タイトル	Report on a free continuous word association test (part 4): Comparing Kruse with WAT10
著者	MUNBY, Ian
引用	北海学園大学学園論集(178): 107-119
発行日	2019-03-25

Report on a free continuous word association test (part 4): Comparing Kruse with WAT10

Ian Munby

INTRODUCTION

In Munby (2008), I investigated the effects of using different cue words from the Kent-Rosanoff list (1910) in a constructive replication of the WAT designed by Kruse et al. (1987), but found no convincing evidence that learner performance was affected. In Munby (2018), with the aim of developing a new, improved version of the WAT (WAT50), I described the process of selecting a new set of 50 cue words and the compilation of two separate, parallel norms lists for stereotypy scoring: one from a group of native speakers of English (the Sapporo L1 norms) and another from a group of highly proficient non-native (Japanese) users of English (the Sapporo L2 norms). Findings from this study showed that measuring learner responses for stereotypy with the L1 norms list yielded stronger correlations between WAT scores and proficiency measures than when responses are measured with the L2 norms lists. The highest correlation was $r = .601$ ($p < 0.01$) between the WAT stereotypy and the TOEIC scores of a group of 82 L1 Japanese learners.

At this stage, there is no evidence that WAT50 is more effective in discriminating learners of different levels than the original test developed by Kruse. For example, the strongest correlations between the WAT scores and the countermeasures in Kruse were $r = .576$, $p < .025$ (WAT A with grammar monitoring). However, we are now in a position to propose a 10-word set of cues, drawn from the set of 50 in the previous study, which, we hypothesize, will produce WAT responses which correlate more strongly with proficiency than the Kruse set of ten cues. In the next section, I shall describe the selection process of these cues for WAT10, the new 10-word set of cues, and the methodology for comparing them with the Kruse cues. The first and the main aim of this study is, therefore, to determine whether or not WAT10, with carefully selected, performance-tested cues and more extensive norms lists based on multiple responses, is more sensitive to proficiency than the Kruse WAT. From this aim, I formulate the following first research question to guide this

study:

RQ1 Does WAT10 yield higher correlations with the proficiency measures than the Kruse WAT?

The two main differences between WAT10 and the Kruse WAT are the norms and the cue words. However, it was not possible to tease apart the effect of the new WAT10 cues from the effect of the new L1 norms compiled in Munby (2018). This was because I had not asked the two groups of 114 participants in the Sapporo L1 norms and the Sapporo L2 norms to supply responses for the set of cue words used by Kruse in addition to the new set of 50 words I had chosen for the purposes of eliciting associations. Note here that norms for the cue words in WAT50 were not on the Kent-Rosanoff list and, therefore, native speaker norms of association for these cue words were not listed in the Postman & Keppel normative data. However, in the study described in Munby (2008), I had collected responses to the Kruse cues from the 50 native speakers in the control group. These subjects were from a similar population to that used for the Sapporo L1 norms lists which I will use here to score the new cue list in WAT10. With regard to the second aim of this study, in order to determine whether or not measuring learner responses with a different set of native speaker norms of association produces different results, I decided to rescore learner responses to the Kruse set using a new norms list compiled from the 50 native speakers in Munby (2018), referred to hereafter as the '2006' norms. This norms list is different from the Postman & Keppel norms because it is current, or up-to-date, based on multiple responses, and the word associations are elicited from the native speakers using the WAT software, i.e. in the same way as responses are elicited from the learners in test conditions. With this analysis, I will address the next research question:

RQ2 Does the Kruse WAT B measure yield higher correlations with the proficiency measures when re-scored with the 2006 norms?

