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

タイトル	Report on a free continuous word association test (part 3)
著者	MUNBY, Ian
引用	北海学園大学学園論集(175): 53-75
発行日	2018-03-25

Hokkai-Gakuen Organization of Knowledge Ubiquitous through Gaining Archives.

Report on a free continuous word association test (part 3)

Ian MUNBY

INTRODUCTION

This paper describes the development of WAT50, a constructive replication of the WAT with new cue words and new norms lists. Note that a summary of this study appears in Higginbotham, Munby, and Racine (2015). As indicated in my two previous papers in this journal (Munby, 2007 and 2008), non-native speaker performance on a multiple response WAT correlates significantly, but only moderately, with standard proficiency tests. The suggestion is that gains in learner proficiency are reflected to a certain extent in the number and type of associations produced in response to a set of target words under timed conditions. However, it is possible that the methodology of both Kruse et al. (1987) and the two replication studies in (Munby, 2007 and 2008), have not lived up to their potential for two reasons, both of which concern the normative data used for measuring the degree of native-likeness of learner responses. First, the normative data used to measure both native and non-native speaker responses for stereotypy (Postman & Keppel, 1970) might be inappropriate (for reasons to be explained in the forthcoming paragraph) and should be replaced with a more suitable norms list of native speaker responses. Second, for reasons that shall be discussed later in this section, a norms list based on highly proficient non-native speakers might be more effective than one based on native speaker responses for the purpose of measuring the word associations of learners of English as an L2.

To begin with, the following three factors suggest that use of the Postman & Keppel norms lists for WAT stereotypy scoring should be discontinued. First, since the normative data is drawn from a collection of single or primary responses to the one hundred Kent-Rosanoff stimuli from 1,000 subjects, it seems likely that these lists fail to tap more distant or remote associations in the native speaker lexicon. For example, the response *cough* to the stimulus *sickness* does not appear on these norms lists. Although *cough* often appears amongst sets of native and non-native responses in both Munby (2007) and Munby (2008), it was never provided as a primary response. Second,

these norms lists may not contain a sufficient number of different responses per cue word to provide an adequate sample of native speaker-like association norms. This may be important in view of the broad range of different responses provided by both native and non-native speakers. With a limited range of response norms, it is possible that entering norms-listed, or scoring rather than non-scoring responses may be attributable to chance rather than language ability. To illustrate this, an experimental norms list was compiled from the data in Munby (2007) and Munby (2008) by combining the multiple responses of the 50 native speakers in the control group and comparing the number and range, or diversity of responses for two cue words: *sickness* and *anger*.

Table 1
Comparison of the total number of associations and the total number of different associations for the cues sickness *and* anger *contained in two norms lists: the Postman & Keppel norms lists, based on single responses and an experimental norms list based on multiple responses.*

Norms lists	Postman & Keppel (n=1000)	Ch3 (n=50)
Total no. of responses for:		
sickness	1,000	542
anger	1,000	512
Total no. of different responses for:		
sickness	76	220
anger	168	287

The evidence in Table 1 indicates that, for two example cues, *sickness* and *anger*, the Postman & Keppel norms contain a smaller number of different responses compared with the experimental norms lists because they are based on single, primary responses only. Third, these norms are outdated with the result that, for example, computer-related responses to cues like *memory* cannot earn points for stereotypy because these meanings did not exist at the time of norms list compilation. Alternatives to the Postman & Keppel norms, such as the EAT (Edinburgh Associative Thesaurus, 1973), are also unsuitable for similar reasons: only single responses are elicited, they are also out of date, and they contain a limited number of responses for each cue word. For example, with only 100 participants providing single responses, the EAT contains only 27 different responses for *sickness*.

The second reason that results of the previous three studies may be misleading is that the non-native speakers, in this case Japanese learners, as their level of proficiency increases, may be approaching the word association performance of highly proficient Japanese speakers of English rather than native speakers. Grosjean (1989) states: "The bilingual is NOT the sum of two

complete or incomplete monolinguals; rather he or she has unique and specific linguistic configuration" (p. 3). In the context of this study, this "unique and specific linguistic configuration" is a bilingual Japanese. Further, Schmitt and Meara (1997) point out that L2 learners will "have different mastery of the various kinds of word knowledge, with formal, grammatical, and meaning aspects probably learned first, and some other aspects, such as collocational behavior and register, perhaps never being mastered at all" (p. 18). Collocational competence is one aspect of ability to produce associations and if learners, even skilled L2 users, never master native-like associational knowledge, it may therefore be more appropriate to measure learner performance against proficient L2 user performance. In this sense, simply having a more extensive norms list compiled from native speaker multiple responses may not be the answer to the problem of non-native participants entering a large number of non-scoring responses. Evidence from the two previous studies (Munby, 2007 and 2008) suggests that around half the non-native responses do not match responses on the Postman & Keppel norms lists. Meara (1983) also notes that learner L2 responses tended to be more varied than those provided by native speakers. If adult native speakers do provide different types of responses from their L2 counterparts, it is plausible that a norms list compiled from highly proficient L2 user responses would be more appropriate. It may therefore produce scores on a free continuous WAT that correlate more strongly with proficiency measures.

There are, therefore, three key aims to this third study which could be described as an interim study to enable future selection of effective set of cues and norms. The first aim is to compile a set of fifty new cue words to gather responses for new normative data for a new WAT, hereafter referred to as WAT50. The decision to begin afresh with new cue words was motivated by a desire to limit the pool of candidate cue words to the 0-1K British National Corpus (BNC corpus, 2007). This was to increase the likelihood that all the cue words would be known to the non-native participants. Since none of the cue words that showed signs of being effective in (Munby, 2007 and 2008) according to the analysis reported in Table 4, such as *sickness* and *dream*, fell within this range of the 1,000 most common word families in the BNC, they were not recycled as cues in WAT50. The decision to select 50 cue words rather than 10 was in recognition of the fact that many items selected through criteria-based screening in (Munby, 2008), such as *joy, memory, spider*, and *window*, had not performed well (see Table 4). Recurrence of this phenomenon was therefore predicted in this study.

