HOKUGA 北海学園学術情報リポジトリ
学校法人北海学園
北海学園大学
北海商科大学

タイトル	Report on a free continuous word association test (part 6): Longitudinal study of WAT20
著者	MUNBY, Ian
引用	北海学園大学学園論集(179): 67-75
発行日	2019-07-25

Hokkai-Gakuen Organization of Knowledge Ubiquitous through Gaining Archives.

Report on a free continuous word association test (part 6): Longitudinal study of WAT20

Ian MUNBY

INTRODUCTION

The aim of this study, the sixth in a series of studies investigating the word association behaviour of adult Japanese learners of English, is to determine whether or not learner performance in WAT20 changes over time. In the previous study (Munby, 2019b), there were significant, positive correlations between the WAT20 scores of a group of 111 Japanese users of English and their scores on a test of receptive vocabulary size (the EVST, Meara & Jones, 1998) and a translation test of controlled productive vocabulary knowledge (adapted from Webb, 2008). However, we should remember that this is a cross-sectional analysis, with only the implication that individual learners are likely to demonstrate gains in both vocabulary test scores and word association test scores when tested before and after a period of L2 study. I decided to test this hypothesis by conducting a longitudinal study to address the question of whether or not WAT20 can reflect changes in proficiency, or lexical processing skills, by measuring the WAT performance of the same group of learners at Time 1 (T1) and Time 2 (T2), following 16 weeks of study. If WAT20 is a valid test of an aspect, or aspects, of L2 ability, we could predict that learner WAT20 scores will improve after a learning intervention. For example, Bachman (1990) points out that: "evidence that indicates a relationship between test performance and the behaviour that is to be predicted provides support for the validity of this use of test results" (p.253). A word of caution is required here. Bachman (1990, p. 183), also recommends that in order to discount the possibility that these learners had merely benefitted from a practice effect, or the experience of taking the same test six months earlier, an alternate form, or parallel version of the WAT with different cues would be required. I decided not to use an alternate version of WAT20 out of concern for the quality of the remaining cues in WAT50 (Munby, 2018) since, as seen in Munby (2019a), cue choice does seem to influence performance.

This longitudinal approach was partially modeled on Schmitt & Meara (1997) who aimed to

measure changes over time in two types of receptive and productive word knowledge: word associations and grammatical suffix knowledge. They compared their subjects' performance on their experimental tests (knowledge of suffixes and associations) with the Levels test, a measure of receptive vocabulary size (Nation, 1990; Nation, 1983). There were three similarities between the methodology of Schmitt & Meara and the methodology adopted in this study. First, as in WAT20, their productive word association test also included 20 cue words. Second, word associations were measured for native-speaker-likeness with native-speaker-generated norms lists. Third, the subjects were young adult Japanese learners. Fourth, they were tested before and after a period of study of similar length (13 weeks in Schmitt & Meara's study compared with 16 weeks in this study).

Nevertheless, this study differed from their study in three ways. First, in this study, there was no measure of grammatical suffix knowledge; nor was the Levels test used. Instead, receptive vocabulary size was measured with the EVST (Eurocentres Vocabulary Size Test; Meara & Jones, 1990) and controlled productive vocabulary knowledge was measured with the translation test (adapted from Webb, 2008). Second, for their productive WAT, norms-based measurement was complemented with native-speaker ratings of associations, while in this study responses are only measured for native-speaker-likeness through norms. Third, their cue words were all verbs, and they were all different from the ones I have used in WAT20. They also included some low-frequency items, such as *disclose, quote,* and *stimulate*, which were selected specifically because they were unlikely to be known by all the non-native subjects taking the test. Fourth, a maximum of three responses were elicited to each cue. Finally, no time limit was imposed on the subjects for task completion. I will compare the results of this study with Schmitt & Meara's study in the discussion section which concludes this paper.

The main purpose of this study is to pick up changes, if any, in the WAT20 performance of a group of non-native subjects over time. This forms the basis of the first research question below.
RQ1 Do the WAT20 scores of non-native speakers change after 16 weeks of study?
In the previous study, significant, positive correlations were found between WAT20 and two measures of vocabulary knowledge: the EVST (Eurocentres Vocabulary Size Test; Meara & Jones, 1990), and the translation test, adapted from Webb (2008). In order to enable a further observation on this finding, I repeat the same correlational analysis in the context of this longitudinal study.
RQ2 Is there a significant, positive correlation between learner WAT20 scores (both number of

response and stereotypy measures) and the countermeasures?

Section 2: THE STUDY

Note the design of WAT20, the treatment of responses and scoring remains unchanged from the previous study.

