Establishing the reliability of word association data for investigating individual and group differences

FITZPATRICK, Tess, PLAYFOOT, David <http://orcid.org/0000-0003-0855-334X>, WRAY, Alison and WRIGHT, Margie

Applied Linguistics 2015: 36/1: 23–50
doi:10.1093/applin/amt020 Advance Access published on 24 September 2013

Establishing the Reliability of Word Association Data for Investigating Individual and Group Differences

[1,]*TESS FITZPATRICK, [2]DAVID PLAYFOOT, [1]ALISON WRAY and [3]MARGARET J. WRIGHT

[1]Centre for Language and Communication Research, Cardiff University, UK,
[2]Department of Psychology, Sociology and Politics, Sheffield Hallam University UK, and
[3]Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Australia
*E-mail: fitzpatrickt@cardiff.ac.uk

This article argues that, across different psychological contexts, the methods of data collection, treatment, and analysis in word association tests have hitherto been inconsistent. We demonstrate that this inconsistency has resulted from inadequate control, in previous studies, of certain important variables including the basis of norm comparisons, and we present a principled method for collecting, scoring, and analysing association responses, to address these issues. The method is evaluated using test and retest data sets from 16-year-old and over-65-year-old twins ($n = 636$), which enable us to (a) compare samples matched for key environmental variables, (b) assess the transferability of norming information between age cohorts, and (c) evaluate the reliability of the scoring protocols. We find systematic differences in the association behaviour of the two age cohorts, indicating the importance of evaluating data only against norms lists that are matched to the target population. Individual association behaviour is found to be consistent across test times, both in terms of response stereotypy and response type.

INTRODUCTION

For over a century, word association (WA) tasks have been used to investigate the content and organization of words and concepts in the mind. In early studies, the focus was conceptual, with responses interpreted as indicators of general behaviours (e.g. Galton 1879; Jung 1910) and, by extension, being used to diagnose psychological abnormality (e.g. Sommer 1901; Kent and Rosanoff 1910). More recently, WA studies have adopted a lexical focus, and have investigated the development and organization of the mental lexicon and the influence of specific variables on lexical access. In applied linguistics, interest has most often been on the integration of L2 items into the lexicon, and the ways in which WA responses might reflect the development of L2 proficiency (e.g. Kruse *et al*. 1987; Wolter 2002; Henriksen 2008; and, for an overview, Meara 2009). However, the findings of these L2 studies have been inconsistent and inconclusive, and in this article we propose that this is on account of an assumption about the nature of WA patterns that increasingly appears to be unsafe. It is an assumption that also pervades the L1 WA research context.

Most studies of WA in the L2 have evaluated learners' responses against 'native speaker norms'. The rationale is one of demonstrating that as proficiency increases, WA behaviour becomes more like that of an adult native speaker. However, recent investigations (e.g. Fitzpatrick 2007; Zareva and Wolter 2012) have questioned the validity of assuming there is a coherent norm behaviour in native speakers, with Fitzpatrick finding that 'not only do [native speakers] vary in the actual words they produce, they also seem to vary in the types of association they make' (2007: 327). On the other hand, consistency was found in the WA behaviour of individuals, both diachronically in the L1 and also synchronically across two languages (Fitzpatrick 2007; 2009).

A review of studies from outside mainstream applied linguistics, specifically from psychology, reveals that the idea of a 'normal' WA behaviour also anchors research and practice there. WA methods have been used (with informants operating in their L1) to investigate the effects on association behaviour of age, personality, psychosis, and cognitive function. While this indicates a recognition that there are individual differences in the L1 population, the focus has not been on capturing a range of normal behaviours so much as on interpreting the behaviour of an individual in relation to assumed normal responses. Specifically, norms lists are used here, just as they are in L2 research, as the core point of reference. We propose that it is perhaps for this reason that these L1 studies also present equivocal findings.

The methodology we present in this article was developed to maximize the opportunity to capture the nature of variation within L1 populations, and thus reveal the extent and nature of 'normal' WA behaviour as a reference point for research in both the L1 and L2 domains. The methodology was informed by theories of the mental lexicon and by previous WA research, and drew on a large sample of respondents ($n = 636$). We evaluated the approach by exploiting several distinct features of our data set. First, the informants were pairs of twins, making it possible to build two matched subsets of data. Secondly, a subgroup of informants completed the WA task at two separate test times, enabling us to assess reliability of response behaviour. Thirdly, the informants fell into two distinct age categories: 16-year-olds and >65-year-olds. This enabled us to examine the capacity of the methodology to capture differences between subpopulations that might inform future assumptions about reference norms.

In addition, we had data for the informants regarding their zygosity (i.e. whether they were identical or non-identical twins) and their performance on a range of cognitive tests. However, these elements are not discussed in this article because they are not relevant to the methodology itself.

In sum, our aim is to resolve the problem highlighted by Schmitt: 'It is clear that association data provides insights in the organization of the mental lexicon . . . . . . .and it seems that this approach is still waiting for a breakthrough in methodology which can unlock its undoubted potential' (2010: 248). In the remainder of this section, we review the extent of variation in the management and analysis of WA data in a number of influential studies. The next two sections describe our data set and analytic procedures. After this, we present and evaluate our method for measuring WA responses by stereotypy, and we demonstrate evidence that norms lists must be selected appropriately for the test population. Finally, we address the inherent complexities of categorizing responses by type. Both the norms and categorization measures are tested for reliability, using matched samples and longitudinal retests.

A review of approaches to the management and analysis of WA data

WA protocols are attractive to the researcher for a number of reasons. They offer a relatively quick and straightforward method for gathering rich language data. The data they elicit are freely produced, but consist of discrete lexical items, or word pairs (cue→response), which lend themselves to quantitative analysis more readily than do discursive language data. They are also congruent with well-established psycholinguistic and applied linguistic theories, such as Connectionism and Latent Semantic Analysis (see Ellis 1998), the Bilingual Interaction Activation model (e.g. Dijkstra and van Heuven 1998), and other models of word knowledge and lexical storage and retrieval (e.g. Marslen-Wilson 1987; Nation 2001). Tracking changes in WA responses can inform the study of a dynamic growing lexicon, in which links are being created and strengthened, and this is reflected in the amount of WA literature published since the 1950s relating to the development of L1 (Ervin 1961; Entwisle 1966; Nelson 1977) and L2 (Meara 2009). Furthermore, since the 1980s attention has been increasingly paid to the application of WA protocols to the study of lexical attrition (Gewirth *et al.* 1984; Gollan *et al.* 2006).

