in Munby (2019b) revealed that they sometimes did not supply responses that occurred to them. For example, one native decided not to enter more than two names of types of gases for the cue *gas* because it suggested one-track thinking. Another happily married native participant was shocked to find himself entering the response *divorce* for *marry*, and decided to ensure that the next response was one that his wife would approve of and entered *love*.

Finally, returning to the first assumption, if responses did genuinely appear automatically to

the subject, as in "makes you think of", then there would be no need for subjects to enact task-completion strategies to produce them. There is evidence of strategy use in the absence of freely occurring responses in this study. For example, in the free comments section on the survey, one non-native subject wrote: "I tried to write words which came after or before the cue word. For example, if the cue was *free* my response was *time* and *paper*".

(ii) Subjects were also advised that there were no right or wrong answers. In the interview protocol, the advanced male subject also says that he used the strategy of trying to find a response by making a sentence including the cue word. For *lead*, he thought of *company*, *friend*, and *person* but he claimed he did not write them because in a sentence the indefinite article "a" would be required. In other words, since "lead friend" is grammatically incorrect, he felt the response was also wrong. He claimed to reject other responses that occurred to him on these grounds "so many times".

(iii) Subjects were also advised: "not to worry about spelling mistakes". Another finding that is similar to (ii) above is that some subjects do worry about spelling, despite what they are told. In the comments on the surveys, one subject writes: "I thought of some English words but I didn't write them because I didn't know the spelling". One native subject also confessed to not supplying responses that occurred to him due to concerns with spelling such as *complementary* or *complimentary* for the cue *free*. This is another way in which the researcher does not gain access to the words that come most immediately to the subject's mind.

(iv) They were also advised against "chaining away from the cue word" as in *cat* (cue), *mouse* (response 1), *cheese* (response 2), *biscuit*, *cake* etc. Despite this, evidence of chaining, or responding to the previous response rather than the cue word, surfaced in the interviews. For example, in response to *gas*, the advanced female subject provides *air*, *smoke*, *factory*. I suspected it was a chain, and she confirmed: "smoke and factory are connected in my mind". The phenomenon of the chain reaction sometimes seems to be related to the production of story-related responses. The intermediate subject claims to have entered the following responses to *police* based on a speeding ticket scenario: *traffic*, *policy*, *point* [points deducted from a license for speeding] *car*, *drive*, *arrest*, *bad*, *person*, and *people*.

In sum, the notion that a researcher can gain unmediated or unedited access to a learner's lexical store with this format appears undermined by evidence from the protocol. There is a

mismatch between what the subjects are told to do and what they actually do, and this stems from two weaknesses in the test task. The first is that there are probably too many rules to bear in mind. In other words, a subject may find it difficult to remember that there were no right or wrong answers, not to worry about spelling mistakes, and to avoid: (i) proper nouns, (ii) entering responses of more than one word, (iii) and "chaining away" from the cue word. A further possible consequence of these instructions is that more cautious subjects will spend longer editing their responses, which I am asking them to do.

A more serious concern is that the scoring system is unethically based on the instruction that there are no right or wrong answers when test-takers' responses are scored as such according to whether they appear on a norms list (right) or not (wrong). In this sense, I am doubtful that the multiple response WAT is based on an ethical construct. One solution would be to inform the subject how the WAT would be scored to inject some transparency into the purpose and assessment of the test. In other words, instead of asking subjects to provide as many responses as possible, subjects could be invited to supply as many native-like, norms-listed responses as possible.

Section 5: CONCLUSION

Through a questionnaire and interviews, this study aimed to gain a deeper understanding of a range of factors that have the potential to affect performance in WAT20. Investigation of learner attitude to WAT20 included questionnaire items gauging the degree of reported liking, effort expended, and reluctance to enter responses. These factors were not found to influence performance to a great extent. In contrast, there is clear evidence that lower-level subjects suffer from L1 dependency in producing associations, and this is consistent with models of the bilingual lexicon such as Kroll & Stewart's Revised Hierarchical model (1994). Further, overall, the non-native subjects believe that their performances on the test reflect their L2 lexical processing ability in terms of the number of words they know, and their ability to access them fluently. However, insights gained from the interview protocol reveal problems concerning how some subjects interpreted the instructions for the WAT. For example, it is apparent that they do not always provide the responses that occur to them. Further, there is evidence of chaining of responses where subjects respond to their previous response rather than the cue word. In sum, while the subjects agree that WAT20 is testing what it is supposed to be testing, in the light of these observations, the following study explores the need for and the effect of changing the task instructions.

REFERENCES

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education*. London: Routledge.

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

Dörnyei, Z. (2010). *Questionnaires in second language research. construction, administration, and processing*. New York: Routledge.

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.

Munby, I. (2007) Report on a free continuous word association test. *Gakuen Ronshu*, The Journal of Hokkai Gakuen University 132, 43–78.

Munby, I. (2008) Report on a free continuous word association test. Part 2. *Gakuen Ronshu*, The Journal of Hokkai Gakuen University 135, 55–74.

Munby, I. (2018) Report on a free continuous word association test. Part 3. *Gakuen Ronshu*, The Journal of Hokkai Gakuen University 175, 53–75.

Munby, I. (2019a) Report on a free continuous word association test. Part 4. Comparing Kruse with WAT10. *Gakuen Ronshu*, The Journal of Hokkai Gakuen University 178, 107–119.

Munby, I. (2019b) Report on a free continuous word association test. Part 5. Further development of WAT20. *Gakuen Ronshu* The Journal of Hokkai Gakuen University. no. 179. 51–66

Munby, I. (2019c) Report on a free continuous word association test. Part 6. Longitudinal study of WAT20. *Gakuen Ronshu* The Journal of Hokkai Gakuen University. no. 179, 67–75

Nation, I.S.P. (2007). Fundamental issues in modeling and assessing vocabulary knowledge. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp.35–43). Cambridge: Cambridge University Press.

Sakui, K. & Gaies, S. J. (1999). Investigating Japanese learners' beliefs about language learning. *System*, *27* (4), 473–492.

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*, 315–329.

Vivas, L., Manoiloff, L., García, A. M., Lizarralde, F., & Vivas, J. (2018). Core semantic links or lexical associations: Assessing the nature of responses in word association tasks. *Journal of Psycholinguistic Research*, *48*(1), 243–256

Appendix 9: WAT attitude and awareness questionnaire

Questionnaire. Name_____ Date _____

Please read the statements and check the boxes 5-1.

5 = Strongly Agree

4 = Agree

3 = Neither agree nor disagree

2 = Disagree

1 = Strongly disagree

	5	4	3	2	1
1. I like this kind of activity					
2. I prefer writing English sentences					
3. I tried as hard as I could					
4. I sometimes didn't write a word in case it was the wrong response					
5. If I knew more words I'd be better at the test					
6. If I could think of words more quickly I'd be better at the activity					
7. I couldn't think of enough words associated with the cue, so sometimes I just wrote down any English words I could think of.					
8. When I try to think of a response, I translate into Japanese in my head					
9. Sometimes I couldn't write a word because I didn't know the English word.					
10. It was more difficult to think of responses for some words than others.					
11. Sometimes my mind went blank. I got stuck					
12. I knew the meaning of all the cue words					
Do you have any other comments about the test, or about how you completed it?					