8.2.1 Subjects, test design, and administration

69 of the 111 non-native subjects in the previous study, hereafter referred to as Test Time 1 (T1), were retested after 16 weeks of study (Test Time 2, or T2) following exactly the same procedure as in T1. They were drawn from five groups of first and second year Japanese students all of whom were majoring in English. Since 19 of the subjects had participated in the re-test in the previous study, thereby possibly affording them an advantage, their scores are not included in the analysis. None of the remaining 50 subjects had participated in the re-test reported in the previous study, and I use the scores of these 50 subjects from the T1study. While a total of 111 subjects had participated in T1, the pool of participants was reduced to 50 in T2 due to the absence of (i) 10 mostly higher level students who had been tested individually (i) one higher level group of 23, (iii) 19 lower level subjects who had participated in the retest, and (iv) 9 absentees from the five groups who did not attend the T2 testing sessions. Although the period of time between T1 and T2 was six months, I should point out that this period included nearly ten weeks of summer vacation time. Further, it is hard to quantify the exact number of hours of L2 study these students engaged in. They participated in about 16 weeks of course work, with six weeks of study before the summer break and 10 weeks after. Course work included a minimum of four ninety-minute lessons a week of skills-based lessons conducted entirely in English with native speaker instructors. They also completed about 2 hours a week of required independent study, and some of them took extra optional English classes. In total, they had completed around 100–200 hours of L2 instruction and independent learning per individual.

The test materials, task instructions, and scoring procedure remained unchanged from the previous study. To begin with the test materials, WAT20, with the same software used in all previous studies, presented subjects with the following 20 cue words in this alphabetical order: AIR BECOME BREAK CHOICE CHURCH CUT FREE GAS HEART KEEP KIND LEAD LINE MARRY PACK POINT POLICE SORRY SPELL SURPRISE.

Each session began with an orientation of how to use the software, and an explanation of the

rules in Japanese. 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. They were then invited to type in as many single English words as possible, up to twelve, in response to the cue word on the screen within 30 seconds of thinking time. They were told (i) that the timer deactivated while responses were being typed, (ii) that there were no right or wrong answers, (iii) not to worry about spelling mistakes, (iv) that they should press *enter* immediately after typing each response, and (v) not to use dictionaries. They were also advised to avoid (i) proper nouns, (ii) entering responses of more than one word, and (iii) "chaining away" from the cue word. The example of *cat* (cue), *mouse* (response 1), *cheese* (response 2), *biscuit*, *cake* was given. The participants were not told how their responses would be scored, and they were not warned in advance that they would be taking the test. In fact, it was not described as a test, but as a language learning activity. Immediately following WAT20, they took the EVST, a computerized test of receptive vocabulary size. The test takes about 10 minutes to complete and involves simply clicking on "yes" or "no" to indicate knowledge, or lack of knowledge, of around 150 lexical items that appear on the screen. Finally, they took the translation test of controlled productive vocabulary (adapted from Webb, 2008). This was scored in the same way as in T1, with one point awarded for each correct translation of a selection of 160 Japanese kanji (characters) in different frequency ranges. Misspellings were not penalized. All three tests were taken in a single session, as in T1.

2.2 Treatment of WAT responses and scoring

Following the WAT protocol established in Munby (2019a, 2019b), I corrected spelling mistakes and discarded proper nouns and a very small number of unidentifiable responses that were not listed in dictionaries. In cases where the same response was entered more than once to the same cue word, the repeated responses were deleted. Multi-word responses were clipped to single words, maintaining any single word which was listed in the normative data. The responses were then scored in two different ways: (i) a number of response measure (WAT A), and (ii) a stereotypy measure (WAT B). For the stereotypy measure, a score of one point is awarded for each response that matches a response on the norms lists of native speaker responses created in Munby (2018), excluding idiosyncratic responses.

Section 3: RESULTS

In this section, I report the descriptive statistics for this study (Table 1) and the Pearson correlation analysis (Table 2).

RQ1 Do the WAT20 scores of non-native speakers change after 16 weeks of study?

In line with expectations, on average, the results in Table 1 indicate that the group of 50 learners made gains in all four measures. A one-tailed paired t-test shows that, for WAT A, WAT B, and the translation test, mean scores at T2 were significantly higher than for T1. However, with the EVST, gains were not significant. This indicates that this group made significant gains in their ability to produce associations in greater number and native-like quality following a period of study. Individual performances are represented on the scatterplots in Figures 1 and 2 below. These show that the majority of the subjects placed above the diagonal or median line, indicating

Table 1
A comparison of the means and standard deviations of all test scores for T1 and T2 (n = 50)

	WAT A		WAT B		EVST		Translation	
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Mean	95.62	104.80	44.40	49.96	3375	3627	89.48	95.64
SD	37.53	40.72	13.13	13.75	1229	1161	17.84	18.80
Difference	+9.18		+5.56		+252		+6.16	
Maximum	240		240		10,000		160	
t value	1.385, p<.01		2.175, p<.0001		0.925, p = 0.07		3.108, p<.0001	

WAT A = number of responses, WAT B = stereotypy measures.