Typically, these studies have used one of two broad analytical approaches to the measurement of data. One entails examining the stereotypy of responses, that is, how similar an individual's response is to those in a reference set. The other approach examines the nature of the relationship between the cue and the response. Some studies combine the two approaches. The choice of analytic approach depends on the research question being addressed and the theoretical assumptions underlying the research. For instance, stereotypy approaches, which rely heavily on the similarity between a respondent's responses and 'normal responses', have been used in the context of cognitive and psychiatric disorders. Approaches categorizing the type of link between cue and response tend to be used to map patterns of variation in normal populations.

Research findings are of course dependent on the research questions and choice of analytic approach. However, a number of other factors also potentially impact heavily on the interpretation of data, so that different data-gathering procedures and materials may compromise the meaningfulness of cross-study comparisons. In addition to sample size, which influences the robustness of any quantitative empirical study, potential methodological variables to consider include the following:

- Mode of elicitation: Cues may be read or heard, and responses spoken, written, or typed.
- Cue choice: The number of cues in the WA task contributes to validity in the same way as population sample size. Less easy to quantify, but possibly even more important, is the way in which cue items are selected. Possible contributors to uncontrolled variation are word frequency, word class, imageability and the age at which the word was acquired. In addition, adequate attention has to be paid to the tendency for certain words to consistently cue a particular response, such as a highly probable collocate (e.g. bread→butter).
- Norms lists: Studies using stereotypy measures depend on norms lists against which to score the responses of the target population. While some studies compile norms lists from the study participants themselves or create bespoke norms lists (e.g. Miller and Chapman 1983; Hirsh and Tree 2001), most use existing lists such as the Postman-Keppel lists (Postman and Keppel 1970) or the South Florida Association Norms (Nelson *et al.* 1998). This second approach may not always allow for the possibility that responses are influenced by cohort characteristics such as generational differences, geographical location, and so on.
- Treatment of responses: Researchers vary in their treatment of response items. Some correct spelling, some lemmatize responses, and problematic responses such as non-words, multi-word responses and blanks are dealt with in different ways.

Thus, although it would seem reasonable, when deciding on a specific methodology for a WA study, to replicate the protocols most commonly used in previous research so as to maximize opportunities for cross-study comparability, a brief review of studies that have used WA methods reveals little commonality of approach. The studies listed in Table 1 have been selected to represent the main variables investigated through WA data: age, cognitive function, personality, and psychosis. The studies with the highest number of citations have been selected for each variable, using the Publish or Perish database (Harzing 2007). As the table shows, there is considerable between-study variation in the selection of cues and norms lists, and in the treatment and analysis of responses, affording little methodological guidance to the researcher. This is exacerbated by the fact that many of these articles report strikingly little methodological detail. Most offer no justification for methodological or procedural decisions, and little or no reference to the way data have been collected, treated, and analysed relative to other comparable studies. There are exceptions to this of course, notably in the early studies of first language development (Ervin 1961; Entwisle *et al.* 1964). Even when studies addressing the same research question and using the same theoretical

Table 1: Subjects, cues, norms lists, response treatment, and measures used in the most cited WA studies investigating age, cognitive function, personality, and psychosis

Study and variable		Subjects	Cues	Norms list	Treatment of responses	Measures
Age	Entwisle *et al.* (1964) **The syntactic-paradigmatic shift in children's WAs**	500 × children aged 5–11	24 high-frequency words: 8 nouns; 8 adjectives; 8 verbs	n/a	Grammatical analysis; subjective judgement made of 'transitional probabilities'	(1) Syntactic/non-syntactic (by age and word class) (2) Homogeneous/heterogeneous (by form class) (3) Form class of response words
	Ervin (1961) **Changes with age in the verbal determinants of WA**	23 × kindergarten 10 × 1st grade 52 × 3rd grade 99 × 6th grade	46 cues in vocabulary range of youngest children, 39 of which elicit antonyms or co-ordinates	n/a	Principled classification according to grammatical class, sequential analysis	Paradigmatic (strict grammatical interpretation)/syntagmatic (strict grammatical and text-informed interpretation)/clang
	Hirsh and Tree (2001) **WA norms for two cohorts of British adults**	45 × young adults 45 × older adults	90 concrete nouns and items likely to elicit concrete nouns	Compiled from participant responses	Plurals lemmatized	(1) Dominant/unique/shared responses (2) Response variation (3) Propositional-relational/hierarchical/categorical
Cognitive function	Gewirth *et al.* (1984) **Altered patterns of WA in dementia and aphasia**	38 × demented 17 × aphasic 22 × normal	16 cues from Palermo and Jenkins (1964): 4 nouns; 4 verbs; 5 adjectives; 3 adverbs	Palermo and Jenkins (1964)	No information given	(1) Popular/unpopular (popular = top 3 on norms list) (2) Paradigmatic/syntagmatic/idiosyncratic/identity (identical or similar to cue)/null
	Gollan *et al.* (2006) **WA in early Alzheimer's disease**	18 × probable AD 18 × elderly normals	52 cues from Nelson *et al.* (1998): 26 eliciting strong and 26 eliciting weak associations	Nelson *et al.* (1998)	In multi-word responses, most strongly associated word is scored; responses lemmatized to strongest association	(1) 'Mean response strength' of individual (according to per cent of normative population giving same responses) (2) Semantic/form/both semantic and form/multi-word/unrelated/non-word
Personality	Gough (1976) **Studying creativity by means of WA tests**	45 × research scientists 66 × engineering students	100 Kent and Rosanoff (1910) cues	Russell and Jenkins (1970)	No information	Close/remote associations, defined as given by following percentages of norm group: >50 per cent; 25–50 per cent; 10–25 per cent; 1–10 per cent; <1 per cent

