タイトル	Report on a free continuous word association test (part 8): Changing the task instructions for WAT20
著者	Ian, MUNBY
引用	北海学園大学学園論集(188): 143-152
発行日	2022-07-25

Report on a free continuous word association test (part 8): Changing the task instructions for WAT20

Ian MUNBY

INTRODUCTION

In Munby (2019b, 2019c), there were indications that WAT20 demonstrated a degree of validity and reliability in its ability to differentiate learners of different levels of proficiency. For this reason, I decided to maintain the same set of cue words and norms lists used in these two previous studies. However, I made one major change to the methodology for this study. The task instructions that preceded the WAT in my previous studies were: "When you see or hear a word, it makes you think of another word. What English words do the prompt words make you think of? There are no right answers or wrong answers. Please type in as many responses as possible to each prompt word, up to a maximum of 12". As such, the WAT is a theory-based diagnostic test that elicits samples of associative behaviour with minimal constraints on the type of responses provided. In other words, the aim is to obtain a sample of connections between words in the mind of the learner in a naturalistic or unmediated condition and to use these for the purpose of diagnosing and assessing L2 lexical development. Note that this is the standard approach to eliciting responses in most word association research. For example, Van Rensbergen et al. (2016) describe response elicitation as follows: "In a word association task, participants respond with the first word(s) that come to mind after reading a certain cue word" (p1646).

For the purposes of this study, however, the task instructions were changed to: "I asked a large group of native speakers to provide single word responses in English to a set of cue words and made a big list. Type in the responses which you think the group of native speakers provided. You will score one point for each response that appears on the norms list". The key difference is that WAT20 now becomes a test of awareness of native speaker word associations rather than a test of free association. The reasons for making this change were due to the following issues. First, as mentioned in the conclusion to Munby (2022a), there was evidence in the post-task interview protocol that subjects do not always provide the responses that occur to them naturally. In other

words, free association does not necessarily tap into subconscious connections between words in the mind. Further, in the same protocol, there was evidence of chaining of responses where subjects respond to their previous response rather than the cue word. It was expected that revealing the purpose of the task, to provide norms-listed native speaker responses, would minimize the risk of chaining. Finally, it was possible that learners may demonstrate their ability to produce native-like responses more effectively if they are made aware of the way the test is scored.

With this new test task orientation towards awareness of native speaker-like, rather than free, associations, I decided to reintroduce an adapted version of the third system of scoring used by Kruse et al. (1987): the weighted stereotypy, or WAT C. In Munby (2007), I pointed out several problems associated with the system of weighting the scoring of responses for the degree of native-like stereotypy employed by Randall (1980), Kruse et al. (1987), Schmitt (1998b), and Wolter (2002). This dissatisfaction stemmed from this view expressed by Schmitt. "Although asking for 'the first word which comes to mind' is implicitly assumed to tap into the strongest mental connections between words in the mind, subjects will not necessarily give the most 'typical' response initially. They may well give an idiosyncratic response first and a very typical one second" Schmitt (1998b, p.391). In other words, testees will not necessarily provide native-like responses unless specifically asked to do so. Since in the study reported here, the task is to provide what the learners feel to be native-like responses, it is possible that higher-level learners may produce more native-like, or indeed more common native-like, responses compared with their lower-level counterparts due to the availability of larger lexical resources. They may also be able to demonstrate greater sensitivity to native-like patterns of association, having greater experience of, or more extensive exposure to L2. This leads to the second aim of the study: determining whether or not a weighted system of scoring (WAT C) yields a closer correlationship with the countermeasures than the other two WAT measures: the number of response measure (WAT A) and non-weighted stereotypy (WAT B). Details of this new weighted stereotypy scoring system are provided in the next section. In this study, for the first time, I also decided to present the cue words in lower case instead of upper case (e.g. 'air' instead of 'AIR'), since two lower-level subjects had reported experiencing difficulty in reading the cue words in large caps in Munby (2022a).

The aim of this study is to determine whether or not there is evidence of a relationship between L2 learner awareness of native-like associational networks and proficiency in the context of new task instructions for WAT20. Note here that the intention is not to make a direct

comparison of learner WAT performance with the new instructions with the old. A different methodology would be required here. However, I will examine the results for general signs of validity and reliability in the function of WAT20 under the new conditions. From this aim, I formulate the following two research questions:

RQ1 In the context of the new task instructions, is there a significant, positive correlation between learner WAT20 scores (both number of response and stereotypy measures) and the countermeasures?
RQ2 With modified instructions, which WAT measure: the number of responses (WAT A), non-weighted (WAT B), or weighted stereotypy (WAT C), yields the highest correlations with the translation test and cloze scores (if positive and significant)?