Regarding scoring for associational stereotypy, one change from the three previous studies (Munby, 2007, 2008, and 2018) was that idiosyncratic responses (or responses only provided by one informant) on each of the norms lists (Sapporo L1 norms, the 2006 norms, and Postman & Keppel) were removed from the normative data. As discussed previously, this was because idiosyncratic responses are not essentially norms since they are not common to at least two respondents in the norming group. However, this position is debatable for two reasons. First, an idiosyncratic response, although in no sense a norm, can still be classed as native-like if provided by just one native speaker. Second, a response may be a norm in one group, but only idiosyncratic in another due to the size and characteristics of the group. In previous studies, scoring participant responses

with three different uses of the norms lists produced different results on the non-weighted (WAT B) and weighted stereotypy (WAT C) measures in terms of the correlations obtained with the proficiency countermeasures. Indeed, scoring responses for stereotypy with idiosyncratic responses included in the Postman & Keppel normative data (Use 3 of the norms) yielded higher correlations with proficiency than when responses were scored with idiosyncratic responses removed (Use 2), or with all but the 12 most common responses to each cue removed (Use 1).

In view of uncertainties in the debate as to what exactly constitutes a norm or a native-like association, and that different uses of the norms lists had been found to produce different results, I decided to rescore all WAT stereotypy measures with idiosyncratic responses included in the normative data. The aim was to determine if this rescoring affected correlations with proficiency measures. The results of this analysis shall be used to answer the third research question.
RQ3 Does the inclusion of idiosyncratic responses in the normative data yield higher correlations between WAT B scores and the proficiency measures?

Finally, in order to gain more information about which elements of L2 ability the multiple response WAT is measuring, I introduce a new countermeasure: a vocabulary test of controlled productivity based on Webb (2008), hereafter referred to as the translation test. Details of the design of this test appear in Section 2.2. Since the WAT also measures ability to produce words, I assumed that this translation test would correlate significantly and positively with it. Additionally, I decided to use two other proficiency measures: the cloze and the TOEIC. The fourth research question is based on the new measure, the translation test:
RQ4 Does the translation test correlate significantly and positively with the Kruse WAT and WAT10?

Section 2: THE STUDY

In this section, I shall begin by describing the cue selection process for identifying the 10 most effective cues for WAT10 from the set of 50 cues in the previous study, WAT50. I shall then provide details of the subjects, the test design (including the methodology for comparing the two WATs) and administration, and the treatment of responses and scoring. I also provide details of the additional proficiency measure, the translation test designed by Webb (2008), a carefully chosen and adapted measure of productive vocabulary knowledge to be used in conjunction with two previously used proficiency measures: the TOEIC and the cloze.

2.1 Selection of the 10 best cue words for WAT10

In order to establish the right conditions to allow for the comparison of the Kruse WAT and WAT10, I decided to run a WAT that combines the ten cues from the Kruse study with the ten best, or most effective, of the new cues in the previous study (WAT10) into a single WAT of 20 cue words. Note that although responses to the cue *man* were dropped from the scoring of the set of 10 cues in the Kruse study, I decided to gather and score learner responses for this prompt word because Kruse had intended to gather responses for this cue in their study. Responses to *man* were dropped from the final calculations because of an unspecified mistake. With the aim of selecting a comparison set of the 10 best cues from the previous study, each cue word from WAT50 in Munby (2018) was treated as an individual or separate test where subjects can score a maximum of 12 points by entering 12 responses that match responses on the L1 norms list. I decided to use the L1 norms list for the WAT stereotypy measurement because it produced higher correlations with the proficiency measures than the L2 norms list. In this way, for each cue word a Pearson correlation can be calculated between the stereotypy scores of each subject and her TOEIC score. In this analysis, TOEIC scores were used because they yielded higher correlations than the cloze test scores with WAT50 in Munby (2018). The implication here is that aspects of proficiency measured by the WAT are more similar to those tested by TOEIC than to any other test. Note that the same analysis was performed with individual cue words in Munby (2018). I add, once more, that the results should be treated with caution since this is a rudimentary way of assessing cues.