Study and variable		Subjects	Cues	Norms list	Treatment of responses	Measures
	Merten and Fischer (1999) **Creativity, personality, and WA responses**	40 × 'artistic' professionals 40 × schizophrenics 40 × normals	25 common nouns	Normative sample from Merten (1992)	No information	For each of three conditions: free association, common, and individual response conditions: (1) Number 'common' responses (primary response) (2) Number of 'individual' responses (not on norms list)
Psychosis	Merten (1993) **WA responses and psychoticism**	46 × psychiatric hospital staff and non-medical professionals	25 common nouns	Normative sample from Merten (1992)	No information	For each of three conditions: free association, common and individual response conditions: (1) Number 'common' responses (primary response) (2) Number of 'individual' responses (not on norms list)
	Miller and Chapman (1983) **Continued WA in hypothetically psychosis-prone college students**	60 × probable psychosis-prone 21 × controls	32 cues from Kent and Rosanoff (1910) lists with >10 consensual associations	Norms compiled from 120 male students	No information	(1) Popular (>25% in norms)/common (at least one occurrence in norms)/idiosyncratic (not in norms) (2) Of idiosyncratic responses: individual (non deviant)/unusual (deviant)

Table 2: Subjects, cues, norms lists, response treatment, and measures used in the most cited WA studies investigating L2 proficiency

Study	Subjects	Cues	Norms list	Treatment of responses	Measures
Fitzpatrick (2006) **Habits and rabbits: word associations and the L2 lexicon**	40 × learners of English (mixed L1) 40 × native speakers of English	60 cues selected from the Academic Word List (Coxhead 2000)	N/A	Post-task interviews to confirm motivation for response	Divided into three categories: meaning-, position-, form-based or erratic, and into 17 subcategories
Kruse *et al.* (1987) **A multiple WA probe in second language acquisition**	15 × Dutch learners of English 7 × native speakers of English	10 cues selected from Postman and Keppel (1970)	Postman and Keppel (1970)	No information	(1) Number of responses (2) Weighted stereotypy: according to where each response appeared on norms list (3) Non-weighted stereotypy: according to whether response appeared on norms list
Meara (1978) **Learners' WAs in French**	76 × female English learners of French	French translations of 100 Kent and Rosanoff (1910) cues	Rosenzweig (1970) (female list)	No information	(1) Primary response same as norms (2) Primary response which occurs in norms list (3) Primary response not in norms list
Namei (2004) **Bilingual lexical development: a Persian-Swedish WA study**	100 × Persian-Swedish bilinguals aged 6–22: 50 Swedish L1 aged 6–18; 50 Persian L1 aged 6–19	Persian and Swedish translations of 100 Kent and Rosanoff (1910) cues	N/A	1) Phonemically transcribed 2) Translated into English	Categorized as clang/syntagmatic/paradigmatic/misunderstanding
Söderman (1993) **Word associations of foreign language learners and native speakers**	112 × Finnish learners of English: 28 each from 7th grade; Gymnasium; 1st yr university; advanced learners Expt 2 only: 28 × native speakers of English	Expt 1: 100 Kent and Rosanoff (1910) cues Expt 2: 64 cues (mostly adjectives): 32 frequent; 32 infrequent	N/A	No information	Categorized as clang/syntagmatic/paradigmatic/other
Wolter (2002) **Assessing proficiency through WAs: is there still hope?**	30 × Japanese learners of English 42 × native speakers of English	20 verbs from Edinburgh Associative Thesaurus, excluding items eliciting dominant primary response or high number of idiosyncratic responses	Edinburgh Associative Thesaurus (Kiss *et al.* 1973)	1) Multi-word responses reduced to head word 2) Responses lemmatized	(1) Non-weighted scoring: according to whether response is on norms list (2) Weighted scoring: according to number of native speakers who had given the response

assumptions are compared, there is little consistency of approach, as seen in Table 2, which lists the most cited experimental studies using the production of WA responses to investigate L2 proficiency.

The methodology reported in the following sections of this article is able to shed light on the potential impact of some of the previously uncontrolled variables listed above. We held constant the variables of mode of elicitation and cue choice, to explore the impact of norms sets and categorization. Future research will be able to focus on the first two variables, using the findings from this study to anchor the latter two.

THE DATA SET

The opportunity to use WA data from twins arose in the context of our collaboration, since 2007, with a research team engaged in two large-scale twin studies: the Genes for Cognition Study and the Older Australian Twins Study (Wright and Martin 2004; Sachdev *et al*. 2009; see http://genepi.qimr.edu.au/ for further details).[1] WA tasks were included in a battery of cognitive performance tests with the ultimate aim of exploring the roles of genes and environment in the relationships between different measures of linguistic and non-linguistic performance. For the norms lists and stereotypy analyses, the data are from 192 participants: 48 twin pairs aged 16 years and 48 twin pairs aged >65 years. The categorization of association types used the responses of 540 of the 16-year-old twins. Responses from a subset of the younger participant group ($n = 36$), who performed the task twice, were used to assess the reliability of both the stereotypy and the categorization methods. All participants in all analyses were native English speakers. The older twins were recruited through the Australian Twin Registry or publicity, and the 16-year-olds through schools and word of mouth. The studies were subject to the strict ethics procedures of medical research. Participants completed the WA task as part of a suite of physical and cognitive tests during either a half (16-year-olds) or 1-day-long visit to the research unit, located in a hospital.

The WA task consisted of 100 cue words,[2] controlled for the impact of frequency by randomly selecting them from the 2 k and 3 k bands of the British National Corpus, http://www.natcorp.ox.ac.uk (thus representing the second and third thousand most frequent words in English usage). Words from the first thousand band were not included because previous research shows that frequently encountered words tend to produce strong dominant responses (Meara 1983) and a proliferation of predictable responses would mask potential differences between participants. On the other hand, restricting cue selection to the 2 k and 3 k bands (50 cues from each) ensured that cue items were familiar enough for the respondents to offer an association to them. The cues and their dominant responses are listed in Appendix. For a full set of responses (excluding idiosyncratic responses) see Supplementary Appendix. Although we did not explicitly control imageability or age of acquisition in the cues selected (see earlier note that these might affect responses), regression analyses indicated that these characteristics of the cue did not predict stereotypy or response category.

The cues were presented in two columns of 25, on two pages. Next to each cue was a space for the participant to write a response.[3] Participants were instructed to write down the first word they thought of when reading each cue, and were told that there were no right or wrong answers. An excerpt from a completed task is shown in Figure 1. Participants were allowed up to 10 min to complete the task, and all participants finished it within this time.

PREPARING THE DATA FOR ANALYSIS

The data were presented to the analysts with only identity codes that did not indicate gender or twin pairings. The handwritten responses were transcribed into an excel file. To enable automatic searches, spelling was corrected, but only where the intention was clear (e.g. *controll* and *controle* were corrected to *control*). However, instances of possible spelling mistakes were not corrected if the response was a real word. For example, one participant wrote *backed* for the cue word *bean*. Although it is extremely likely in this particular case that the intended response was *baked*, many other cases rendered much less clear relationships between what was actually written and what might have been intended (e.g. both *council* and *counsel* are plausible as associates for the cue *session*). So, to avoid a kind of second-guessing that would have imposed the analysts' own WA preferences, a blanket policy was adopted of treating real word responses at face value.

