タイトル	Report on a free continuous word association test (part 5): Further development of WAT20
著者	MUNBY, Ian
引用	北海学園大学学園論集(179): 51-66
発行日	2019-07-25

Report on a free continuous word association test (part 5): Further development of WAT20

Ian MUNBY

INTRODUCTION

The findings of the previous study (Munby, 2019a) indicate that the new version of the WAT (WAT10) functioned more effectively than the Kruse WAT as a means of measuring proficiency through learner ability to produce multiple native-like word associations in timed conditions. The two key differences between the Kruse WAT and WAT10 thaht contributed to this finding were the introduction of a new set of cue words, and new native speaker norms of word association for measuring responses. However, there was one element of the testing instrument that required reconsideration, namely the number of items to present in the set of cue words. In the previous study, in order to directly compare the Kruse WAT with the new cues and norms, the number of cues was reduced from 50 (Munby, 2018) to 10. In this present study, the options were either to maintain the number of cues in the set at 10, or to reduce it, or to increase it by removing cue words or adding new ones from the selection in WAT50 in Munby (2018).

To inform the decision on the number of cues to present in the WAT, I considered both the literature on L2 productive word association tests and evidence from the data gathered thus far. To begin with the literature, besides eliciting different numbers of responses to each cue, researchers have also used an astonishing variety of numbers of cues in their word association tests. To list some examples, Lambert (1956) used 32 cues, Riegel & Zivian (1972) 40, Ruke-Dravina (1971) 4, Randall (1980) 50, Erdmenger (1985) 1, Schmitt & Meara (1997) 20, Schmitt (1998a) 17, Fitzpatrick (2006) 60, Wolter (2001) 96, Wolter (2002) 30, and Zareva (2005) 73. They also asked for different numbers of responses for each cue. Further, little has been written about the merits or demerits of using a smaller or larger number of cues with the notable exception of Kruse et al. (1987). In reaction to the low test-retest correlations achieved in their study, they commented: "Greater reliability would be achieved by extending the test" (p.152), which I take to mean extending the number of cues in the test. Since investigating the reliability of the new WAT

was one of the aims of this study, this was the first reason I decided to increase the number of cues in the set. For example, Cronbach's Alpha test requires comparison of two equal subsets of cues and response sets. Having a larger number of cues in each subset was likely to improve chances of confirming internal reliability. The second reason was that 10 cues, the number used in Kruse, may not have been sufficient to obtain a representative sample of learner associative competence or performance. For example, it was possible that some subjects, particularly lower level ones, may either produce too few responses to analyze, or not even know the core meaning of one or more cue words. Although this could be an argument for including easier cue words in the set, not knowing the core meaning of a cue word may not necessarily be the reason for a subject not providing any responses to it. Further, in previous studies, I noticed that a few subjects had misread cues, suggested by responses such as *dawn* for the cue *surprise* (misread as *sunrise*). Both situations seem likely to result in lower, or unrepresentative, WAT scores and influence results to a greater extent with a set of ten cues than with a larger set. The problem was that the analysis of individual cue performance in Munby (2018) hinted that some cues wielded less discriminatory power than others. For example, the top ten cues used in WAT 10 were selected according to their individual Pearson correlations (WAT B with TOEIC). These ranged from $r = .509$, $p < .01$ for *choice*, the most effective cue word, to $r = .425$, $p < .01$ for *sorry*, the tenth most effective. While the next set of 10 (ranking 11–20) produced correlations ranging from $r = .415$, $p < .01$ (*line*) to $r = .384$, $p < .01$ (*free*), the third set (21–30) produced a range of $r = .384$, $p < .01$ (*happen*) to $r = .327$, $p < .01$ (*ready*). Alternatively put, if the set were extended to thirty cue words, for example, some of the lower-ranking new cues may elicit response sets which do not differentiate well between learners of different levels, thereby reducing correlations for the whole set. Since there was a risk that extending the set of cues by a large number may result in a less sensitive testing instrument, I decided to increase the cue set to twenty items. From a practical point of view, a twenty-cue WAT was also appealing since this was the total number of cues to which the subjects responded in the previous study in Munby (2019a), allowing ample time to complete the countermeasures in one 90 minute session.

The aim of this study is to examine issues concerning the validity and reliability of WAT 20, the new WAT with 20 cue words. To begin with validity, since this is a new WAT, I decided to compare the performance of a learner group with that of a control group of native speakers for a third and final time in this series of studies. I also assess the concurrent validity of the test by correlating WAT20 scores with two measures of vocabulary knowledge: the EVST (Eurocentres Vocabulary Size Test; Meara & Jones, 1990), and the translation test based on Webb (2008). Details

of these measures are included in the next section.