The 10 cue words that correlated the highest with the TOEIC scores were *choice*, *pack*, *break*, *air*, *police*, *keep*, *church*, *heart*, *lead*, and *sorry*. This set includes two cues that had been shown to be problematic in Munby (2018). First, *church* tends to elicit a number of proper nouns, and *lead* occasionally elicits responses related to *read* due to miscue. However, despite the issues described above, I decided to maintain them in the set due to evidence of their performance in the correlational analysis.

2.2 Subjects, test design, and administration

The subjects were 71 Japanese learners of English at tertiary level and included both first and second year students ranging in level from elementary to intermediate. This group of subjects has a similar profile to the group of subjects who participated in the previous study Munby (2018), with the main difference being that the mean proficiency score of this group was slightly lower (see Table 1 in the results section). For example, in Munby (2018), the mean cloze score was 18.5 (*SD*

7.2), and the mean TOEIC score was 539.2 (*SD* 137), while in this study scores for the equivalent measures were 15.9 (*SD* 8.8) and 508 (*SD* 140) respectively.

In this WAT, the subjects were presented with one set of 20 cue words combining the 10 Kruse cues and the new WAT10 cue words, marked in bold, in this order:

MAN **AIR** HIGH **BREAK** SICKNESS **CHOICE** SHORT **CHURCH** FRUIT **HEART** MUTTON **KEEP** PRIEST **LEAD** EATING **PACK** COMFORT **POLICE** ANGER **SORRY**

I interleaved cues from the two sets to limit any presentation order effect that might afford some advantage to either set. The task was to enter as many responses as possible, up to 12, to each of the 20 stimulus words. Subjects were advised to avoid proper nouns and chaining, to provide only single word responses, and to refrain from using dictionaries. As usual, there were two practice items: *gas* and *marry*.

There were three proficiency measures. First, there was the same 50-gap cloze test that has been used in all three previous studies. Second, as in Munby (2018), there was the annual in-house TOEIC test (listening and reading comprehension) which is marked and scored by the company ETS (Educational Testing Services). Finally, as mentioned in the introduction, I introduced a vocabulary test of controlled productivity for the first time. This is a test of productive vocabulary knowledge based on Webb (2008) where the task is to write English translations for a series of 120 single words of varying levels of word frequency written in L2 (Japanese). The aim was to determine whether or not this translation test of productive vocabulary produces higher correlations with the WAT than the regular proficiency tests (TOEIC and the cloze test). I felt that a test of productive vocabulary knowledge of this kind was more likely to tap into the same kind of vocabulary knowledge as the WAT, where the task is also to produce a series of single words in L2. The 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. In the original version of this translation test there were 180 items, but I decided to shorten this to 120 items (40 in each band) because there are 36 *katakana* words that are easily translated as transliterations of English loan words. 17 of these appear in the third band (3,401st–6,600th). This test was scored in 2 ways, as in the original paper by Webb: sensitive (soft), with spelling errors allowed, and strict, or hard, where only perfectly spelled answers are accepted. With the "soft" measure, I am interested in minimal productive word knowledge where credit is given for producing a recognizable, but not necessarily accurate

translation of the target word. The "hard" measure was found to produce lower correlations with proficiency measures than the "soft". For this reason, only scores for the "soft" scoring are reported. This matches the fact that the WAT is also "soft" scored since spelling errors are allowed.

The subjects took the WAT first, with 10 minutes allowed for instructions and demonstration, as in previous studies. The test took 20–30 minutes. On completion of the WAT, all subjects completed the cloze test. At the end of the week when the WAT and cloze sessions were conducted, the subjects took the TOEIC test. The same subjects completed the vocabulary test (15 minutes allowed) at the beginning of the following week's classes. Failure to complete two of the countermeasures, the TOEIC (15 subjects) and the translation test (6 subjects) meant that the final pool of subjects who took all four tests in this analysis was reduced to 71.