While the majority of responses (>95 per cent) were single words, participants occasionally wrote two or more words or a short phrase. Where phrases could be construed as formulaic sequences with a single coherent meaning (Wray 2002), they were transcribed as written. When multi-word responses did not represent strings in this way,[4] two procedures were used to shorten them. The first, appropriate where two separate one-word responses had been offered, was to truncate responses at punctuation (comma, slash, etc.). Thus, *bomb/explosion* was transcribed as *bomb*. The second entailed deleting function words, particularly conjunctions (*and, or, with*), pronouns (usually *I*), and infinitive *to*.

NORMS LISTS AND STEREOTYPY MEASURES

Use of norms lists

Stereotypy determines how similar a participant's responses are to those of a comparison group and thus entails the use of a normative response corpus. As can be seen in Tables 1 and 2, many previous studies have used published norms lists.

Selecting a norms list that has already been created, published, and used in other studies can be a useful shortcut in stereotypy analysis. However, a norms

Age: Date: Total time taken:

Please write down the first word you think of when you read each of the words listed below.
There are no right or wrong answers.

abuse	child	joint	bones
agenda	Plan	landlord	bossy
annoy	nuisance	loss	life
attack	injure	mathematics	confusing
bean	green	miner	dirty
blame	fault	nail	varnish
bread	soft	nurse	carer
candidate	perfect	owe	payback
cheese	crackers	permit	allow
cloud	rain	plug	drain
concentrate	focus	prevent	accident
cope	life	pudding	yummy
cupboard	drawers	reflect	think
delay	buses	repair	mend
diet	healthy	rock	concert
domestic	housework	sand	grainy
effort	try	session	class
establish	facts	sin	bad
extension	house	source	food
fence	white	store	items
fraction	layers	swear	oath
gold	jewellery	thick	cream
heaven	god	tour	guide
ideal	best	variety	pack
instance	moment	weak	link

1

Figure 1: Excerpt from data set

list will only be reliable as a point of reference if it is able to transcend the impact of variables characterizing subpopulations. Until more is known about how different variables affect WA behaviour, researchers should be cautious about using independently gathered norm data as the reference point. The best way to address this issue is to create a norms list specifically for the study at hand, reliably to reflect the maximum possible number of characteristics of the study population. In this way, it will be possible to develop an understanding of the differences in such norms across populations and the contribution that those differences make in the interpretation of data. Accordingly, as outlined below, in this study separate norms lists were compiled for the two populations under investigation—16-year-olds and >65-year-olds, and it was these lists that were used to calculate stereotypy scores (see below).[5]

Each norms list represented the associations of 96 participants in the respective age group. The lists were created by compiling a full list of the responses for each cue word, and counting up how many times each response was given. To do this, it was necessary to determine a definition of 'word'. For example, some scholars count every different word form as a different response (so that *walk* is different to *walked* or *walking* or *walker*), while others group such responses together as versions of the same lemma. The decision we took here was to lemmatize inflectional variants but not derivational ones. Specifically, words that corresponded to level 2 of Bauer and Nation's (1993) description of word families were considered the same. In practice, that meant affixes producing plural nouns or verb participles were ignored, so that *cat* was considered the same as *cats*, *think* the same as *thinking*, and *walk* the same as *walked*. Derivational affixes, though, were retained, so that *health* and *healthy* were considered different responses as were *teach* and *teacher*. The justification for this decision was that while any kind of lemmatizing potentially impacts on gaining a full understanding of collocational behaviour (compare *attack* and *attacked* as responses to *heart*), the impact of not lemmatizing is arguably greater because it considerably reduces the incidence of common responses across the population. The key consideration is consistency and transparency, so that the way is clear for future empirical interrogations of the potential impact of the decisions taken.

The norms lists were finalized by ordering the responses according to their frequency for that cue word, along with a record of those frequencies.

Scoring for stereotypy

Previous studies have scored stereotypy in different ways (see Tables 1 and 2, last column), variously awarding 'stereotypy' points

(a) for any response in the top 3 (or 5) in the norms list
(b) for each percentage point of the norming population giving the response
(c) according to percentage bands of the norming population giving the response
(d) according to the ranking of the response on the norms list
(e) for any response that appears anywhere in the norms list
(f) for a response that is the dominant response on the norms list.

In this article, we focus on a method using procedure (f), as this represents the measure most commonly used in the studies cited in Tables 1 and 2.[6] It should be noted, though, that the decision about which stereotypy measure to use will be dependent on the context of that particular study. In L2 research, for example, where participants typically have limited lexical resources, method (e) above might be more appropriate. Using scoring method (f), a response was considered 'stereotypical' if it was the most frequently recorded response on the norms list for the participant's age cohort. Participants scored 1 point for every stereotypical response, and all their other responses scored zero. For cues where two (or more) responses were equally popular, a point could be scored for either response.

The data used in this analysis were from participants who had provided responses to >90 per cent of the cues. In studies like this one, which use relatively frequent cue words from the participants' L1, and where participants are adults with no cognitive impairment, blank responses are rare. However, in other contexts, a proliferation of blank responses might affect the analysis of some data sets, and appropriate methodological adjustments (typically the exclusion of data sets with more than *n* blank responses, or scores calculated on proportional rather than raw counts) have to be implemented.

Assessing the validity of the norms list approach

To assess the effect of norms list characteristics on the profiling of the data, age was used as a variable. The 192 participants were split on the basis of age and twin birth order (1 or 2) to create four groups (young twin 1, young twin 2, older twin 1, older twin 2).[7] A separate norms list was created for each group after the procedures described above, with each norms list therefore representing the responses of 48 participants. The prediction here was that differences between groups matched for age would be smaller than those not so matched.

Using the four separate norms lists as the reference, four stereotypy scores were calculated for each participant, according to the procedure described above. The first score was calculated from the norms lists to which the participant had contributed (i.e. a young twin 1 was given a point for every response that was a dominant response on the norms list compiled from all young twin 1 participants). The second stereotypy score was calculated from the norms list of responses from the group of the same age, different twin number (i.e. young twin 1 was given a point for every response that was a dominant one on the young twin 2 norms list). The third and fourth stereotypy scores were calculated from the norms list of twin 1 in the other age group, and twin 2 in the other age group. The four stereotypy scores therefore represent the similarity to 'own list', 'same age, other twin', 'twin 1, other age group', and 'twin 2, other age group' norms. Group mean stereotypy scores and standard deviations are presented in Table 3.