The following two research questions are based on these aims.
RQ1 Is there a difference between the performance of native and non-native speakers on WAT20?
RQ2 Is there a significant, positive correlation between learner WAT20 scores (both number of response and stereotypy measures) and the proficiency countermeasures?

Reliability of the test is assessed from two different angles. First, I examine the split half reliability and calculate Cronbach Alpha reliability estimates for the learner WAT scores. Second, I examine test-retest reliability in learner WAT performance, as in Kruse and the approximate replication (Munby, 2018). The following two research questions are based on these aims.

RQ3 Does the WAT demonstrate internal reliability?
RQ4 Are non-native speaker WAT results consistent between test and re-test?

SECTION 2: THE STUDY- WAT20

In this section, I provide details of the subjects, the test design and administration, the treatment of responses and scoring.

2.1 Subjects, test design, and administration

The subjects were 111 young adult Japanese learners who took the tests in groups of 1–23. They represented a wide range of levels from first to fourth year university students, and post-graduate advanced users who had studied abroad for between 9 months and seven years. A few of these subjects would have qualified as highly proficient users of the kind who contributed to the Sapporo L2 norms list in Munby (2018). In addition, with a view to answering RQ1 (Is there a difference between the performance of native and non-native speakers on WAT20?) a control group of thirty-seven native speakers of English also took the WAT. None of these subjects had contributed to the norms lists by correspondence in Munby (2018). With the exception of four participants, all of the native-speaker participants were based in Japan, most on a long-term basis. The breakdown by nationality was as follows: fifteen from the USA, nine Canadians, six Australians, four British, two from Ireland, and one from New Zealand. Twenty-eight of them were English teachers in Japan and twenty-two had completed post-graduate degrees.

Regarding test materials, the WAT, using 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.
These test items were the twenty most effective cues from the fifty-cue-word WAT in Munby (2018). They include the best ten used in the previous study (Munby, 2019a), along with the next best ten cues from WAT 50 (Munby, 2018). As in the previous study, the cues were selected as a result of the following analysis. Subject stereotypy scores for each cue word are treated as an individual, or separate, test. In this way, a correlation between stereotypy scores for each cue word and the proficiency measures, in this case the subjects' TOEIC scores, can be calculated. The original fifty cue words from WAT50 can then be ranked for effectiveness in discriminating learners of different levels of proficiency, and allow for selection of the best-performing ones. As usual there were two pre-test practice items (*banana* and *dress*).

In order to answer RQ2, "Is there a significant, positive correlation between learner WAT20 scores (both number of response and stereotypy measures) and the countermeasures?" I used two vocabulary knowledge tests as proxy for proficiency measurement. TOEIC scores were unavailable for this study. The first was the EVST (Eurocentres Vocabulary Size Test; Meara & Jones, 1990), a computerized test of receptive vocabulary size that had not been used before in the empirical work in any of this series of studies. 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 one by one on the screen. The target words are chosen from a range of word frequencies, approximately 10 from each 1,000 word range from 0–10K. Points are also deducted for signaling knowledge of about 50 nonsense words which are mixed in with the real words. Upon completion, an estimate of the number of words known appears on the screen. Note this is a global figure, extrapolated out from the sample tested. I decided to use the EVST as a proficiency measure in this study for three reasons. First, it has been used by other researchers in published work in the field of vocabulary testing (e.g. Meara & Fitzpatrick, 2000). Second, the yes/no format has undergone extensive validation (e.g. Eckymans et al. 2007). Third, for practical reasons, the EVST was preferred to other tests of L2 vocabulary such as the VLT (Vocabulary Levels Test, Nation, 1983) which takes up to an hour to complete.

The second proficiency measure was an extended version of a paper-based translation test of productive vocabulary knowledge (adapted from Webb, 2008) that had yielded high correlations with the WAT measures in the previous study. In this test, the task for the subjects is to translate

into English a list of 160 Japanese *kanji* representing a range of frequency bands with increasingly rare words. In the previous study, three sets, or columns, of 40 *kanji* were used. To recap, this set of 120 single content words in L1 Japanese are translations of English words with 40 items from the full range of each of the following three frequency bands: 701st–1900th, 1,901st.–3,400th, and 3,401st–6,600th. Since a number of subjects scored close to the maximum score in the previous study, I felt that this test may not have the power to adequately differentiate the higher level students in this study from their lower level peers. For this reason, I decided to add a column of 40 Japanese kanji, with 10 each from the 6–7,000 K, 7,000–8,000 K, 8,000–9,000 K and 9,000–10,000 K levels of the BNC. 20 minutes were allowed to complete the test. I decided not to use the cloze test because there was no time available to complete it in one session together with the WAT, the two vocabulary tests, and a survey which I describe below. Additionally, the non-native subjects completed a survey of their attitudes and reactions to the WAT. The design principles and results of this survey shall be reported separately, along with insights gleaned from recorded post-task interviews with a small number of both native and non-native participants.