2.3 Treatment of responses and scoring

Since the aim was to compare the performance of the two sets of cues, responses for the Kruse cues and WAT10 cues were sorted for separate scoring. For each subject, a "number of response" score (WAT A) was obtained by summing the number of responses entered for each of the two sets of cue words. I obtained stereotypy measures from a straight count of the number of responses that matched words on the following norms lists. First, following the methodology of the original study by Kruse et al. (1987), the Kruse cue words were scored with the Postman & Keppel norms lists (1970). Second, the Kruse cue words were re-scored with the 2006 norms. Third, the WAT10 cue words were scored with the Sapporo L1 norms from the previous study in Munby (2018). Note that these responses were scored against the L1 norms list, not the L2 norms list because the former was found to produce higher correlations with the proficiency measures in the previous study. As mentioned in the introduction, all stereotypy scores for both WATs were initially scored with idiosyncratic responses removed from the normative data, and then a second time with idiosyncratic responses included to compare the effects on correlations.

Section 3: RESULTS

In this section, I report the descriptive statistics for WAT10 and the Kruse WAT (Table 1) Results of the initial correlational analysis appear in Table 2 and results of the reanalysis of scoring for stereotypy, with idiosyncratic responses included in all normative data, appear in Table 3. I then address the four research questions. The following are the keys to the abbreviations used in the columns of all Tables 1–3.

Key to abbreviations in the Tables 1–3

WAT A	The number of responses measure
WAT B	The native-like stereotypy measure
P&K	[WAT B stereotypy scored with] the Postman & Keppel norms lists
2006	[WAT B stereotypy scored with] the 2006 norms
Sapporo L1	[WAT B stereotypy scored with] the Sapporo L1 norms lists generated from native speakers that was created for measuring responses for stereotypy in Munby (2018)
(1)	[WAT B stereotypy scored] without idiosyncratic responses in the norms
(2)	[WAT B stereotypy scored] with idiosyncratic responses in the norms

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

		Mean	SD	Hi	Low	Max
Number of responses	Kruse WAT A	53.8	19.8	113	11	120
	WAT10 A	49.2	20.4	109	10	120
Stereotypy scores	Kruse WAT B P&K (1)	23.0	7.7	45	5	120
	Kruse WAT B 2006 (1)	25.9	8.6	43	8	120
	WAT10 B (1)	21.4	8.4	51	3	120
	Kruse WAT B P&K (2)	27.1	8.9	50	7	120
	Kruse WAT B 2006 (2)	31.2	10.2	54	9	120
	WAT10 B (2)	26.8	10.4	63	3	120
Proficiency measures	TOEIC	508	140	915	255	990
	Cloze	15.9	8.8	42	1	50
	Translation test	82.0	14.4	116	52	120

With reference to Table 2, correlational analysis indicates that the WAT10 stereotypy measure (WAT B) yields a closer relationship to all three proficiency countermeasures (TOEIC, cloze, and the translation test) than the Kruse WAT when scored with either the Postman & Keppel norms or the 2006 norms.

Table 2
Correlations between the word association test scores and the proficiency measures for the Kruse cues (scored with the Postman & Keppel norms and the 2006 norms) and the WAT10 cues and norms without idiosyncratic responses.

Cue list	Kruse	WAT10	Kruse	Kruse	WAT10
Measures	A	A	B (P&K)	B (2006)	B
TOEIC	.459**	.371**	.534**	.553**	.700**
Cloze	.425**	.310**	.520**	.528**	.662**
Translation	.533**	.394**	.606**	.604**	.676**

Pearson 1-sided p-value: All significant at **$p < 0.01$.

In contrast, the number of response measure (WAT A) for the Kruse cues correlates more strongly with all three proficiency countermeasures than the corresponding measure for the WAT10 cues.

Table 3

Correlations between the word association test scores and the proficiency measures for the Kruse cues (scored with the Postman & Keppel norms and the 2006 norms) and the WAT10 cues and Sapporo L1 norms with idiosyncratic responses included in all normative data.