However, with 50 cue words, it was also predicted that a minority would merit inclusion in

smaller set of effective cues for use in future studies. The second aim is to compile two separate norms lists for this new set of cues with responses from two groups of participants: a group of native speakers of English and a group of highly proficient non-native (L1 Japanese) users of English. Since it is conventional to include reference to a place in the name of a norms list, as with the Edinburgh Associative Thesaurus (Kiss et al., 1973) and the Birkbeck norms (Moss, 1996), I decided to name the former "the Sapporo L1 English norms" and the latter "the Sapporo L2 English norms" (Munby, 2014). However, for the sake of brevity, I will often refer to the lists as L1 norms and L2 norms. The third aim is to run WAT50 with a group of learners, using these two new norms lists for separate stereotypy scoring.

From these aims, I formulate the following research questions to guide this study:

RQ1 Is there any evidence in learner WAT50 performance that the new set of cue words functions in the expected way according to the criteria by which they were selected?

Our next research question (RQ2) is motivated by a concern with the potential for the large number of cues in the test set (50) to invite a problem with a fatigue effect (Bachman, 1990, p. 24 & p. 160). In other words, learner performance on WAT50 may be affected by tiredness towards the end, therefore undermining the analysis of the results.

RQ2 Is there any evidence of a fatigue effect?

RQ3 Which norms list, the Sapporo L1 English norms or the Sapporo L2 English norms, yields the best match with learner responses?

RQ4 Which norms lists, the Sapporo L1 English norms or the Sapporo L2 English norms, yields the highest correlations with proficiency?

RQ5 Is 12 responses the optimal number of associations to elicit for each cue word?

The fifth and final research question for this study (RQ5 above) concerns the number of responses to elicit for each word. The decision to elicit up to 12 responses per cue word was a radical departure from the methodology employed by Randall (1980), who elicited a maximum of 5. Wolter (2005) described the twelve-response maximum as "probably too many responses" (p. 22). The methodology section (4.3) describes a technique for establishing the optimal number of responses in WAT50.

Section 2: SELECTION OF CUE WORDS

In this section, I describe the criteria for selection process that led to the creation of a new set of 50 stimuli for WAT50.

2.1 Cue word selection criteria

Cue selection criteria from the previous study were modified slightly for this study. Criterion (f) from the previous study (Is the stimulus likely to generate too many predictable responses?) was omitted because it was excessively subjective and proved difficult to work with in making judgments about cue words. This item was replaced by a new criterion (f) for reasons explained below.

(a) The stimulus is known to even the lowest level subjects taking the tests.
The cues *mutton* and *priest* were unknown to many subjects in Munby (2007), resulting in many subjects being unable to produce any associations for these cues. Although the list of candidate words will be restricted to the 0-1K band of the BNC, some of these 0-1K words, such as *vote* may be unknown to many lower level participants. Personal intuition based on extensive experience of teaching in Japan was used to determine which words were likely to be known and which were not.

(b) The stimulus does not seem likely to produce a "dominant primary" response, such as an adjective or other word that produces its polar opposite (e.g. *high-low*) or a noun which is marked for sex which tends to produce its polar opposite in response (e.g. *king-queen*).

(c) The stimulus is not likely to generate responses through highly predicable lexical subset relationships, such as *fruit-apple*.

(d) The stimulus word is not a proper noun. Some words on the 0-1K BNC list are proper nouns such as *Germany* and *America*.

(e) The stimulus is not likely to elicit proper nouns, such as *river-Mississippi, city-Minneapolis, ocean –Pacific*.

(f) The stimulus does not have a phonological equivalent in the L1 Japanese, or the potential to cause confusion because of the existence of a similar sounding word or loan word. For example, words such as *trouble/travel (trouble* is used in the two word loan word combination *engine trouble*) or words with /l/ and /r/ (*fly/fry*) shall be avoided.

(g) The stimulus is not a function word. For example, prepositions were eliminated because there

was a risk they might generate other similar function words as responses. This reduced the set of candidate items to the following form classes: nouns, verbs, adjectives and adverbs

2.2 Cue selection method

The initial candidate base of cues was selected from the 0-1K range of the British National Corpus (BNC corpus, 2007). Note that this allowed for a much larger pool of cue words (1,000) than the list of 100 in the Postman & Keppel norms to which I had been restricted in the previous study. After each one of the 1,000 words was screened, only 125 candidate cue words fit all the above criteria. These were screened for overlap. Overlap is defined as a phenomenon where one cue word shares, or is perceived as having the potential to share, an excessive number of responses stimulated by another word. Also, common responses should not include another cue word. This leads to the final intra-list selection criterion: none of the cue words elicits responses which are also listed as common responses to other cues according to the EAT (1973). A common response is defined as a response making up 6% or more of the total responses. For example, *body* stimulates the response *soul* on 10% of occasions, which means that the cue *heart*, producing *soul* on 7% of occasions, cannot be included in a set containing the cue *body*. This proved very difficult to realize in practice and the following exceptions were therefore made: *up* (6 cases), *out* (4), *of* (2), and *me (*2). Since three of these responses are prepositions (*up*, *out*, and *of*) and *me* is a pronoun, I felt that they did not represent a risk of semantic overlap that was found with other cues such as *heart* and *body*. The final 50 cue words were chosen at random from the remaining set of 125 and replaced continually until all overlaps, with the exception of the above, were filtered out.