Three patterns are apparent. First, twin 1s and twin 2s have similar mean scores irrespective of the norms list. This is consistent with the assumption that there would be no material differences between first- and second-born twins in the context of stereotypy score. Secondly, the levels of stereotypy

Table 3: Mean scores (and standard deviations) for four measures of WA stereotypy

Participant group ($n = 48$ per group)	Comparison norms list			
	Own list	Same age other twin	Other age twin 1	Other age twin 2
Young twin 1	28.31 (6.68)	25.21 (6.39)	18.71 (7.18)	19.81 (7.01)
Young twin 2	27.31 (6.12)	27.33 (6.98)	18.54 (5.60)	19.38 (6.02)
Older twin 1	27.10 (10.36)	25.00 (9.69)	19.02 (7.35)	17.85 (7.32)
Older twin 2	26.00 (8.05)	24.71 (7.78)	20.65 (6.35)	18.81 (6.08)
Overall mean	27.18	25.56	19.23	18.96

for any given condition of comparison (i.e. the figures in each column) are similar, which indicates that the four groups' responses are related to each other in a consistent way. Thirdly, all participants' responses are more typical of their own age group than of the other age group, as shown by the lower mean stereotypy scores when using the norms derived from the other age twin lists.

To test the significance of the observations derived from these descriptive statistics, stereotypy data were entered into age (2) by twin (2) by norms list (4) repeated measures analysis of variance analyses by subjects and by items. Age and twin were entered as between subject variables in the analysis by subjects, and as within subject variables in the analysis by items. 'Norms list' was treated as a within subjects variable in both analyses. Greenhouse-Geisser corrections were applied to all analyses including the norms list factor, as it violated the assumption of sphericity. The analysis was conducted to establish whether (i) the choice of norms list for comparison had a significant effect on the stereotypy scores of the participants and (ii) whether there were overall differences in stereotypy levels between age groups or twin pairs once norms list factors were taken into account. The main effect of norms list was significant by subjects and by items [F_1 (3, 564) = 136.948, MSe = 25.730, $p < .001$, $\eta^2 = .421$; F2 (3, 297) = 69.319, MSe = 22.181, $p < .001$, $\eta^2 = .412$]. Bonferroni-corrected follow-up *t*-tests ($\alpha/6 = .0083$) revealed that mean 'own list' and 'same age group' stereotypy scores (27.18 and 25.56) were both significantly higher than those calculated from the other age group norms lists (19.23 and 18.96). The mean 'own list' stereotypy score (27.18) was significantly higher than stereotypy on the other norms list from the same age group (25.56), as is predictable given that participants' responses by definition all appear on their own norms list, and thus potentially contributed to the dominance of that response. The small difference in mean stereotypy in relation to other age twin 1 and other age twin 2 lists was not significant. The main effects of age and twin number did not reach significance, and no interactions were significant.

This analysis demonstrates the importance of using age-appropriate norms lists in the study of WA stereotypy. Participants gained an advantage of more than six stereotypy points (average 25.56 versus average 19.1) when scored against age-appropriate lists. There are several possible reasons for an age-related difference in the norms lists. One is that certain changes in WA selection strategies occur as a function of ageing. A second is that each generation has its own preferred set of vocabulary and/or associations. The first explanation predicts that the 16-year-olds' responses would, over time, come to resemble more closely the norms of the 65+ years age group. This means that the appropriacy of norms for new experimental groups could be calculated as a gradation on the basis of age. The second explanation predicts that the 16-year-olds would, in 50 years time, display norms rather similar to those they produced in teenage, but that a new cohort of

Table 4: Test–retest—stereotypy scores with correlation coefficient

$n = 36$	Test 1			Test 2			Correlation
	Min	Max	Mean	Min	Max	Mean	
Stereotypy	4	42	23.86 (8.371)	8	39	23.78 (7.388)	.855*

* $p < .01$.

16-year-olds at that time would produce new norms. A third possibility is that age and generation interact, such that as one gets older one attends to different concepts and words in the environment, as a function of one's changing interests and common activities, themselves influenced by prevailing generational cultural preferences. This more complex explanation, if correct, would predict that neither of the norms lists developed in this study would be a good match for the 16-year-olds when they got to 65+ years. Common to all three explanations is the caution about using as a reference point any norms list that is not derived directly from the target population.

Assessing the reliability of the stereotypy measure

For a measure to be considered reliable, it should produce comparable results at two test events using the same participants, always assuming participant performance is a stable factor. Key reasons why participant performance might not be replicable are practice effects including memory for the previous iteration (if the test events are close in time) and developmental or attritional changes in the participant's underlying organization of response options (if the test events are temporally very distant). The interval between test events here was ~3 months, which was considered large enough to minimize practice effects without reflecting substantial inherent changes in lexical knowledge or organization.

Thirty-six of the younger participants provided the data for this analysis, having completed the WA task on two separate occasions. Following the finding reported above, age-appropriate norms lists were used to score participants' responses for stereotypy. Table 4 presents descriptive statistics for stereotypy test and retest scores.

Mean scores were broadly similar across test times, with a significant positive test–retest correlation indicating consistency in WA behaviour over time. A calculation of repeated responses revealed that this consistency in scoring is not explained by participants producing the same responses to the same cues at each test time: on average identical responses were only produced for 25.5 of the 100 cues (Table 5).