Cue list	Kruse	Kruse	WAT10
Norms list	P&K	2006	Sapporo L1
TOEIC	.552**	.599**	.667**
Cloze	.526**	.561**	.622**
Translation	.617**	.636**	.646**

Pearson 1-sided *p*-value: All significant at **$p < 0.01$.

RQ1 Does WAT10 yield higher correlations with the proficiency measures than the Kruse WAT? However, since WAT10 stereotypy-proficiency correlations are the highest of all correlations in Tables 2 and 3, we can conclude that the new WAT (WAT10) is more sensitive to proficiency than the Kruse WAT. Further, the analysis in Table 3 indicates that this closer WAT10-proficiency link, evidenced in higher correlations on the stereotypy measure, also holds true when responses are scored with idiosyncratic responses included in all normative data.

RQ2 Does the Kruse WAT B measure yield higher correlations with the proficiency measures when re-scored with the 2006 norms?

With reference to Table 1, rescoring responses to the Kruse WAT with the 2006 norms lists resulted in higher WAT B scores. A paired samples one-tailed t-test shows that this difference is significant at $t = 3.047$, $p < 0.0001$ without idiosyncratic responses in the norms, and at $t = 4.372$, $p < 0.0001$ with idiosyncratic responses in the norms. Table 2 indicates that the 2006 norms yields a slight increase in Kruse WAT B-proficiency correlation strength with the TOEIC and Cloze test, but not with the translation test.

RQ3 Does the inclusion of idiosyncratic responses in the normative data yield higher correlations between WAT B scores and the proficiency measures?

A paired samples one-tailed t-test between the means in Table 1 indicates that rescoring the WAT B stereotypy measures, with idiosyncratic responses included, produces significantly higher scores for all three comparisons: (i) the Kruse WAT B scored with Postman & Keppel ($t = 6.928$,

$p < 0.0001$), (ii) the Kruse WAT B scored with the 2006 norms ($t = 7.352$, $p < 0.0001$), and (iii) WAT10 ($t = 7.327$, $p < 0.0001$). With reference to Table 3, for the Kruse cues, with idiosyncratic responses included in the Postman & Keppel norms, correlations between stereotypy measures and all three countermeasures increase slightly in strength. The same strengthening of correlations was also observed when the Kruse cues are re-scored for stereotypy with the 2006 norms including idiosyncratic responses. In contrast, with WAT10, WAT10-stereotypy correlations weaken as a result of rescoring with idiosyncratic responses included.

RQ4 Does the translation test correlate significantly and positively with the Kruse WAT and WAT10?
The results in Tables 2 and 3 indicate that the translation test yields significant and positive correlations with both the Kruse WAT and WAT10 with both WAT A and WAT B. The results also indicate that there is closer link between the translation test scores and the WAT10 stereotypy (WAT B) measures than with the cloze measure, but these are lower than the WAT10 stereotypy-TOEIC correlations.

Section 4: DISCUSSION

The main aim of this study was to determine which WAT, WAT10 or the Kruse WAT, yields a closer link to proficiency. The evidence indicates that, as predicted, it is WAT10. This difference is most apparent when we compare TOEIC test scores and stereotypy measures for the original Kruse WAT and WAT 10 (Table 2), and represent them on scatterplots (see Figures 1 and 2).

WAT10 differs from the Kruse WAT because (i) the cue words were first selected through a set of principled criteria and then chosen from a set of 50 based on their performance in Munby (2018) and (ii) the norms lists (Sapporo L1 norms) were current, based on multiple responses, and more extensive, in terms of the total number of responses listed, than the Postman & Keppel norms used in Kruse. In this section, I begin with some discussion concerning contributing factors underlying this finding. I also consider what aspects of L2 ability WAT10 may be measuring in the light of the results of this study and the previous three studies (Munby, 2008, 2018).