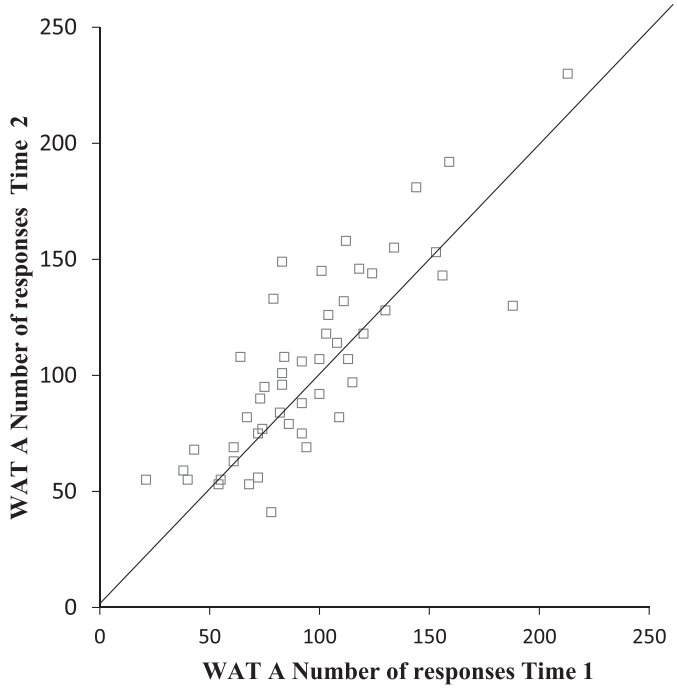

Figure 1 Comparison of WAT20 number of response scores in T1 and T2

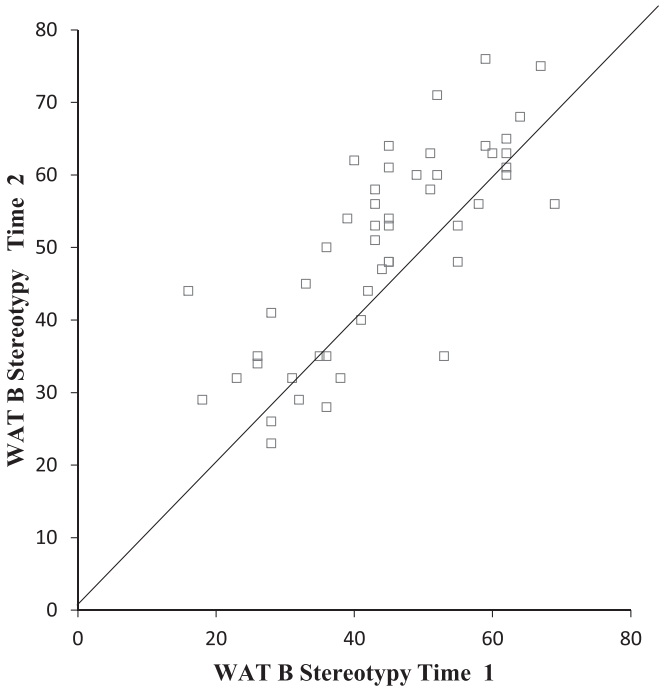

Figure 2 Comparison of WAT20 stereotypy scores at T1 and T2 (after 16 weeks of study).

Table 2
Pearson correlations for WAT20 between all scores for WAT A, WAT B, EVST, and the translation test (n = 50) at Time 1 (T1) and Time 2 (T2).

	EVST (T1)	Translation (T1)	EVST (T2)	Translation (T2)
WAT A	.289*	.311*	.105 ns	.135ns
WAT B	.357**	.519**	.156 ns	.464**

Key to rows: WAT A = number of responses, WAT B = stereotypy measures
Pearson 1-sided *p*-value:
**. Correlation is significant at the 0.01 level (1-tailed).
*. Correlation is significant at the 0.05 level (1-tailed).
ns = not significant.

that they scored higher on this WAT at T2 than at T1 in both the number of response measure and the stereotypy measure. However, both the association scores and the vocabulary countermeasure scores of some subjects decreased after a period of intervention.

RQ2 Is there a significant, positive correlation between learner WAT20 scores (both number of response and stereotypy measures) and the countermeasures?

Table 2 shows that there are positive, significant, but modest correlations between both WAT measures and the two countermeasures in all cases in Time 1. However, at T2, the only significant correlation is between WAT B (the stereotypy measure) and the translation test. Overall, all correlations among the WAT and the countermeasures are lower at T2 than at T1.

Section 4: DISCUSSION

This section includes commentary on the results of this study: (i) in the light of the research questions, (ii) in comparison with the findings of previous studies in this series, and (iii) similarities and differences between this study and Schmitt & Meara's study.