Each session began, as usual, with an orientation of how to use the software and an explanation of the instructions 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* etc. 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. Non-native subjects took the WAT first and then completed the questionnaire. Immediately following this, they took the EVST and then the translation test in a single session lasting nearly 90 minutes. In order to answer RQ4 (Are non-native speaker WAT results consistent between test and re-test?) 39 of the 111 non-native subjects participated in a retest of the WAT following a two week interval.

2.2 Treatment of WAT responses and scoring

In the same way as the two previous studies, Munby (2018 and 2019a), I corrected spelling mistakes and discarded proper nouns and a 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, the repeated responses were deleted. Multi-word responses were clipped to any single word that appeared on the norms list. For example, with the response: *police officer* for the cue *police*, one point is awarded for the response *officer*. The responses were then scored in two different ways: (i) a number of response measure, (ii) a stereotypy measure. For the stereotypy measure, a score of one point is awarded for each response that matches a response on the Sapporo L1 norms lists generated from native speaker informants in Munby (2018). Idiosyncratic responses were removed from the norms lists for two reasons. First, idiosyncratic responses are not strictly norms. Second, this use of the norms yielded higher correlations with proficiency measures in Munby (2019a) than when matches with idiosyncratic responses were included in the stereotypy scoring.

Section 3: RESULTS

In this section, I first present the descriptive statistics for this study (Table 1) and the correlational analysis (Table 2) with a view to answering the first two research questions concerning the validity of the WAT. To address the final two research questions concerning the reliability of the WAT, I then report the results of the split half reliability and calculate Cronbach Alpha reliability estimates for the learner WAT scores. Finally, I examine test-retest reliability in learner WAT performance.

RQ1 Is there a difference between the performance of native and non-native speakers on this WAT?

With reference to Table 1, on average, native speakers outperform non-natives. One-tailed unpaired t-tests confirm that the difference between the number of response scores for the two groups is significant at $t = 4.199$ ($p < .0001$) and significant at $t = 6.790$ ($p < .0001$) for stereotypy. Figures 1 and 2 feature a comparative representation of the distribution of scores for both WAT measures for the two subject groups: natives and non-natives. The numbers of non-native speakers scoring above the native mean in WAT A and WAT B are 14 and 2 respectively. However, none of the native speakers scored below the non-native mean in either WAT measure.

RQ2 Is there a significant, positive correlation between learner WAT20 scores (both number of

Table 1.
A comparison of the means and standard deviations of all test scores for all subjects.

	Mean	*SD*	Hi	Low	Max
WAT A (Non-native speakers)	105.4	51.8	229	16	240
WAT A (Native speakers)	182.8	37.2	235	107	240
WAT B (Non-native speakers)	47.8	22.6	127	5	240
WAT B (Native speakers)	104.8	20.4	141	68	240
EVST	3583.0	1564.0	8400	750	10,000
Translation test	94.6	25.2	151	41	160

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

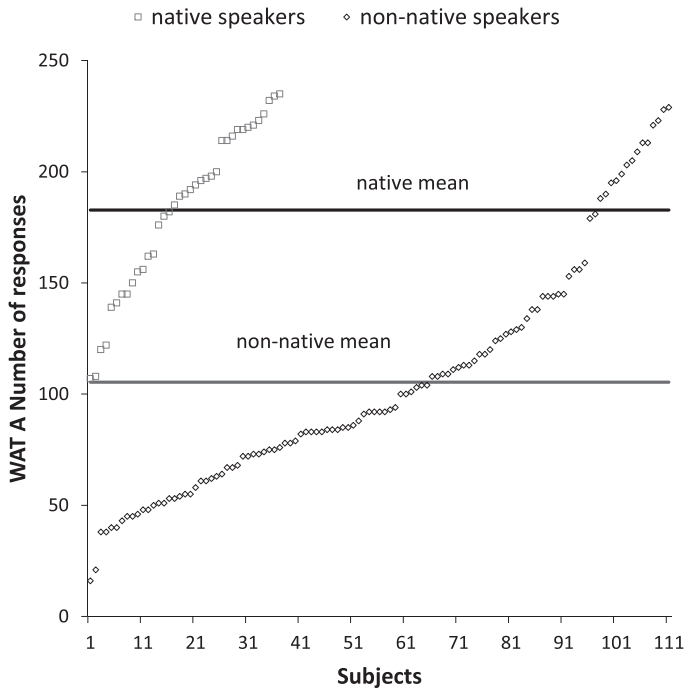

Fig. 1. Distribution of non-native and native speaker scores for the number of response measure (WAT A).