Section 2: METHOD

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

2.1 Subjects, test design, and administration

The subjects were 38 young adult Japanese learners who took the tests in groups of 1–19. They represented a wide range of levels from first to fourth-year university students and a post-graduate advanced user who had studied abroad for four years. Regarding test materials, the WAT, with the same software used in all previous studies, presented subjects with the following 20 cue words in this alphabetical order with two pre-test practice items (*boat* and *act*).

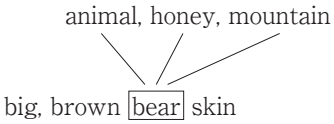
air become break choice church cut free gas heart keep kind lead line marry pack point police sorry spell surprise

These cue words are exactly the same as those used in Munby (2019b, 2019c) with the only difference being that they were presented in small case rather than capitals.

The new instructions (given in Japanese) are as follows:
i) “I asked a large group of native speakers to provide single word responses in English to a set of cue words and made a big list. One of the cue words was: *bear*. What responses do you think the group of native speakers provided? List your responses on a piece of paper. Do not worry about spelling mistakes and avoid proper nouns. You have one minute”. Note that I opted for a pencil and

paper practice - in addition to the two practice items mentioned above- because the training session had three distinct stages, involving stopping and starting in order to draw mind-maps for example. In practice, this was difficult to achieve with the WAT software.

ii) I then drew the following mind-map on the whiteboard using pre-prepared example responses from the norms lists:



I began by writing the position-based responses before and then after the cue word to indicate a linguistic relationship. I also explained that the responses could have a meaning-based relationship. For example, a bear is an animal that likes honey and lives in the mountains. Meaning-based responses are written above the cue word to suggest that the relationship with the cue word is semantic, or at least non-position-based or non-syntagmatic.

iii) This procedure is repeated with the following cue word: *fair*.
I chose the following example responses from the norms lists and wrote them on the board: *equal*, *unfair*, *fairly*, *fairness*. I then explained that a response could be a word with the same meaning (such as *equal*), a word with an opposite meaning (i.e. an antonym) in the case of *unfair*, or be derived from the same form (such as *fairly* or *fairness*).

iv) "Now write down on a piece of paper as many responses as you can to the cue *power* which you think are on my list. You have 30 seconds".

v) "Time up. Now here is the list of native speaker responses to *power*. How many did you guess?" Subjects' papers were not collected or scored. One reason for having the students check their responses against the norms list was to allow them to notice the large number of different scoring responses available for the cue *power* (87 in all). An additional reason was that it presented the testees with an opportunity to notice that proper nouns such as *Obama* and *USA* were not listed as scoring responses to the cue *power* because they were proper nouns. Additionally, the subjects could notice that all the native responses were single words. A final reason was that it allowed them a chance to appreciate that chaining away from the cue word might minimize their chances

of scoring points.

vi) "Now we're going to do the same activity using special software".
Then the original orientation and instructions from previous studies follow. Note that the following line: "There are no right or wrong answers" was omitted. As in previous studies, they were told (a) that the timer deactivated while responses were being typed, and (b) not to worry about spelling mistakes since they would be corrected and a point awarded if their response was recognizable. Upon completion of WAT20, the subjects took the translation test of productive vocabulary (adapted from Webb, 2008) that was used in Munby (2019b, 2019c). Following this, the subjects took the same 50 item cloze test that has been used in the studies reported in Munby (2007, 2008, 2018, 2019a).
As in all previous studies, the participants took the tests without prior knowledge that this was how they would be spending their class time. The WAT, the translation test, and cloze tests were taken in that order in one session lasting 90 minutes as usual.

2.2 Processing and scoring methods

Responses were processed in the same way as in previous studies. 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, the repeated responses were deleted. Multi-word responses were clipped to any single word that appeared on the norms lists and were counted as scoring on the stereotypy measure. Note here that I am using the Sapporo L1 English norms without idiosyncratic responses to score learner associations, as in Munby (2018, 2019a, 2019b, 2019c).

As in Munby (2007, 2008), there are now three scoring systems for WAT20, instead of two as in Munby (2018, 2019a, 2019b, 2019c). The first is the number of response measure (WAT A), a straight count of the total number of responses provided minus proper nouns, unidentifiable responses, or repeated responses to the same cue word. The second is the non-weighted stereotypy measure (WAT B). As in Munby (2019a, 2019b, 2019c), one point is awarded for each non-idiosyncratic response that is listed on the Sapporo L1 English norms. These norms were compiled from the responses of 114 native speakers in Munby (2018) and were the same lists used for scoring stereotypy in Munby (2018, 2019a, 2019b, 2019c). For the reasons outlined in the introduction, I decided to reintroduce a system of scoring for weighted stereotypy (WAT C). The aim was to determine whether or not the subject's ability to produce responses matching the more