Table 2
Final list of 50 cue words

AIR	CHOICE	GAS	MEAN	SCIENCE
BEAR	CHURCH	HAPPEN	MOVE	SET
BECOME	CLASS	HEART	NATURE	SHARE
BLOW	CROSS	HOSPITAL	PACK	SORRY
BREAK	CUT	KEEP	PART	SPELL
BOAT	DRAW	KILL	POINT	STAGE
CALL	DRESS	KIND	POLICE	SURPRISE
CASE	FAIR	LEAD	POWER	TIE
CATCH	FIT	LINE	READY	WORLD
CHANCE	FREE	MARRY	RULE	USE

Section 3: COMPILING NORMS LISTS FROM WORD ASSOCIATION DATA: DESIGNING AND IMPLEMENTING PROTOCOLS.

This section describes:

- The participants
- The word association task and the data collection method
- Treatment of responses
- The method of norms list construction

I conclude with some comments on the two norms lists

3.1 Participants

In this section, I describe how I selected the two groups of participants for compiling the two norms lists. The native speakers of English were chosen from among personal friends and colleagues. As for the group of highly proficient non-native (L1 Japanese) users of English, the greatest challenge was ensuring that members were highly proficient. For one, it was not possible to conduct supervised proficiency testing of suitable L2 subjects to justify their inclusion in the study because there were an insufficient number of suitable candidates living locally. Indeed, prospective participants lived in many different parts of Japan and some lived outside the country. An alternative to supervised proficiency testing was self-rating of English skills. I had started by asking for self-rating but it soon became evident that many respondents were excessively modest and under-rated their proficiency so this approach was abandoned. The method finally adopted was to construct a set of criteria based on use and experience of English to justify participant inclusion in the L2 group. The set of criteria was devised from analysis of how my English-speaking Japanese colleagues and acquaintances had acquired "native-like" skills in English. Some have lived, or are living, in English-speaking countries for extensive periods. Others are teaching, or have taught, English. Others had never lived abroad nor taught English but had acquired high degrees of fluency through: (i) using English for academic purposes, such as scientific researchers who publish papers in English, (ii) using English professionally in the international workplace, such as EFL publishers' representatives, or (iii) using English socially, with English-speaking spouses, for example. The resulting definition of a highly proficient Japanese user of English was a person who:

(i) has lived or has been living abroad (English-speaking country) for a year or more,
or (ii) is teaching English or has taught English,
or (iii) has extensive experience using English socially, in the international workplace, or for

academic purposes.

The final pool of 114 L2 subjects was drawn from among friends, colleagues from my workplace and professional organizations such as JALT (Japan Association of Language Teachers) and JACET (Japan Association of College English Teachers). The following is a group profile of participants in each group drawn from the personal information provided on the task sheet (see Table 3 below).

Table 3
Group profile of survey participants

	n=114 for each group	L2	L1
Age	Average age	43	47
Gender	Male	38	76
	Female	76	38
Country of residence	Resident in Japan	99	69
	Resident outside Japan	15	45
Highest level of education	University graduates	110	110
	High School graduates	4	4
Dominant occupation	Teacher	70	88
Number of L2 participants who are living or have lived in an English speaking for a year or more		85	
Mean number of years spent in English-speaking countries		4.8	
Number who teach or have taught English		88	
Number who often use English for academic purposes*		95	
Number who often use English with family or friends		56	
Number who often use English for business		68	

* Note this number automatically includes all those who are presently teaching English.
By nationality, the breakdown of the L1 group is: USA (34), Canada (33), Britain (32), Australia (13), Ireland (1), and New Zealand (1). The 15 L2 participants living outside Japan live in the following countries: Canada (6), USA (3), Britain (2), Indonesia (1), Brazil (1), Germany (1), and Samoa (1).

3.2 The word association task

Word association task forms were sent out and collected as e-mail attachments. In the task instructions, participants were asked to:

1) Provide 5 different responses to each cue word.
2) Avoid proper nouns.
3) Use English words only.
4) Avoid responses of more than a single word.
5) Be unconcerned about making spelling or typing errors.
6) Refrain from consulting dictionaries, online references tools, or friends

However, some participants in both groups provided incomplete response sets, with fewer than

5 responses in each set, proper nouns, and responses of more than a single word. The suggestion is that they either ignored the instructions or found it difficult to control their response behaviour. With respect to the latter explanation, the premise underlying a word association task is that the items produced are those immediately, or automatically, activated by the cue. In this sense, it is unreasonable to expect a "response filter" to be strictly applied in all cases, even with only 5 rules to bear in mind.

3.3 Treatment of responses

Three factors were taken into account when processing responses. First, norms lists and response treatment guidelines in three other published norms lists: the EAT (1973), the Postman & Keppel lists (1970), and the Birkbeck word association norms (1996) were studied since the same issues would likely have surfaced during compilation. Also, if useful comparisons are to be made between this new norms list and these other three, we should consider using comparable treatment protocols. Second, discussion of issues underlying treatment of problematic responses in the literature, especially Wolter (2002), was carefully considered. Third, I needed a processing tool to assist with the task sorting 228 complete sets of 5 responses for each of the 50 cue words. I chose Tex-Lex Compare tool (Cobb, 2007) because it sorted and counted the number of tokens and types for response sets to each cue word.