response and stereotypy measures) and the countermeasures?

Table 2 shows the relationship between the scores for the two WAT measures and the two vocabulary measures. Each correlation is significant and positive. Correlations between the stereotypy measure and the vocabulary measures indicate that test-takers with larger productive vocabulary knowledge (Translation test) and receptive vocabulary size (EVST) tend to produce a larger number of responses that match responses on a native-speaker generated norms list.

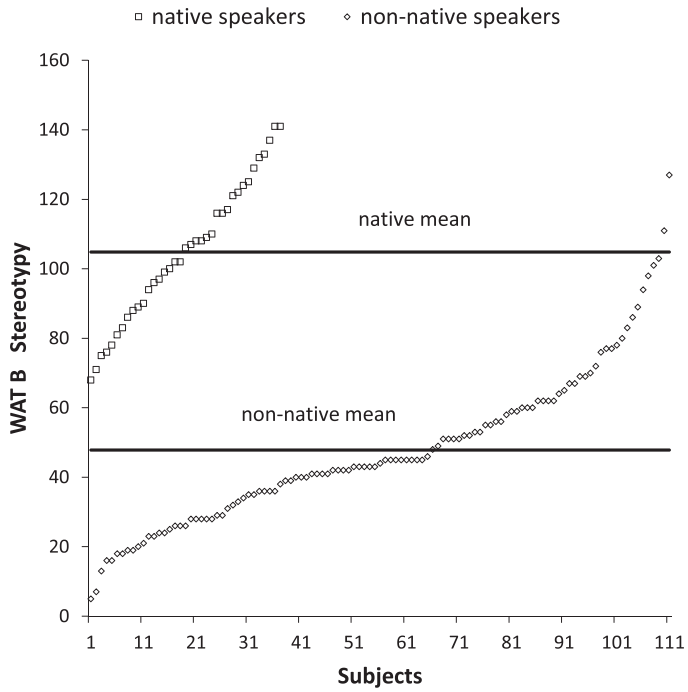

Fig. 2. Distribution of native speaker and non-native scores for the non-weighted stereotypy measure (WAT B).

Table 2
Pearson Correlations among WAT, EVST, and translation test scores

	EVST	Translation test
WAT A	.576**	.625**
WAT B	.706**	.791**
EVST		.808**

Key to rows: A = number of responses, B = stereotypy measure
Pearson 1-sided p-value: All significant at **p < 0.01

RQ3 Does the WAT demonstrate internal reliability?

In order to rule out the possibility that the subjects were producing responses in greater quantity and native-like quality in response to some cue words than for others, affecting the balance of the set, I performed a split-half reliability test to check results for internal consistency. In this analysis, following Bachman (1990, p.173) two parallel sets were constructed for both WAT measures. The "odd" set consisted of an analysis of the responses and scores for the odd numbered cues (first, third, fifth, etc, or *air, break, church* etc). The even set comprised an analysis of the responses and scores for the even numbered cues (second, fourth, sixth, or *become, choice, cut*).

With reference to Table 7.3, the correlations for both WAT measures indicate that the items in

Table 3
Means and standard deviations of odd- numbered and even-numbered split sets of 10 items each for the WAT Measures (n = 111), Pearson correlations between the two sets, and one-tailed paired t-test.

	ODD Mean (*SD*)	EVEN Mean (*SD*)	Correlations	t-value
WAT A	53.72 (26.0)	51.68 (26.3)	.961**	t = 1.481**
WAT B	24.12 (11.7)	23.67 (11.7)	.860**	t = 0.382 ns

Key to rows: WAT A = number of responses, WAT B = stereotypy measures
Pearson 1-sided p-value: Significant at **p < 0.01
ns = not significant.

Table 4
Mean scores, standard deviations, theoretical maximum for all tests, and 1-tailed paired t-test between means of WAT A and WAT B at Test Time 1 & 2 (n = 39).