common native speaker responses on the norms lists would be reflected in stronger correlations with proficiency countermeasures. Instead of adopting the system used by Kruse et al. (1987), or indeed any previous researchers, I decided to award 2 points instead of one for any response listed among the twelve most common responses on the norms list for each of the twenty cue words. This is because a maximum of 12 responses is being elicited for each cue word. In this weighted system of scoring, one point was also awarded to each response matching a response in the norms that fell outside the list of the most common responses. The new scoring system addresses the disproportionate nature of scores in the system of weighted stereotypy used by Kruse which I discussed in Munby (2007). One problem with this system was that some responses were provided an equal number of times by the native speakers on the norms list at the twelfth most common rank. For example, for the cue *point* there were five responses occupying the twelfth most common response rank. These were *arrow, indicate, rude, show, view*. This "overshoot" of the twelfth most common response affected 6 cue words, meaning that a total of 252 responses for the 20 cue words, instead of 240, were awarded two points. With this scoring system, the theoretical maximum weighted stereotypy score was 480 as opposed to 240 for the original non-weighted stereotypy score. This could only be achieved by supplying the maximum number of responses allowed (12) to each of 20 cue words (240 in total) that matched the most common native responses (scoring 2 points each).

2.3 Results

In this section, I report the descriptive statistics for all measures in this study in Table 1 and the correlational analysis in Table 2.

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

Non-native speakers (n = 38)					
	Mean	*SD*	Hi	Low	Max
WAT A	122.55	52.31	229	17	240
WAT B	58.55	24.30	103	12	240
WAT C	85.74	36.00	152	18	480
Translation	98.95	22.70	147	51	160
Cloze	12.97	8.83	31	1	50

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

Table 2
Correlations between all scores for the WAT, cloze, and the translation test for non-native speakers (n = 38).

	Cloze	Translation
WAT A Number of responses	.352*	.495**
WAT B Non-weighted stereotypy	.620**	.725**
WAT C Weighted stereotypy	.618**	.727**
Cloze		.788**

Pearson 1-sided *p*-value (1-tailed):
**. Correlation is significant at the 0.01 level.
*. Correlation is significant at the 0.05 level.

Section 4: DISCUSSION

In this section, I provide answers to the research questions put forward in the introduction and comment on these results in relation to findings in my earlier studies. The final part of the discussion addresses the merits and demerits of the new task instructions. With respect to RQ1 (In the context of the new task instructions, is there a significant, positive correlation between learner WAT20 scores -both number of response and stereotypy measures- and the countermeasures?) the analysis in Table 2 indicates that there is. Regarding the two stereotypy measures (WAT B and WAT C), these results suggest that more proficient learners are better able to perceive correctly and provide native-like word associations in WAT20 than their lower-level peers. In line with findings from Munby (2019a, 2019c) the translation test produces higher correlations with all WAT measures than the cloze test. This allows for the conclusion that WAT20, with both old and new instructions (in this study), has more in common with the translation test than the cloze test in terms of the shared aspects of L2 ability that it is measuring.

The following is in answer to RQ2: "With modified instructions, which WAT measure: the number of responses (WAT A), non-weighted (WAT B), or weighted stereotypy (WAT C), yields the highest correlations with the translation test and cloze scores (if positive and significant)?" With reference to Table 2, there is hardly any difference at all in the strength of correlations with the countermeasures between WAT B (non-weighted stereotypy) and WAT C (weighted). Admittedly, while correlations between weighted stereotypy and translation test were slightly higher than the equivalent correlations with non-weighted stereotypy, the reverse was true with the cloze.

The suggestion is that the new system of weighting the stereotypy scores does not represent an improvement on the non-weighted system of scoring, even where the aim for the testee is to produce native-like responses. However, both weighted and non-weighted stereotypy measures are better predictors of controlled productive vocabulary knowledge (translation test) and cloze test performance than the number of response measure. There is now an accumulating weight of evidence across all 7 studies in this series of papers of the following two phenomena. First, the non-weighted stereotypy measures generally correlate significantly and positively with the countermeasures. Note that the only exception was with the EVST in Test time 2 in Munby (2019c). From this, we can conclude that young Japanese adult learners of English as an L2 generally approach "native-like" associative behavior as their ability in English increases. Second,

the WAT stereotypy measure is always more sensitive to proficiency than the WAT number of response measure. In contrast to the findings of Kruse, this supports the use of native-speaker norms to assess the quality of learner associations. Nevertheless, regarding scoring associative responses for stereotypy, Fitzpatrick and Thwaites (2018) conclude with a note of caution that "despite .... sophisticated and dogged attempts by .... researchers to find a method of stereotypy scoring that can reliably evaluate learner proficiency or word knowledge, this is likely to be an unattainable goal" (p.14). Among reasons given is the claim that usage-based assessment of responses is essentially dependent on an "individual's linguistic system and is the result of their unique experience with language" (p.14).