Table 4
Summary of types of response requiring special treatment

RESPONSE TYPE
1) Blank responses
2) Proper Noun
3) Non-English words
4) Multi-word units (MWUs) and hyphenated words
5) Spelling and typing errors
6) Non-alphabetical symbols
7) Incomplete words, morphemes, suffixes and affixes
8) Repetitions of response or cue word within a 5 item response set
9) Punctuation marks
10) Single letters
11) Acronyms
12) Transcription of noises, or non-standard onomatopoeic words
13) Spelling varieties (American and British English)

Clearly there is some tension between the desire to leave all responses in their original form,

without changing them in any way, and the necessity of applying, and strictly adhering to, rules in treating them. Since it was clear that the vast majority of participants had followed the instructions, it was felt that allowing proper nouns, for example, would compromise task conditions. In sum, both maintaining consistency of response treatment and attempting to reflect the intention of the participants as faithfully as possible were concerns of equal importance. The non-standard responses listed in Table 4 were treated in the following ways:

1) Blank responses

While 5 responses for each cue were requested, a very small number of respondents failed to complete the sets. In two cases, both from the L2 group, a large number of spaces (33% and 10% respectively) were left blank throughout so data from those participants was discarded. In a few other cases, 1-5 words were missing from each set of five responses and asterisks were inserted in these blank spaces adopting a procedure followed in the Birkbeck lists. 15 asterisks were inserted into one survey and this was the maximum allowed.

2) Proper nouns

Since participants had been asked to avoid proper nouns, when they were provided, they were also replaced with an asterisk. For the purposes of this research, I define a proper noun as a noun which requires capitalization of the initial letter and appears *only* in this form according to the Merriam-Webster online dictionary. *Jesus*, *Christ*, *Sunday*, *Christian*, *Christianity*, and *Christmas* were very common in response to *cross* and *church* and were removed, along with *Presbyterian*. However, *catholic*, *protestant*, *pope*, *baptist*, *tao* and *methodist* were accepted for inclusion in the list since the items are either described in the dictionary as "often [therefore not always] capitalized" or are listed even once without the initial capital letter. In this way, for the cue *church*, 34 responses were discounted from the L1 list, and 113 from the L2 list out of a total of 570. In fact, their inclusion as cues in this study was the clearly the result of a screening error wherein cues which seemed likely to elicit proper nouns - criterion (e) - were supposed to have been eliminated from the final set of 50 according to the selection criteria. On this basis *church*, the cue which prompted most of the above responses, should not have been included as a cue word. Some nouns such as *Bill* in response to *kill* were rejected because it was assumed that they referred to the movie title *Kill Bill* where Bill is a short form of the name William, even though *bill* exists as a common noun, or non-proper noun. In contrast, *china* is accepted as a response for *break* since it was probably intended as a common noun as in *bone china*. However, there are clearly potential pitfalls connected with this policy of assuming the reasons behind the associations made by the

participants in the survey and subjects in the WAT.

3) Responses that were not standard English words, or from another language.

Surprisingly, only one word that is clearly not English was provided by one respondent. This was the response: *fertig* (German) in response to *ready*. This response was replaced with an asterisk. The German word *weltanschauung* was provided by one scholar from the L2 group. Since it is in the Merriam-Webster dictionary, it was accepted.

4) MWUs (Multi-word units), hyphenated words, and contractions

Although participants in this survey were advised in the instructions to provide single word responses and to avoid entering responses of more than one word, multi-word units were very occasionally provided. Here follows a discussion of the issue of how MWUs are treated in this norms list and why. Typically, these were combinations of two words such as: *a horse* in response to *lead*, or three words, *of no return* in response to *point*, or even 4 or 5 words such as *as I'll ever be* in response to *ready*. In the Postman & Keppel lists (1970) and the EAT (1973), these MWUs are very rare but they are accepted unaltered. For example, *Foshay Tower* was provided in response to *high*, in the former list, and *pop group* in response to *move* in the latter. However, in the Birkbeck norms lists (1996) MWUs are hyphenated. One solution recommended by Wolter (2002) is to maintain the head word of the MWU and delete the remaining words. Unfortunately, it is not always clear what the most essential meaning-bearing or content item is, but this process of clipping MWUs was adopted as intuitively as possible. The issue is complicated further by participant use of brackets to indicate what the main single response was. While this seems helpful in some cases, the participants' chosen responses are not always the headword, nor do they necessarily match the most commonly chosen word. For example, in response to *use* one participant writes *full* (*useful*). It was decided to accept *useful* as the response. With MWUs that contained the prompt word, such as *break in* for *break*, the prompt word was removed from the response and *in* was entered. All hyphens in hyphenated items are replaced and joined with an underscore (_) because of the behaviour of the processor, which otherwise separates these items. Finally, the processor also separates words containing apostrophes such as *I'm* in response to *sorry*. These responses are accepted as single words (*I_m*) and apostrophes are also replaced with an underscore. *Can't* is also represented as *can_t, won't* is altered to *won_t*, and possessives are indicated in the same way as in *children_s*.

5) Spelling and typing errors.

Participants in the surveys were advised: "Please don't use a dictionary or any online tools or reference books (or friends) to help you. If you make a spelling or typing mistake, I will correct it for you". Unfortunately, with misspelled items it is not always immediately clear what the intended word was, although the associative context serves as an invaluable guide. In the Postman & Keppel Lists (1970) and the EAT (1973) lists there are a small number of misspelled items such as *wakefulleness* in response to *sleep* in the former. In this study, it was decided to correct misspelled items and alter them to a form that was believed to have been intended. This represents a divergence from the practice followed in the Birkbeck lists (1996:4) where possible spelling errors that produced a real word were not changed. Examples of alterations made include one case of *alter* in response to *church* that was altered to *altar*, and *coincident* altered to *coincidence* in a few cases in response to *chance*. The response *patience* for *hospital* was left as it was, after some thought.

6) Non-alphabetical symbols

Some responses were numbers such as the emergency number 999, and the Japanese equivalent, in response to *call*. These were removed because they do not appear in the dictionary (and could not be reduced to a single meaning-bearing headword), while twenty-two in response to *catch* was maintained as a single word that is listed in the dictionary. The digits *22* were provided by some participants and these were altered to *twenty_two* because the processor renders digits as "number". Combinations of numbers and letters such as *Co2* and *O2* in response to *air* were transcribed as the words *carbon_dioxide* and *oxygen*. Similarly, the response *$* to the cue *gas* was changed to *dollar*. One L2 participant entered *j*b* in response to *blow* and this was altered to *job*.