As stated in the introduction to this study, it was not possible to compare directly the performance of the Kruse and the WAT10 cue words because the norms lists used to score these WATs for stereotypy were different. This limits the extent to which we can draw conclusions on whether the cues or the norms lists have driven the higher correlations. Since the 2006 norms are

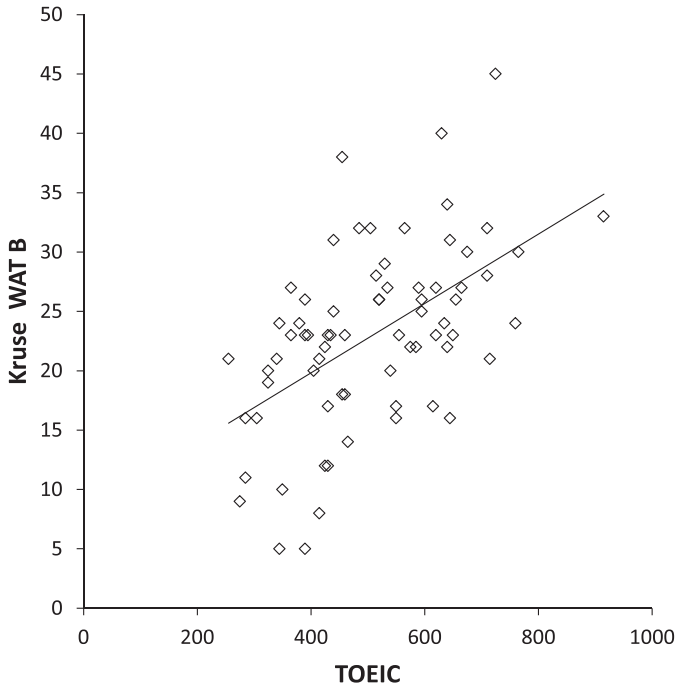

Figure 1 Comparison of Kruse WAT stereotypy scores and TOEIC scores ($r = .552$, $^{**}p < .01$).

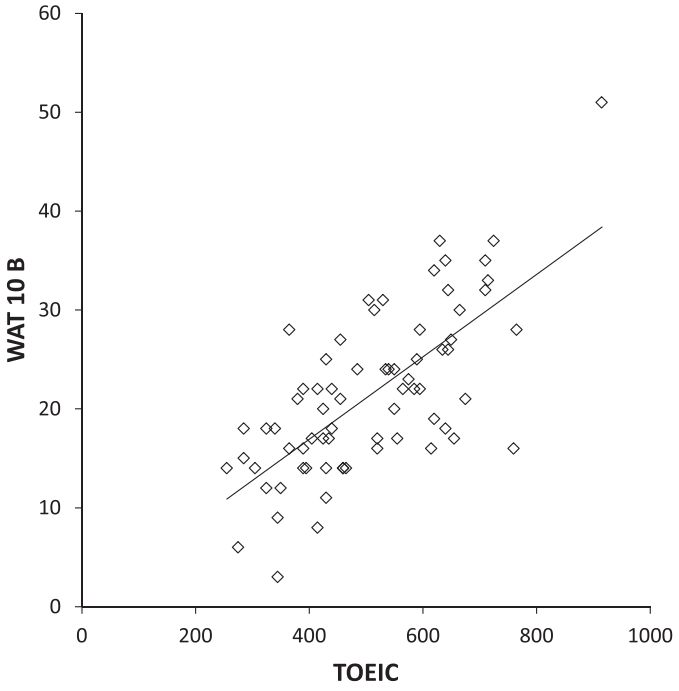

Figure 2 Comparison of WAT10 stereotypy scores and TOEIC scores ($r = .700$, $^{**}p < .01$).

similar to the Sapporo L1 norms, and the 2006 norms produce higher correlations for the Kruse cues than the Postman & Keppel norms, the implication is that both may contribute. However, with reference to Tables 2 and 3, since the difference between the level of WAT B-proficiency correlations for the 2006 norms and the Postman & Keppel norms is smaller than the difference between equivalent correlations for WAT10 stereotypy-proficiency, we could tentatively conclude that the norms lists contribute less than the cues in WAT sensitivity to L2 ability in this study.