I conclude this discussion with some thoughts about the way the new instructions and orientation procedure (using mind-maps) may alter the underlying construct of the WAT. I also consider whether or not they are appropriate ways of addressing the main aim and research question of this series of studies (to establish whether or not there is a link between WAT performance and proficiency). Although there was no evidence of problems arising from the new instructions in this WAT, from a theoretical standpoint, several factors need to be taken into consideration. In the introduction, I pointed out some possible advantages to inviting the subjects to write responses that they felt native speakers would provide. One such advantage is that the new instructions allowed for transparency in both the purpose of the test and its scoring mechanism. It also provided the test-takers with an opportunity to appreciate the reasons underlying recommendations to avoid practices such as chaining away from the cue word. In this way, the word association activity bears a closer resemblance to a language test involving, or at least inviting, the use of test-taking strategies. For example, in comments volunteered on the questionnaire reported in Munby (2022a), some subjects reported using strategies anyway rather than simply waiting for responses to become available or activate themselves. For instance, one 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*".

However, we need to be cautious about our assessment of the new instructions for two reasons. First, if the aim is to produce native-like associations, the instructions may be imposing some constraints on the test-taker with the result that it is no longer, strictly speaking, a test of free association. On the other hand, neither is it a word association test under restricted conditions of the kind investigated by Riegel & Zivian (1970) wherein subjects are asked to provide "following words", or collocations, for example. In theory, the new instructions focus on awareness of native

speaker associative behavior and access a different kind of L2 language knowledge. Additionally, this kind of awareness is probably important for language learners, especially since becoming more native-like may be the stated goal of the learner. The problem is that since the L2 lexicon will not necessarily be the same as the L1 lexicon, as pointed out by Grosjean (1989) and Meara (1996a), the new instructions may not have any effect on the number and native-like quality of responses that a learner produces.

The second theoretical issue is that the system of scoring for stereotypy is underpinned by the notion that there really are "correct" or "incorrect" associations. For example, informing a learner that the response *important* to the cue *point* is not scoring may be damaging. In terms of their lexical development, learners may become reluctant to use the combination *important point*, although most native speakers would be likely to accept it as a common collocation. In this study, the situation was avoided because scored WATs were not returned to the subjects. On the other hand, it seems wrong that the old instructions simply encourage the subjects to provide as many responses as possible when it is the quality of their responses rather than their quantity that best indicates the learner's level of proficiency. In this respect, instructing the learners to focus on the quality of their responses by attempting to achieve matches with native speaker norms may be justifiable and better suited to the purpose of the task. All in all, since establishing the optimal conditions for the WAT to reflect level of proficiency is the stated goal of this series of studies it seems essential to determine which set of instructions, old or new, are most effective in eliciting samples of learner associations that best reflect their abilities. However, besides simply discontinuing one set and proceeding with the other, the most interesting issue is that it is possible to measure two different constructs: learner associations and learner perceptions of native associations. These might each relate differently to proficiency, and reveal something different about proficiency, or different facets of it.

Section 5: CONCLUSION

The aim of this study was to investigate whether or not WAT20 with new task instructions functioned in a similar way to WAT20 in Munby (2019b, 2019c). Correlations between learner WAT20 scores (both number of response and stereotypy measures) and the proficiency countermeasures were significant and positive. However, the re-introduction of a system of scoring for weighted stereotypy (WAT C) did not yield higher correlations with proficiency than the non-weighted (WAT B) system. In fact, there was almost no difference in the correlations they produced. As in all previous studies, the number of responses (WAT A) was less sensitive to

proficiency than the stereotypy measure. Regarding the relative merits and demerits of the new instructions, in view of the fact that there are both ethical and practical reasons for revealing the purpose of the test, as stated in the introduction, I would be encouraged to use these new instructions in future studies when comparing WAT20 performance with proficiency. Before doing so, I would investigate whether or not the task instructions do influence learner WAT performance in a new study.

REFERENCES

Fitzpatrick, T., & Thwaites, P. (2020). Word association research and the L2 Lexicon. *Language Teaching*, *53*(3), 237–274.

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, *36*, 3–15.

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

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

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 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 179, 67–75

Munby, I. (2022a). Report on a free continuous word association test. Part 7. Qualitative analysis of WAT20 behavior based on post-task questionnaires and interviews. *Gakuen Ronshu* The Journal of Hokkai Gakuen University 188, 125–142. July, 2022.

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.

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

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

Van Rensbergen, B., De Deyne, S., & Storms, G. (2015). Estimating affective word covariates using word association data. *Behavior Research Methods*, *48*(4), 1644–1652.