Table 5: Response items repeated at test time two (maximum 100)

$n = 36$	**Min**	**Max**	**Mean**	**Standard deviation**
Repeated items	8	54	25.53	9.667

WA RESPONSE TYPE MEASURES

WA behaviour has also conventionally been assessed in terms of the types of link between the cue and the response. In early studies of this nature, analyses of the links were based on the Saussurian definitions of syntagmatic and paradigmatic relationships. A distinction was made between pairs of words that co-occur in text (syntagmatic, e.g. *van-drive*) and pairs of words that can be substituted for one another without changing the grammaticality of the sentence (paradigmatic, e.g. *van-train*). A third category, known as 'clang', was later added to this framework to represent responses based on the form of the cue, typically phonological (e.g. *van-fan*). Of the studies summarized in Tables 1 and 2, some (e.g. Ervin 1961; Gewirth *et al.* 1984) use variations of this framework and terminology, and there has more recently been a partial shift towards a change in terms to increase transparency, for example, 'collocational', 'semantic', and 'phonological'. Developments in cognitive linguistics relating to the categorization of sense relations (e.g. Croft and Cruse 2004), insights from natural language processing research (e.g. latent semantic analysis, Landauer *et al.* 1998), and the development of large-scale lexical databases such as WordNet (Miller 1995) have some potential to challenge and inform WA categorization systems, especially in the case of semantic (paradigmatic) connections. However, the recurrence in WA data of syntactic (usage-based) and orthographic/phonological associations has endorsed the continued inclusion of categories that accommodate these, such as the syntagmatic and clang categories in the conventional classification system.

These broad categories have revealed some qualitative differences in the response behaviours of children and adults (Nelson 1977). However, category comparisons between responses of other participant groups have been less conclusive, with studies sometimes producing contradictory findings (see Meara 2009 for a summary of these in relation to L2 investigations). Fitzpatrick (2006), also focusing on L2 WA processes, proposes a categorization based on a word knowledge framework (Nation 2001), which specifies subtypes of association response within each main category. She argues that this fine-grained approach provides greater insight into how learners of English engage with words. Her studies of distributions across these subcategories reveal differences between WA behaviour of L1 and L2 users of English, and between L2 users of different proficiency levels, which had hitherto been masked by the broad category approach (Fitzpatrick 2006, 2009).

Categorization of responses

The system of categorization used in the present analysis was based on Fitzpatrick (2006), and informed by the findings of subsequent studies (Fitzpatrick 2007, 2009; Fitzpatrick and Izura 2011; Higginbotham 2010). Key features of the revised system are, first, a rationalization of the number of subcategories, so as to ensure definitions are clear and the number of responses for each type is large enough for formal analysis. Secondly, the framework allows for responses to be coded as a potential combination of multiple links. For example, *knife* is commonly followed by *fork* in general usage (a collocation), but they are also items from the same lexical set (cutlery). In previous WA categorization systems, the researcher would be forced to make a choice as to which of these reasons was more likely. Here, the response can be classified as being both *lexical set* and *cue-response collocation.* It is advantageous to be able to recognize this level of complexity in light of the finding that participants are particularly quick to respond when the cue and the response are linked in more than one aspect (Fitzpatrick and Izura 2011).

The new framework comprises 14 subcategory headings in total, and is summarized in Table 6, with examples drawn from data in the present study.

Scoring WA responses using categories

The rationale when devising a categorization framework is to sustain a balance between consistency and common sense, while adequately accommodating all the responses. This is not an easy task, nor an exact science, because the analyst's belief that a participant probably had a reason for giving a particular response is not always enough to create a warrantable assumption about the link. To avoid second-guessing, the balance of power must lie with consistency. In this study, two specific procedures were used to maximize such consistency. First, to ensure that the raters were not influenced by the respondent's previous behaviour patterns, or by the popularity of a particular response across the sample, the categorization was done by cue not by participant. Thus, the complete list of responses to each cue was compiled into a single list, and duplicate answers were deleted, so that each response was listed only once per cue word. The relationship between cue and response was thereby neutralized, meaning that when raters were assigning responses to categories, they were not tempted to think 'this person has given a lot of collocations already so this is probably one too', or 'only one person said this so it's likely to be an erratic response'.

The complete set of responses to all the cues was categorized by two raters separately, according to the definitions above. Once the categorization had been completed by both raters, the scoring of responses was compared, revealing that 76.9 per cent of response items had been assigned to the same category in the initial coding. A further 22.8 per cent of the classifications were agreed

Table 6: Subcategories used to classify WA responses

Subcategory	Definition	Example
Synonym	Cue and response are synonymous in some situations	*Delay → impede* *Fraction → portion* *Establish → build*
Lexical set	Cue and response share a hyponym, or one word in the pair is an example of the other; includes antonyms	*Bean → pea* *Bean → vegetable* *Permit → deny*
Other conceptual	Cue and response are related in meaning, but are not synonyms or in the same lexical set	*Fence → field* *Sin → prayer* *Nurse → illness*
Cue-response collocation	Cue is followed by the response in common usage; includes compound nouns	*Fence → post* *Rock → roll* *Swear → word*
Response-cue collocation	Cue is preceded by the response in common usage; includes compound nouns	*Fence → electric* *Candidate → Nominate* *Plug → spark*
Cue–response and response–cue collocation	Cue could precede or follow the response in a common phrase(s)	*Rock → hard*
Affix manipulation	Cue is the response with the addition, deletion or changing of an affix	*Irony → ironic* *Abuse → abusive* *Plug → unplug*
Similar in form only	Cue and response are similar in orthography and/or phonology but do not share meaning	*Fence → hence* *Weak → week*

Subcategory	Definition	Example
Two-step association	Cue and response appear linked only through another word	*Weak* → *Monday* (via *week*) *Owe* → *mine* (via *own*)
Erratic	The link between cue and response seems illogical. Includes repetition of the cue	*Wolf* → *and* *Heaven* → *heaven*
Lexical set *and* cue–response collocation		*Bread* → *cheese* *Gold* → *silver* *Heaven* → *hell*
Lexical set *and* response–cue collocation		*Cheese* → *bread* *Nurse* → *doctor*
Synonym *and* cue–response collocation		*Torch* → *light*
Synonym *and* response–cue collocation		*Shove* → *push*

after a short discussion and close reference to the definitions. The non-alignments in the initial categorization of these responses were usually attributable to one rater missing a possible sense of the cue word. For example, one rater had missed the fact that *routine* could mean 'dull, boring, and monotonous', while the other missed the meaning of *establish* as 'to prove'. This highlights the necessity for multiple raters, particularly given the demands on raters to pay close attention to such large amounts of data. Agreement about the categorization of a small number of responses (0.3 per cent) could not be reached even after discussion. In these cases, a third party was consulted, and the link identified by the third party was used to arbitrate between the two options. During the categorization process, two cue words were found to be problematic, in that participants commonly mistook them for a (near-) homophone. *Miner* was mistaken for *minor*, and responded to as such, and *instance* was responded to as *instant*. These cues and the responses they elicited were excluded from the categorization analysis.

Using a spreadsheet, the responses were allocated their category type, and the instances of each category were summed to create individual response profiles.