Finally, it is not clear why the removal of idiosyncratic responses from the Sapporo L1 norms lists results in the strengthening of WAT10-proficiency correlations. This was not expected due to evidence presented in Munby (2018), where scoring learner responses with norms including idiosyncratic responses yielded higher correlations between stereotypy measures and proficiency measures than the same norms without idiosyncratic responses. However, for the purposes of future experiments, it would seem wise to proceed with the new set of cue words from WAT10 and to measure responses with the new L1 norms lists without idiosyncratic responses.

To begin the next part of the discussion regarding what WAT may be measuring, the results of this study are in line with the findings of previous studies in the following two ways. First, the WAT stereotypy measures for both Kruse and WAT10 yield higher correlations than the number of responses measure with all proficiency measures. The findings indicate that quality of responses, as measured by norms of native speaker association (stereotypy), reveals more about a learner's L2 ability than quantity of responses. Nevertheless, we can say that higher-level learners still tend to produce more responses than their lower level peers in timed conditions. We could interpret this in three different ways. First, one could take this as evidence that the number of response measure is a measure of fluency of lexical production. In other words, with gains in vocabulary knowledge, proficiency, or experience in language use, higher level learners are generally able to demonstrate more fluent or salient access to L2 vocabulary in their lexicons than their lower level peers. Secondly, production of larger numbers of responses could be interpreted as a function of a larger "word pool" or L2 lexicon. Finally, one could argue that connections between items in the lexical store of the higher level learner are more organized, and it is the native-like quality of these connections that allows for higher access speed. It is also possible that learner performance on the number of response measure is determined by a combination of the above factors and degree of motivation.

Comparing these results with the previous study (Munby, 2018) the TOEIC test once again produces higher correlations with the WAT stereotypy measures than the cloze test. The puzzle here is that one might expect the cloze and WAT10 to measure a larger number of elements of L2 ability common to both tests than the TOEIC and the WAT. This is because the cloze measures learner ability to produce words, while the TOEIC does not. There are also two ways in which this study sheds more new light on what this WAT may be measuring. First, the finding that the translation test yields correlations at a similar level as those produced by the TOEIC test indicates that the learners with larger L2 productive vocabularies will perform better on this WAT than learners who know fewer words. From a practical point of view, with only 15 minutes required to complete the translation test compared with 2 hours for the TOEIC, this is an encouraging development.

CONCLUSION

As it stands, there is evidence in all four studies reported so far that the WAT correlates positively and significantly with some standard proficiency measurement formats, such as listening and reading comprehension (as in TOEIC), cloze, and a test of productive vocabulary knowledge. The degree to which one can predict scores on the WAT from learner performance on these proficiency measures appears to depend to a great extent on the choice of cue words and the quality, or suitability, of the norms used to measure learner responses. Note that, according to Cohen et al (2000), correlations ranging from 0.65 to 0.85 "make possible group predictions that are accurate enough for most purposes" (p.191). Correlations between WAT stereotypy scores and the TOEIC, cloze, and translation test scores fall into this range with WAT10. The implication is that, with gains in proficiency, learners of English tend to move towards patterns of native speaker-like organization in associative performance. This tentative conclusion is consistent with claims that learning an L2 involves the gradual building of lexical networks that approach those of native speakers in terms of structure and dynamics.

REFERENCES

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

Jenkins, J.J. (1970). The 1952 Minnesota word association norms. In L. Postman & G. Keppel (Eds.). *Norms of word association*, (pp.1–38). New York: Academic Press.

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.

Munby, I. (2007) Report on a free continuous word association test. *Gakuen Ronshu*, The Journal of Hokkai Gakuen University 132, 55–74.

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.

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