	Test Time 1	Test Time 2		
	Mean (*SD*)	Mean (*SD*)	Maximum	*t* value
WAT A	93.72 (53.03)	115.03 (58.13)	240	t = 2.114, p < 0.0001
WAT B	44.74 (27.55)	52.28 (28.90)	240	t = 2.164, p < 0.0001
EVST	3400 (1550)	-	10,000	
Translation	93.6 (26.7)	-	160	

Key to rows: WAT A = number of responses, WAT B = stereotypy measure.

each sub-test set of 10 items were assessing word association ability in a similar way. Further, results of a t-test indicate that there is no significant difference between the means of the two sets for the WAT B measure. However, the same does not hold for WAT A where there is a significant difference between the means. For a further estimate of internal consistency, I calculated Cronbach's alpha. This yielded $\alpha = 0.949$ for WAT A, and $\alpha = 0.930$ for WAT B, suggesting that this WAT displays a high level of internal reliability (Bachman, 1990, p.184).

RQ4 Are non-native speaker WAT results consistent between test and re-test?
With reference to the mean scores for both WAT measures in Table 4, there is no doubt that, on average, the non-native subjects perform better on Test 2 compared with Test Time 1.
Results of a paired samples (dependent variables), one-tailed t-test between the means of both WAT measures at Test Time 1 and 2 are also significant, indicating that these gains are consistent, perhaps due to a practice effect that benefitted the majority of non-native participants. It is also worth noting that even at Test Time 2 the non-native means are lower than the native means. Further, test-retest correlations are at a level where it can be concluded that this WAT produces consistent results between test and retest with this group of non-native subjects (see Table 5).

Table 5
Pearson one-tailed correlations for test-retest reliability for WAT measures (n = 39).

WAT A: Number of responses	.844**
WAT B: Stereotypy	.927**

Significant at **p < 0.01

Key to rows: WAT A = number of responses, WAT B = stereotypy measures

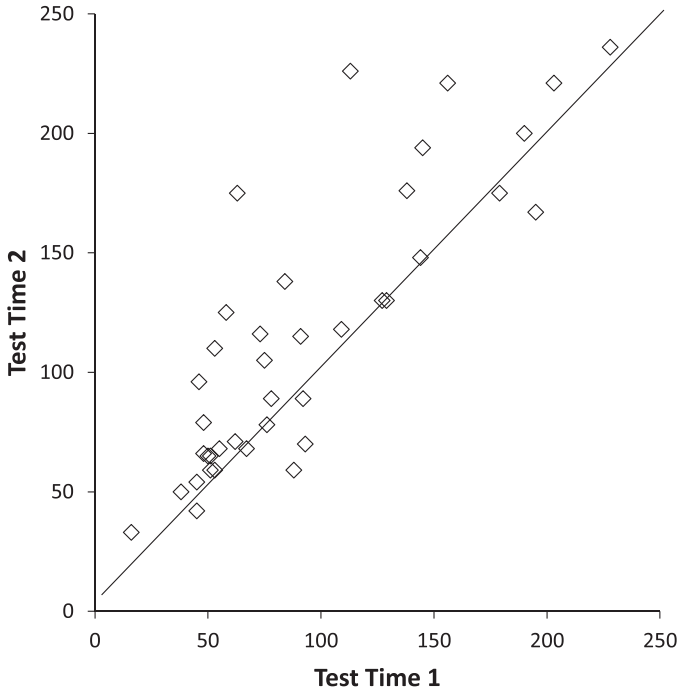

Figure 3 A comparison of non-native WAT A (number of response) scores in WAT Test Time 1 and Test Time 2.

See Figures 3 and 4 for scatterplot representations of test-retest performance on WAT20.
In summarizing the main results of the assessment of the reliability of the test in this study, the WAT measure did produce reliable data.

Section 4: DISCUSSION

In this section, I discuss the results of this study in relation to the research questions, and attempt to give reasons for the outcomes of RQ2: "Is there a significant, positive correlation between learner WAT20 scores (both number of response and stereotypy measures) and the countermeasures?" I also draw comparisons with the findings of the four previous studies (Munby, 2007, 2008, 2018, 2019a), and with the study by Kruse et al. (1987).