7) Incomplete words, morphemes, suffixes and affixes.

If a word is recognized as incomplete or only part of a word, such as a morpheme, suffix or affix that could be used to form a recognizable word in combination with the cue word, it is combined. However, this is a digression from the practice followed in the Postman & Keppel lists and the EAT. For example, *ish* in the former is listed as a response to *sickness*, and *ment* in the latter appears among the norms for *move*. Actually *–ment* and *ment* are listed separately because of the hyphen in the former. In this list, *ment*, in response to *move*, was changed to *movement* in one case and *–t* for *kill* was changed to *kilt* as intended, I assume. *Ful* for *power* was adjusted to *powerful*. Similarly, *heartedly* for *heart* is not a complete free-standing word and was changed to *whole_heartedly*.

8) Repetitions.

If the same response is entered more than once in the same set of 5 response words, the first is maintained but subsequent items are replaced with asterisks because participants had been asked to supply five different responses to each cue word. If a cue is entered as a response (in two cases, the response *surprise* was supplied for the cue *surprise*, as in the surprise party choral refrain "surprise, surprise!") it is left unaltered.

9) Punctuation marks.

Punctuation marks such as exclamation and question marks, are not altered but they are removed by the processor. There were no cases of responses being only exclamation or question marks, as in the EAT, although in this survey some responses like *boo* to the cue *surprise* are followed by an exclamation mark, while others are not.

10) Single letters,

A and I are the only single letter English words not rejected by the processor, and these are both accepted as words. *U-* in response to *boat* was changed to *u_boat*. However, all other single letter responses were rejected such as *c* for *fair*, *o* for *free*, and both *t* and *x* for *cross* since it was felt they did not constitute words.

11) Acronyms.

Acronyms such as TV and CD are accepted as long as they appear in the dictionary and do not represent proper nouns.

12) Transcription of noises, or non-standard onomatopoeic words

For the cue *bear*: *garaoo*, *gaaaaaooo*, and *groar* were supplied but, since they do not appear in the dictionary, they are treated in the same way as spelling mistakes and altered, to *growl* in this case.

13) Spelling varieties (American and British English)

Both *apologize* and *apologise* were supplied for *sorry* in roughly equal measure. However, in recognition of the fact that 67 of the 114 L1 respondents were North American (see Table 3), I decided to change all British spellings to American orthographic forms. Other items requiring alteration to American spelling included *theatre* (changed to *theater*), *colour* (*color*), *centre* (*center*), *tyre* (*tire*), and *favorite* (*favourite*).

A final issue concerning treatment of responses in the normative data is the incidence of reading miscues. There were a total of 7 cases of participants, all of them in the L2 group, apparently suffering reading miscues. These concerned the cues *chance* (mistaken for *change* in 3 cases), *fair* (mistaken for *fire* in two cases), and one case each for *blow* (*below*), and *bear* (*beer*). Initial reaction was that this represented a challenge to their assumed status as highly proficient non-native speakers of English, especially since it had not occurred even once with the NES group but had occasionally happened with learner groups in Munby (2007 and 2008). However, records show that in two of the above cases, the subjects had each spent about 10 years in the USA, one completing a PhD there. All responses resulting from apparent miscues were accepted.

3.4 The method of norms list construction

Following treatment, the responses on each completed survey form were then copied in vertical columns on two Excel spreadsheets- one for the Sapporo L1 English norms group and one for the Sapporo L2 English norms group. On each spreadsheet the response sets for each of the 50 cue words, or 114x5 = 570 tokens in response to each cue, were then copied and entered into the Tex-Lex Compare tool (Cobb, 2007)) for sorting and counting. If 570 tokens were registered in a set by the processor, the asterisks (representing blank or rejected responses) were removed because they were not norms. Following this, the lists were copied and pasted into a column on another spreadsheet- the norms lists- where the responses are listed complete with distribution information. The number of asterisks is listed at the foot of each column. The following is a description of how to read the norms lists. On the spreadsheet containing the completed norms lists, the cue word is listed repeatedly in all cells in the left column (e.g. column A for the first cue word *air*) adjacent to the responses. For the same cue *air*, the Sapporo L1 English normative responses are listed in column B, and the Sapporo L2 English normative responses are listed in column C to enable convenient comparison. This format is repeated for each of the 50 cue words, in three-column sets. The most common response for each cue is listed at the top in row 1. For example, the most common response to *air*, was *plane* in both groups. The numbers “49” and “50” mean that 49 L1 and 50 L2 participants provided this response.

3.5 The L1 and L2 norms: comments on the finished product

Regarding the two new norms lists (L1 and L2), there were two positive indications that they fulfilled their function effectively. For example, one problem I identified with the Postman & Keppel norms lists (1970) was that they were insufficiently extensive with an average of 100.22 different responses per cue for the Kruse cues (Munby, 2007) and 133.22 for the cue set in the

constructive replication (Munby, 2008). The two new norms lists each feature a larger mean number of different responses per cue word (L1 186.34 and L2 174.06).

Section 4: THE STUDY

To recap briefly on Section 5.1, the research questions for this study were:

RQ1 Is there any evidence in learner WAT50 performance that the new set of cue words functions in the expected way according to the criteria by which they were selected?

RQ2 Is there any evidence of a fatigue effect?

RQ3 Which norms list, the Sapporo L1 English norms or the Sapporo L2 English norms, yields the best match with learner responses?

RQ4 Which norms lists, the Sapporo L1 English norms or the Sapporo L2 English norms, yields the highest correlations with proficiency?

RQ5 Is 12 responses the optimal number of associations to elicit for each cue word?