Assessing the reliability of the categorization system

Having categorized participants' responses according to the process described above, an assessment of the reliability of this method was undertaken. The aim was to establish whether, irrespective of specific items in responses, the distributional patterns of response types were replicable—these patterns are the basis on which observations might be made about differences in participant profiles. Data from the 36 test–retest participants were used. Responses were categorized according to the framework in Table 6, and profiles were produced for all participants at time 1 and time 2. The mean number of responses in each subcategory is presented in Table 7, along with test–retest correlation coefficients (categories represented by, on average, less than one response per participant are not listed). Of the six main subcategories, significant positive correlations were observed for all but the erratic response category. High scorers on a given category in the initial test were likely to be high scorers on the same category in the retest.

As observed in connection with the stereotypy analyses reported above, this consistency cannot be attributed to participants providing identical response items at each test time (Table 5); the consistency here is in the type of response given, not the item itself.

Assessing the validity of the category clusters: a principal components analysis

As mentioned previously, a common analytic approach to WA data is to cluster responses into semantic, collocational, and form-based groups, and indeed the

Table 7: Test–retest—mean category scores and correlation coefficients (categories represented by an average of <1 response per participant are not included)

$n = 36$	**Test 1**	**Test 2**	**Correlation**
Synonym	17.17 (8.062)	14.61 (6.478)	.721*
Lexical set	5.81 (2.877)	6.06 (3.189)	.521*
Other conceptual	51.42 (9.749)	52.28 (9.254)	.824*
Cue–response collocation	10.86 (6.095)	12.25 (5.406)	.724*
Response–cue collocation	6.47 (3.247)	6.97 (2.932)	.518*
Erratic	1.06 (1.548)	1.22 (1.606)	.259

* $p < .001$.

subcategories proposed by Fitzpatrick were originally presented as subdivisions of these three groups. While there are theoretical grounds for making these distinctions, whether responses actually cluster in this way is an empirical question, which can be explored by submitting WA profile data (i.e. category scores) to a principle components analysis.

Principal components analysis is a technique designed to organize large numbers of inter-correlated variables into clusters such that the information can be described using only a small number of 'components'. This has advantages in terms of statistical power, and avoids multi-collinearity problems when using regression analyses. For example, imagine you have a bowl containing 100 sweets and you ask a child to pick five. There are a large number of possible combinations of five sweets that the child could choose. When asked, the child tells you that he/she decided which sweets to take on the basis of their colour, picking only red ones. A second child chooses five sweets from the bowl, and also takes only red sweets, but this child tells you that his/her decision was based on flavour. As there is a strong correlation between the colour and flavour of sweets, the identical selections of these two children, in the context of a larger set of children choosing on other grounds, could not be explained reliably using either of these variables, as both are possible explanations for their choice. A principal components analysis identifies patterns like this in the data set, and suggests a single 'colour–flavour' factor instead. Another child chooses five sweets from the bowl, but his/her strategy is to take the sweets closest to the surface. His/her selection has nothing to do with the 'colour–flavour' factor, and the variance in sweet picking is instead explained by proximity.

This analysis takes the total variance in the WA behaviour and attempts to partition it into linear components. The procedure results in clusters of variables (in this case, WA categories), which explain a proportion of the variance not explained by anything else. If the three major conventional categories are valid, they should manifest as clusters. Our initial categorization matrix contained 14 possible classifications for a response. Response data from 540 participants (all aged 16 years), in the form of response profiles, were entered into a principal components analysis. The sample size was determined to be adequate using the Kaiser–Meyer–Olkin measure (KMO = .51). The data met the sphericity assumption as determined by a significant Bartlett's test statistic [χ^2 (78) = 1069.056, $p < .001$]. The principal components analysis extracted five factors (rotated using the varimax procedure with Kaiser normalization) to explain the data. The rotated component matrix is presented in Table 8. The component labels in the table represent our interpretation of the component clusters; the analysis merely identifies them as discrete components.

Table 8 lists components from left to right, in order of the proportion of variance in the data they account for, with the largest proportion being attributed to the first rows. The first component identified comprises synonym, lexical set, and other conceptual link categories. This can be described as a meaning-based (semantic) component, as a conceptual link between cue and response underlies each of these subcategories. A second component includes both cue–response and response–cue collocations. This can be described as a position-based (collocational) component, as the link is determined by the close occurrence of the two items in language use. The third component comprises form-only, two-step, affix manipulation, and erratic responses. It is suggested that this is a form-based component. In Fitzpatrick's original system,

Table 8: Rotated component matrix (factor loadings below 0.5 have been suppressed)

WA sub-category	Component				
	Meaning	Position	Form	Multi-position	Position plus meaning
Other conceptual	−.822				
Synonym	.717				
Lexical set	.709				
Cue–response		.816			
Response–cue		.672			
Two step			.641		
Erratic			.617		
Affix			.548		
Form only			.535		
Cue–response–response–cue				.788	
Lexical set plus response–cue					−.743
Synonym plus cue–response					.685

only two of these subcategories, *form only* and *affix*, constituted the broad category *form*. The components analysis suggests that two additional subcategories may belong in this group, and a closer analysis of these subcategories provides a principled explanation for this. First, in *two-step* associations, one step is nearly always form-based. This is illustrated by examples such as *bean* → *stork*. Here there has been an intermediate association involving the collocation *stalk*, a homophone (similar in form only) of the response *stork*. Secondly, the *erratic* response category encompasses potential spelling mistakes (i.e. form errors). The fact that these two categories load on the same component supports the notion that the *bean*→ *stalk/stork* response type might indeed be caused by erratic spelling (similarly, the *bean*→*backed* example cited earlier in this article). Component 4 includes only the cue–response–response–cue collocations; note that these did not load with the other position-based categories, though given that few of these responses were produced (<0.5 per cent), it is unwise to speculate about the reason for this.

The final component includes dual-link associations: synonym plus cue–response collocations and lexical set plus response–cue collocations. The separation of these associations from the main groups supports Fitzpatrick and Izura's (2011) finding that dual-link associations are particularly strong and quick to retrieve, and do not behave in the same way as either semantically or position-based responses. The last two components contribute an extremely small proportion of the total variance, and indeed items with these double links were uncommon in the data.

Specific research questions and hypotheses can demand a focus on particular subcategories (e.g. Fitzpatrick 2006 found that synonyms make a much larger contribution to the semantic category in L1 responses than in L2). However, it is often advantageous, for reasons of statistical analysis, to group data into larger categories, and this principal components analysis has identified a convincing framework for doing so.