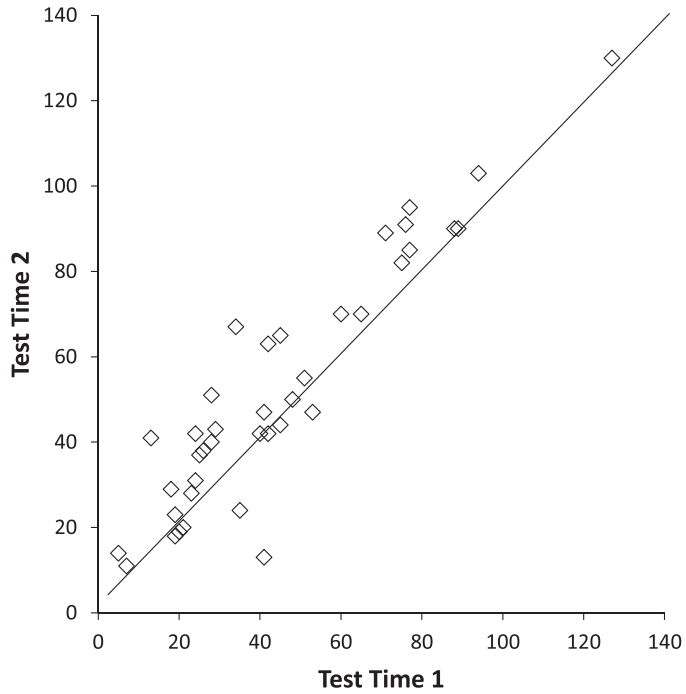

Figure 4. Comparison of non-native WAT B (stereotypy) scores in WAT Time 1 and Time 2.

With regard to RQ1 (Is there a difference between the performance of native and non-native speakers on WAT20?) the results presented here indicate that, on average, native speaker speakers outperform non-natives in both WAT measures. This is in line with findings from the two previous studies (Munby, 2007, 2008) that have also involved the participation of control groups of native speakers. Admittedly, the finding that native speakers are more native-like than non-natives is almost tautological given that the responses of native participants are being matched with native norms. On the other hand, there is also evidence in these two studies and the present study that the WAT scores of some higher level learners are equal to, or exceed the mean of the native speaker group. From this angle, the finding of Kruse that there is not much difference between native and non-native performance is hardly surprising in view of the fact that their Dutch subjects were presumably all at an advanced level of proficiency, and the WAT may not be sensitive at the top end of the proficiency scale. However, while there is a significant difference between the performances of the two groups on WAT20, there is also a large difference between individual performances within the groups. Historically, with native speakers, this variation in degree of stereotypy on word association tasks has been linked to mental disorders and has been used to identify schizophrenia, as in Kent & Rosanoff (1910), for example. To my knowledge, none

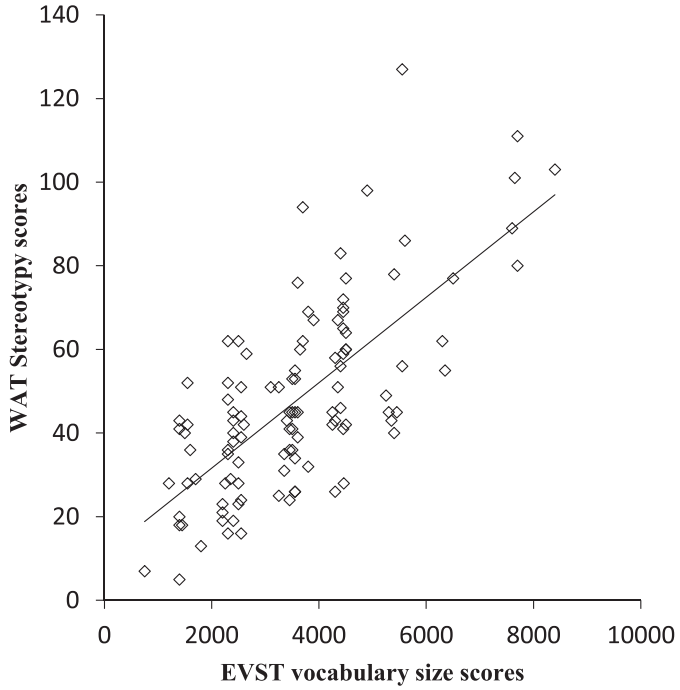

Figure 5 Comparison of WAT stereotypy and EVST scores ($r = .706$, $p < .01$)

of the 37 native participants in this study suffered from any mental disorder. Indeed, it is beyond the scope of this study to account for variation in native speaker performance in both the number of response measure and the stereotypy measure from a psychological viewpoint. As pointed out by Fitzpatrick (2007a, p.320), what is clear is that native speakers are not homogeneous in their performance on word association tasks.