In this section, I shall describe the methodology used to address RQ2 and RQ5. Before doing so, I provide details of the subjects, the test design and administration, and the treatment of responses and scoring. Note that this study classifies as a constructive (or conceptual) replication of Kruse since I "manipulate non-major variables, operationalizations or design features of the original study" (The review panel of the journal *Language Teaching* (2008, p3). As stated in the Munby (2008), these non-major variables are the cue words and the norms used to measure the responses. However, the basic methodology of Kruse remains the same; subjects enter up to 12 responses to a set of cues displayed on computer software. These responses are scored for the total number of responses provided and measured for stereotypy. Finally, these WAT scores are compared with proficiency countermeasures.

4.1 Subjects, test design, and administration

The participants comprised 82 English majors at Hokkai Gakuen University, Sapporo, Japan. Of these, 42 were female and 40 were male. 22 were first years, 51 were second years, 7 were third years, and 2 were in their fourth year. The majority of the subjects (61) were full-time students aged 18-22, while 21 were part-time students enrolled in the night program. Many of them were "mature" students of varying ages. A number of subjects in the day program had participated in both previous studies (Munby, 2007 and 2008). Regarding proficiency countermeasures, all subjects took the annual in-house TOEIC test. This is a two hour test of listening and reading using a multiple choice question format. In addition, one week later, all subjects took the same cloze test

that had been used in (Munby, 2007 and 2008). The WAT was taken three weeks later. The same software (IMO6a) that was used for Munby (2007) and 4 was used once again to display and collect responses for the 50 new cues. However, there was one addition to the task instructions. Subjects were advised not to chain their responses and reminded to respond to the cue on the screen. Two practice cue items (*banana* and *hope*) were also included, as in the previous two studies. It should also be noted that that the test was much longer than in the previous tests with 50 cue words instead of 10. Typically, subjects completed the test in 50-70 minutes, excluding time allowed for instructions and demonstration.

4.2 Treatment of responses and scoring

In this study, responses were treated in exactly the same way as responses supplied by the L1 and L2 groups in the WA survey task. With MWUs, if any one of the words in the MWUs appeared on the norms list, the response was maintained and treated as a scoring response. The only difference in treatment of responses from Munby (2007 and 2008) was that proper nouns were removed from response sets although they had previously been accepted for scoring in the number of response measure. There were three WAT scoring systems:

1) Number of responses. This is a straight count of the total number of responses entered for the 50 cue words.

2) (Non-weighted) stereotypy score. This is a straight count of the total number of responses that matched responses on the norms lists. Note that each set of learner responses was scored twice for stereotypy, once with the Sapporo L1 English norms and once again with the Sapporo L2 English norms lists. Every learner response that appeared even once on the norms lists was given one point for stereotypy. In other words, all matches with both idiosyncratic and non- idiosyncratic responses on the norms lists were also awarded one point each. While Kruse used both weighted and non-weighted measures, in this study, for the first time, no weighted stereotypy score was calculated. This was primarily because, overall, there was no compelling evidence in the two previous studies to suggest that weighted stereotypy revealed a closer link with proficiency measures than non-weighted stereotypy. Further, in a free word association task, there appears to be no possible way to justify a system of scoring wherein different responses score a different number of points depending both on whether they are entered as primary or secondary responses and on their frequency in the norms list. Hereafter in the thesis, we shall refer to "stereotypy scores" which, unless otherwise stated, should be taken to mean non-weighted.

4.3 Investigating fatigue effects and the optimal number of responses

With reference to RQ2 (Is there any evidence of a fatigue effect?), in order to test the effect of fatigue or boredom on performance, the total number of responses to the first half and the second half of WAT50 were compared. In order to address RQ5 (Is 12 responses the optimal number of associations to elicit for each cue word?) stereotypy scores (both L1 and L2) and their correlations were recalculated against both the cloze and the TOEIC measures after removing all the subjects' twelfth row of scoring responses from the Excel files, followed by the eleventh, and then all the way down to a calculation of WAT50 scores based only on the first response.

Section 5: RESULTS

In this section, I address the results of this study in the light of the five research questions.

RQ1 Is there any evidence in learner WAT50 performance that the new set of cue words functions in the expected way according to the criteria by which they were selected?

The first aim of this study was to select a new set of cue words for WAT50 according to criteria that were altered on only two counts from the study in Munby (2008). Here, I assess their effectiveness beginning with criteria (a), "The stimulus is known to even the lowest level subjects taking the tests". In Munby (2008), I interpreted failure to provide any responses to a cue as a potential indicator that the cue is unknown to the learner. In this study, it was encouraging that there were only 45 cases of zero response. In other words, at least one response was provided to each cue by all 82 subjects in all but 45 cases. Further, since these incidences never applied to more than 3 cases per cue, there was no evidence of a problem with any particular cue words, as there was with *mutton* and *priest* in Munby (2007).

Regarding criteria (e), ("The stimulus is not likely to elicit proper nouns"), out of an initial total of 22,720 responses produced by this group of learners in this study, 627 proper nouns were discounted. However, correlations between the number of response scores and the two proficiency measures were calculated twice, once before removal of proper nouns and once after. There was no difference between the two sets of correlations for this measure at the two decimal point level. The problem was that *church*, as with the L1 and L2 norming groups, also elicited a larger number of proper nouns, such as *Jesus* and *Mary*, from the learner group. Regarding part of criteria (f), "Words with /l/ and /r/ (e.g. *fly/fry*) shall be avoided", in a few cases, the cue *lead* was also found to elicit responses apparently related to *read*, such as *book*; this potential problem was not identified in screening and it should have been excluded. Despite the minor problems with criteria (e) and (f), overall the cues were functioning in the expected way according to the criteria

with which they were selected

RQ2 Is there any evidence of a fatigue effect?