The main research question (RQ1) was to determine whether or not the WAT20 scores of non-native speakers change after 16 weeks of study. Accepting the need, highlighted by Bachman (1990), for a parallel or alternate version of this WAT, we can tentatively conclude that WAT20 reflects improvements in the English vocabulary, or indeed proficiency, in this group of learners probably resulting from 16 weeks L2 instruction. This significant mean gain in scores reflects expectations raised by the performance of WAT20 in the previous study. However, these WAT20 gains are mirrored in significant gains in the translation test, but not in the EVST (see Table 1). The second research question (RQ2) was: "Is there a significant, positive correlation between learner WAT20 scores (both number of response and stereotypy measures) and the countermeasures?" The results in Table 2 indicate that while all correlations are positive and significant at Time 1, at Time 2 only correlations between WAT B and the translation test are significant.

Comparing the results of this study with previous studies in this series, I focus on the following three observations that emerge from closer inspection of the statistics in Table 2. First, this is the sixth time in as many studies where the stereotypy measure proves to be a better predictor of proficiency, or both productive and receptive vocabulary knowledge, than the number of response measure. Second, the translation test produces higher correlations with both word association test measures (number of responses and stereotypy) than the EVST, as in the previous study. This might be due to the fact that both WAT20 and the translation test measure productive L2 vocabulary knowledge. The third observation is that correlations between the four measures were not as high as in the previous study. This is almost certainly due to the range of levels being narrower in this study, with only first and second years taking part, and with the absence of the majority of the highest and lowest level subjects. For example, in Munby (2019b) correlations for

the WAT stereotypy measure stood at $r = 0.791$, $p < .01$ with the translation test, and at $r = 0.706$, $p < .01$ with the EVST. According to Cohen et al. (2000), these correlations allow predictions to be made about performance on one test from scores on another for groups of students, but not for individuals. At T2 in this study, correlations are not at a level where such predictions can be made. It is not clear why this is, but the suggestion is that gains made by these subjects are uneven, non-systematic, or not in step with each other. However, there is an alternative explanation for this apparent instability in test scores. This is the possibility that evidence of increases and decreases in the scores of individual subjects reflects problems with the testing instruments used to measure changes in learner ability, rather than the progress or lack of progress in the learners themselves. This may apply to both WAT20 and the countermeasures. In other words, we could say that the testing instruments do not have the sensitivity to measure individual performances with accuracy.

As mentioned in the introduction, the methodology of the study by Schmitt & Meara bears some similarities to this one. Indeed, there were some similarities in the results. The most striking of these is that, overall, the subjects in Schmitt & Meara's study produced more native-like associations after a period of study, suggestive of an approach towards native-like associative behaviour. In addition, gains in association scores in both this study and Schmitt & Meara were matched by mean gains in receptive vocabulary size, although the testing instruments used were different, as commented in the introduction. However, as in this study, Schmitt & Meara find that both the association scores and the receptive vocabulary scores of some subjects decreased after a period of intervention. Clearly, further research is required to determine the reasons for this.

Section 5: CONCLUSION

In this study, I examined the potential of WAT20 to pick up changes in learner associative performance after a period of intervention. Results indicate that after 16 weeks of study, overall, this group of learners achieved higher scores on both WAT measures. Although higher scores were not achieved in every test by every learner, results of a t-test indicated that these gains were significant. This said, these conclusions must be tempered by the fact that the longitudinal gains in WAT scores achieved by this group were no larger than in the test-retest in the previous study where there was a much shorter period between test times (two weeks). For example, the overall longitudinal gains in the mean WAT20 stereotypy scores reported in Munby (2019a), from 44.40 to 49.96, were smaller than the gains achieved in the WAT20 test-retest (44.74 to 52.28). This comparison places question marks over any claim that WAT20 is a valid tool for measuring changes in L2 ability or, at least, developments in the lexical processing ability of a group of

learners. As mentioned earlier, a parallel forms version of this WAT is required to confirm longitudinal gains. In addition, either a period of longer than 16 weeks of intervention, or a more intensive period of L2 study in between tests, would be necessary to improve ability to the extent necessary to satisfactorily assess longitudinal gains in WAT20.

REFERENCES

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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

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

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. *Gakuen Ronshu*, The Journal of Hokkai Gakuen University 178, 107–119.

Munby, I. (2019b) Report on a free continuous word association test. Part 5. *Gakuen Ronshu*, The Journal of Hokkai Gakuen University 179, 51–66.

Nation, I.S.P. (1983). Teaching and testing vocabulary. *Guidelines, 5*(1), 12–25.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Rowley, MA: Newbury House.

Schmitt, N. (1998a). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning, 48*, 281–317.

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

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition, 30*(1), 79–95.