With non-natives, differences in WAT performance are related to level of L2 proficiency. This brings us to the second research question, RQ2 "Is there a significant, positive correlation between learner WAT20 scores -both number of response and stereotypy measures- and the countermeasures?" Across all four previous studies, (Munby, 2007, 2008, 2018, 2019a), and the present one, this is the fifth observation of an important phenomenon: the stereotypy measure is more reflective of a test-taker's proficiency than the number of response measure. Further, correlations between WAT (non-weighted) stereotypy measures and proficiency countermeasures are always significant and positive, whereas this is not always the case with the number of response measure. With reference to the scatterplots in Figures 5 and 6, there appears to be a clear link between WAT performance on the native-like stereotypy measure and the two

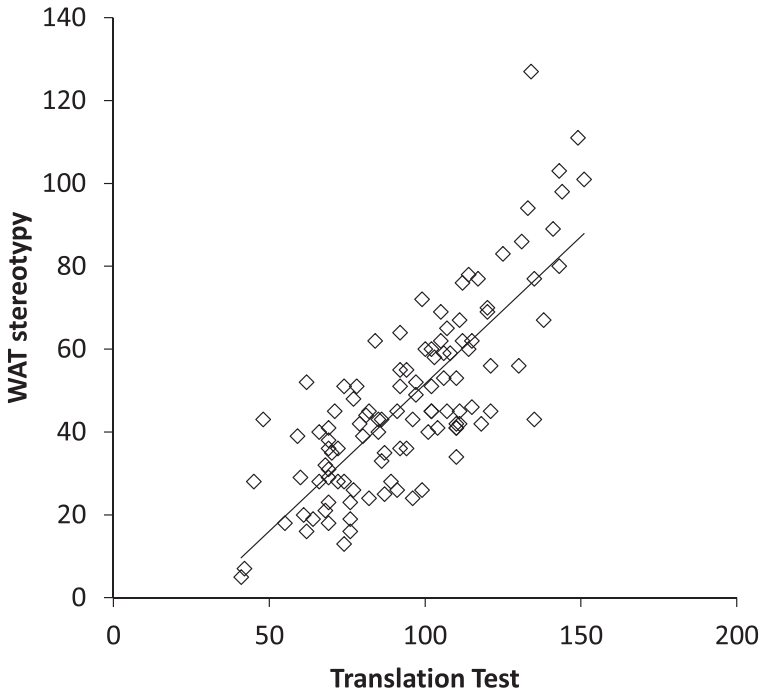

Figure 6 Comparison of WAT stereotypy and translation test scores ($r = .791$, $p < .01$)

vocabulary measures.

These correlations between WAT stereotypy and the two vocabulary tests (EVST and translation test), the proxy proficiency measures, are at a level where the performance of a group on one test can be largely predicted from its performance on the other. According to Cohen et al. (2000): "Nearer the top of the range ($r = 0.65–0.85$), group predictions can be made very accurately, usually predicting the proportion of successful candidates in selection problems within a very small margin of error" (p.202). The finding of correlations in the range described above indicates that WAT20 demonstrates a high degree of concurrent validity with two standard measures of vocabulary knowledge. In other words, WAT20 can differentiate learners of different levels of proficiency in a way similar to the two tests of L2 vocabulary knowledge. While this does not mean that WAT20 is a valid measure of proficiency in itself, it does suggest that WAT20 is measuring L2 vocabulary knowledge among other aspects of L2 ability. A fuller discussion of this phenomenon will appear in a separate study in answer to the question: What is WAT20 measuring?

In the previous study, it was not possible to produce unequivocal evidence as to precisely why WAT10 was more sensitive to proficiency than the Kruse WAT. Similarly, in this study, the same situation prevails in accounting for the higher correlations. For example, in the previous study, WAT10 stereotypy-translation test correlations stood at $r = .646$, $p < .01$, while in this study they are $r = .791$, $p < .01$ for the equivalent measures. I now consider two factors that possibly contributed to this finding. The first is that the non-native group represented a wider range of level of proficiency and that this could partially account for the higher correlations, as Brown (2005, p.161) suggests that it could do. The problem was that the translation test used in this study was different from that used in Munby (2019a) since it included an additional set of 40 low frequency items in the set of 160 items. In order to make a comparison with the WAT10 study in Munby (2019a), I removed scores from this extra set of 40 items and recalculated the mean scores for the group in this study of WAT20. With the 40 new items removed from the translation test, the mean translation score was 84.7 (*SD* 18.6), suggesting that the mean level was indeed slightly higher than the mean of 82.0 (*SD* 14.4) in the previous study. However, a t-test to compare these two sets of translation test scores shows that this difference was not statistically significant.