The analysis detailed in Section 3 indicated that subjects produced 11,148 responses in the first half (mean 445.92, *SD* 73.82) and 10,945 in the second (mean 437.44, *SD* 79.12). A paired t-test to compare the means showed that there was no significant difference, suggesting that fatigue was probably not a factor affecting student performance. Note that this calculation was performed after proper nouns were removed.

RQ3 Which norms list, the Sapporo L1 English norms or the Sapporo L2 English norms, yields the best match with learner responses?

The data in Table 5 below indicates that there is a better match between subjects' responses and the Sapporo L2 English norms norms lists, with a mean stereotypy score of 184.29, than subjects' responses and the Sapporo L1 English norms (mean, 159.3). Results of a one-tailed paired *t*-test produced a *t* value of 9.45 ($p<0.0001$), meaning that this difference is statistically significant. This is not surprising since every single non-native subject scores a higher stereotypy score with the Sapporo L2 English norms list than with the Sapporo L1 English norms list.

Table 5

Mean scores, standard deviations, highest & lowest scores and maximum for all scoring methods of the WAT and proficiency measures (N=82)

	Mean	*SD*	High	Low	Maximum
No. of responses	269.4	107	578	89	600
L1 Stereotypy	159.3	51.4	300	60	600
L2 Stereotypy	184.3	58.8	377	71	600
TOEIC	539.2	137	935	300	990
Cloze	18.5	7.2	40	5	50

RQ4 Which norms list, the Sapporo L1 English norms or the Sapporo L2 English norms, yields the highest correlations with proficiency?

Although it is worth noting that the correlations with proficiency are broadly similar for each norms list score, correlations are higher for the Sapporo L1 English norms stereotypy measure (see Table 6 below). The TOEIC test, used for the first time as a proficiency measure, produces higher correlations with all three of the above WAT50 measures than the cloze test scores.

Table 6
Pearson correlations between WAT scores and proficiency measures

	CLOZE	TOEIC
No. of responses	.389**	.433**
L1 stereotypy	.562**	.601**
L2 stereotypy	.523**	.563**

1-sided p-value: Significant at $^{**}p<0.01$

A scatterplot comparison of the subjects WAT B stereotypy scores and TOEIC scores appears in Figure 1 below.

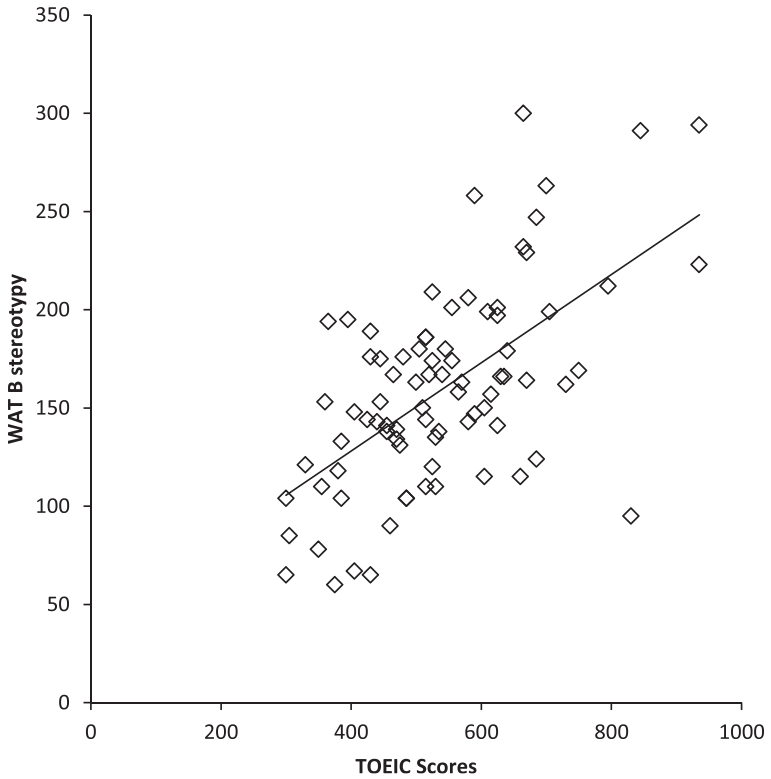

Figure 1 Scatter plot representation of correlations between the stereotypy scores (L1 norms list) and TOEIC scores (r=.601, p<.01)

RQ5 Is 12 responses the optimal number of associations to elicit for each cue word?
The results of the analysis described in Section 4.3 appear in Table 7 below.

Table 7
Recalculation of correlations between WAT stereotypy scores and proficiency measures with varying numbers of responses per cue word counted.

	CLOZE		TOEIC	
No. responses	L1	L2	L1	L2
1	.455**	.414**	.514**	.499**
2	.478**	.432**	.530**	.466**
3	.524**	.447**	.562**	.477**
4	.503**	.456**	.538**	.484**
5	.524**	.485**	.553**	.515**
6	.543**	.515**	.563**	.540**
7	.556**	.520**	.583**	.547**
8	**.566****	**.531****	.593**	.558**
9	.564**	.529**	.595**	.561**
10	.565**	.528**	.599**	.565**
11	.564**	.526**	**.603****	**.566****
12	.562**	.523**	.601**	.563**

1-sided *p*-value: All significant at $^{**}p<0.01$

The highest correlations are highlighted in bold. This analysis illustrates two important points about the most appropriate number of responses to be elicited in this WAT. First, with the cloze scores, correlations are highest with the first eight responses with both L1 and L2 norms. Second, the optimal eight response band discovered in the recalculation against the cloze measure does not hold true with the TOEIC. Here a recount of the first 11 responses appears to be the most effective.