The second possible reason for the finding of higher correlations also concerns the translation test. Further analysis showed that the extended 160 item translation test correlates at $r = 0.791$ ($p < 0.01$) with WAT20 stereotypy. With the 40 new translation test items removed from the scoring, the equivalent correlation stands at $r = 0.769$ ($p < 0.01$). While this analysis indicates that extending the set of items in the translation test was a positive step, the difference in correlational strength was negligible. This leads us to consider a third possibility, that extending the set of cues in the WAT to 20 was the major factor contributing to the finding of higher WAT stereotypy-translation test correlations. However, it is not possible to provide clear evidence that this was the case.

With regard to RQ3 (Does the WAT demonstrate internal reliability?) one of the reasons for extending the set of cues to 20 was to improve conditions for demonstration of internal reliability. Results of this analysis support the claim that WAT20 demonstrates internal reliability, with the exception of the split-half reliability test for WAT A, where a significant difference was found between the means for the odd and even subsets. With RQ4 (Are non-native speaker WAT results consistent between test and re-test?), a retest of WAT20 demonstrated a high degree of reliability. As in Kruse and Munby (2007), this study shows that the non-native subjects increase their WAT scores in the retest, probably benefitting from a practice effect. It may also be possible to attribute

this to increased proficiency due to the fact that these subjects spent two weeks studying intensive English classes in between Test Time 1 and 2. However, as in Munby (2007), the test retest correlations between both the number of response measures and the stereotypy measures indicate that there is a higher degree of consistency in WAT performance than found by Kruse. One key difference emerges here. While the test-retest correlations for the "number of response" measure are higher than the stereotypy measure in Kruse and in Munby (2007), the reverse is true in this study. It is not clear why this is, but it may be a function of having an improved norms list for measuring responses.

Section 5: CONCLUSION

In this study, I began with a decision to continue with the new cue words and new norms lists from Munby (2018). Before examining issues concerning the validity and reliability of the new WAT, I increased the number of cues from 10 to 20. Concerning validity, native speakers were found to outperform non-natives on average, although native speaker performance was once again varied. Nonetheless, correlational analysis produces further clear evidence of a link between non-native WAT performance and proficiency, or L2 vocabulary knowledge, particularly with the native-like stereotypy measure. WAT20 also demonstrated a high degree of internal reliability, and analysis of non-native test-retest performance indicated that it yields consistent results. As mentioned in the description of this study in section 2, I conducted a post-task survey of non-native WAT task attitude and lexical processing awareness through a questionnaire, along with a small sample of recorded interviews with both natives and non-natives. Findings from the survey shall be reported separately, together with a discussion of how these findings informed the decision to make changes to WAT20 for an additional study.

REFERENCES

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

Brown, J. D. (2005). *Testing in language programs*. New York: McGraw-Hill.

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

Eckymans, J., Van de Velde, H., van Hout, R., & Boers, F. (2007). Learners' response behaviour in Yes/No vocabulary tests. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp.59–76). Cambridge: Cambridge University Press.

Erdmenger, M. (1985). Word acquisition and vocabulary structure in third year EFL learners. *International Review of Applied Linguistics, 23*(2), 159–164.

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

Fitzpatrick, T. (2007a). Word association patterns: unpacking the assumptions. *International Journal of Applied Linguistics*, *17*(3), 319–331.

Kent, G. H., & Rosanoff, A. J. (1910) A study of association in insanity. *American Journal of Insanity*, *67*, 37–96 and 317–390.

Kruse, H., Pankhurst J., & Sharwood-Smith, M. (1987). A multiple word association probe. *Studies in Second Language Acquisition*, *9*(2), 141–154.

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

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

Meara, P. M., & Fitzpatrick, T. (2000). Lex30: an improved method of assessing productive vocabulary in an L2. *System, 28*(1), 19–30.

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.

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

Randall, M. (1980). Word association behavior in learners of English as a foreign language. *Polyglot, 2*(2). B4-D1.

Riegel, K., & Zivian, I. (1972). A study of inter- and intralingual associations in English and German. *Language Learning*, *22*(1), 51–63.

Ruke-Dravina, V. (1971). Word associations in monolingual and multilingual individuals. *Linguistics*, *74*, 66–85.

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.

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.

Zareva, A. (2005). Models of lexical knowledge assessment of second language learners of English at higher levels of language proficiency. *System*, *33*, 547–562.