Section 6: DISCUSSION

The main purpose of the study presented here was to design an improved multiple response WAT (WAT50) by seeking to rectify weaknesses and verify assumptions apparent in the original probe by Kruse et al. In doing so, the aim was to establish optimal conditions for this WAT to reflect level of proficiency with adult Japanese learners of English. The cornerstones of WAT50 were the new cue words and the new norms lists, and, in the light of RQs1, 2, and 5, they appear sound. With respect to RQ1, the cue words generally functioned in the expected way according to the criteria by which they were selected, as suggested in the previous section. The new norms lists also clearly represented an improvement on the Postman & Keppel lists (1970), which were in fact compiled in 1952. For example, in terms of quality, and the need for up to date responses, a larger number of learner-generated scoring responses (in both L1 and L2 norms) are related to contemporary consumer items or fashions which did not exist in 1952, such as *call*>*cell-phone, pack*>ziplock, tie>dye, *pack*>CD and *use*>computer. This indicates that the decision to abandon

the Postman & Keppel lists was the right one. Further, with respect to RQ2, although WAT50 typically took between 50 and 70 minutes to complete, there was no evidence of a fatigue effect. Finally, with respect to RQ5, results of the analysis in Table 7 do not warrant reducing the number of responses from 12.

The purpose of this section is twofold. First, I consider possible explanations for why the learner responses yield more matches with the L2 than the L1 norms list (RQ2). Second, I suggest one reason why correlations between learner stereotypy scores and proficiency measures are higher with L1 than with L2 norms (RQ3).

To begin with our first item, with reference to Table 5 in the results section, the finding that the learner responses yield more matches with the L2 than the L1 norms list is especially puzzling in view of the fact that the L1 lists feature a larger total number of different responses to 37 of the 50 cue words. While the prompt *ready* elicited exactly the same number of different responses from each norming group (150), the L2 group produced a larger number of different responses with only 12 of the cue words. One would suspect that there are at least two contributing factors. First, there are a number of responses on the L1 norms list that are not provided by either the learners or the L2 survey respondents. Animal-related responses to *pack*, such as *wolf* or *mule*, are examples of this class of exclusively native response. Second, many responses provided by the learners appear on the L2 lists but not on the L1 lists. For example, in response to *spell*, the form-based response *misspelling* appears on the L2 lists, but not on the L1 lists, and many learners scored points for this response. Further, although a large number of L1 participants (69 out of 114, see Table 3) are living in Japan and are familiar with the culture and language, they may not respond in a Japanese-like way. For example, the response *typhoon* to the cue *blow* was often provided by learners, and is listed twice in the L2 norms, but not at all in L1 norms. In this way, there may be some truth in Kruse's claim that the WAT is influenced by "problems such as …the effects of cultural background knowledge" (1987, p. 153).

Turning to the second item, namely correlations between stereotypy measures and proficiency (Table 6), I can think of only two reasons why the L1 norms list produces stereotypy scores that correlate more strongly with proficiency measures than the L2 norms list. The first is that the L1 norms list mirrors the kind of "native speaker like" English that is tested in the proficiency measures and that the L2 (Japanese) English is a different variety of the language, and influenced by L1. The second possible reason is that with increased proficiency, or increased exposure to

English, learner responses become more native-like.

CONCLUSION

In this study, I began by identifying some theoretical problems concerning the measurement of learner responses with the Postman & Keppel norms lists. I then described the selection of a new set of 50 cue words and compiled two new norms lists by correspondence: one from a group of native speakers of English (L1) and another from a group of highly proficient non-native (Japanese) users of English (L2). In this study, these 50 new cue words were used to elicit responses from a group of 82 learners with the software in WAT50. As in Munby (2007 and 2008) correlations between the WAT measures and the proficiency measures show that the two stereotypy measures (L1 and L2) are a better indicator of proficiency than the number of response measure, although all measures correlate positively and significantly. A further interesting trend found in this study was that the TOEIC test yielded stronger correlations with all WAT50 measures than the cloze test. Although the answers to the cloze test were never provided, since the same test cloze test was used each time and some subjects were repeating it, it is possible that its discriminatory power as a proficiency measure was being compromised. Further, the results of this study indicated that while there is a better match between the learners' responses and the L2 norms lists than the L1 normative data, correlation with proficiency measures is higher for the latter stereotypy measure. Finally, there is some evidence to suggest that the number of responses elicited for each cue word (12) should be maintained. In sum, as in Munby (2007 and 2008), there are clear signs in the study presented here that there is a link between a learner's performance on a multiple response free WAT and her level of L2 ability. However, there is no evidence that the development process of this WAT, leading to the creation of WAT50, has produced a test that is superior in quality to the original WAT designed by Kruse et al. This shall be the focus of the following study.

REFERENCES

The British National Corpus, version 3.2 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available: http://www.natcorp.ox.ac.uk/

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

Cobb, T. (2007). The Compleat Lexical Tutor (v.6 11/07) Available: http://www.lextutor.ca/text_lex_compare/

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

Higginbotham, G., Munby, I., & Racine, J. P. (2015). A Japanese word association database of English. *Vocabulary Learning and Instruction, 4*(2), 1–20.

Kiss, G. R., Armstrong, C., & Milroy, R. (1973). An Associative Thesaurus of English. EP Microfilms, Wakefield.

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

Language Teaching Review Panel (2008). Replication studies in language learning and teaching: Questions and answers. *The Language Teacher*. *41*(1), 1–14.

Meara, P. M. (1983). Word associations in a second language. *Nottingham Linguistics Circular*, *11*, 28–38.

Moss, H & Older, L. (1996). *Birkbeck word association norms*. Hove: Psychological Press. National Language Research Institute.

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. (2014). Sapporo word association norms lists. Retrieved October 11, 2014, from http://sapporo wordassociationnormslists.wordpress.com/

Postman, L. & Keppel G. (1970). *Norms of word association*. New York: Academic Press.

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

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.

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

Wolter, B. (2005). V_Links: A New Approach to Assessing Depth of Word Knowledge. Unpublished PhD thesis, the University of Wales, Swansea.