CONCLUSIONS

We have demonstrated that norms lists differ between age cohorts, and we strengthened the evidence by using two uniquely matched participant groups, enabling within-group comparisons to constitute a point of reference. The implications of this for stereotypy-based measures of association behaviour are clear: norms lists must be selected, or compiled, to reflect the demographic profile of the target population. In this study, we have found an age, or generational, difference, and this has direct relevance, for example, to the way WA tasks have been used in SLA research to assess L2 proficiency: often the experiment group has a somewhat restricted age profile (they are typically university undergraduates), which differs considerably from that of the norming group (see Meara 1978 and Kruse *et al*. 1987 in Table 2). It is possible that other factors such as educational background or gender might also affect response norms.

Using the age-appropriate norms lists, we produced stereotypy scores for all participants, reflecting the number of primary dominant responses (i.e. those at the top of the norms lists) they produced. Large individual differences in stereotypy proved consistent, with a significant test–retest correlation of .855. In terms of response category analysis, a principal components analysis indicated a slightly different grouping of subcategories from that used in previous studies. Again, a test–retest analysis produced significant positive correlations in all main categories.

Taken together, the evidence presented in this study moves the field of WA research forward in a number of ways. First, the test–retest data, the establishment of norming criteria, and the confirmation of category clusters all contribute towards an argument for the construct validity and the reliability of this method of investigation. Secondly, it proposes a principled protocol for the analysis of WA data, facilitating comparison of data sets and making transparent the assumptions and procedures that underpin the methodology and analytic framework. As we have acknowledged throughout, specific research questions may motivate changes to the way association data is measured. For example, measures of idiosyncrasy will complement stereotypy scores, and particular subcategories of association type will be salient to the study of certain variables. The studies summarized in Tables 1 and 2 of this article are evidence that researchers in diverse fields, for well over half a century, have seen the potential of WA protocols to investigate lexical behaviour in conditions of development, decline, and impairment. By understanding the implications of methodological decisions, and by basing further studies on a consistent approach, it will be possible to maximize both the mutually informative nature of inter-study comparisons, and the degree to which findings can be interpreted in a meaningful way.

SUPPLEMENTARY DATA

Supplementary material is available at *Applied Linguistics* online.

ACKNOWLEDGEMENTS

We are grateful to Ann Eldridge, Natalie Garden, Marlene Grace, and Kerrie McAloney of Queensland Institute of Medical Research for collecting the data for this project, and to Cris Izura and Jeremy Tree of Swansea University and Naomi Wray of Queensland Brain Institute for their expert input to data analysis and interpretation.

FUNDING

The study reported here was part of a project supported by a grant from the Economic and Social Research Council, UK (RES-000-22-4012, October 2010–June 2012). Data collection was supported by grants from the Australian Research Council (DP1093900) and the National Health and Medical Research Council, Australia (401162).

APPENDIX

Cues and dominant primary responses from two participant groups: 16-year olds (n = 96) and >65-year olds (n = 96). Cues listed in order of task presentation

CUE	16s	>65s
abuse	hit	child
agenda	plan	meeting
annoy	irritate	pest
attack	hurt	hurt
bean	food	vegetable
blame	accuse	accuse/game
bread	food	butter
candidate	election	election
cheese	yellow	cheddar
cloud	sky	sky
concentrate	think	think
cope	stress	manage
cupboard	food	food
delay	wait	wait
diet	food	food
domestic	house	home
effort	try	try
establish	buildings	start
extension	long	house
fence	gate	post
fraction	math	part
gold	money	silver
heaven	god	hell/sky
ideal	perfect	perfect
joint	bones	knee
landlord	house	rent
loss	lose/sad	gain
mathematics	hard/numbers	sum
nail	hammer	hammer
nurse	doctor	doctor
owe	money	money
permit	allow	allow
plug	bath	sink
prevent	stop	stop
pudding	chocolate	plum
reflect	mirror	think
repair	fix	fix
rock	hard	hard
sand	beach	beach
session	time	time
sin	bad	bad
source	find/information	begin
store	shop	shop
swear	bad	word
thick	thin	thin
tour	guide	holiday
variety	different	different
weak	strong	strong

CUE	16s	>65s
abbey	church	church
alley	dark	lane
astonish	amaze/surprise	surprise
basket	ball	fruit
bond	james	money
bucket	water	water
canal	water	water
certificate	award	paper
click	mouse	shears
concert	music	music
corridor	hallway	hall
curious	wonder	cat
devote	love	love
dominate	power/strong	rule
echo	sound	sound
expose	show	show
fined	money	speed
foster	parents	care
gentle	soft	soft
greed	money	money
hay	horse	stack
hood	jumper	hat
indulge	chocolate	eat
irony	funny	sarcasm
ladder	climb	step
liquid	water	water
manual	car	book
miracle	god	birth/wonder
multiple	many	many
nuclear	bomb	bomb
overtake	car	pass
peak	mountain	top
poison	death	ivy
pride	lions	prejudice
rack	shelf	lamb
rescue	save	save
routine	daily	work
script	play	write
shove	push	push
snap	break	break
spite	hate	nasty
stiff	hard	hard
suicide	death	death
symbol	sign	sign
terrace	balcony/school	house
torch	light	light
tumble	fall	fall
vandal	graffiti	graffiti
wander	walk	roam
wolf	dog	dog

NOTES

1 Previous outputs from this collaboration include Mollet *et al*. 2010; Mollet *et al*. 2011.
2 Two of these cue words, and the responses they elicited, were subsequently excluded from analyses
3 The WA task was presented in written rather than spoken mode for three reasons. First, it was not feasible to collect both written and spoken responses from the same informants, unless in the same short timeslot of the same day, when fatigue and/or repetition effects would confound the results. The data were collected as part of a larger study, with little scope to manipulate the order of presentation or to extend the overall time taken for the WA element. Given this constraint, the main consideration was which mode to prefer. The written mode was preferable because, secondly, a team of research assistants was involved in data collection, and it would not be possible to guarantee consistency of delivery of spoken cues. And thirdly, the majority of WA studies in applied linguistics use written data, and using that same elicitation method maximized the relevance of our study to others. Clearly the mode of delivery is a significant variable, and future research needs to extend to a methodical comparison of the responses from participants under both conditions.
4 For a practical approach to justifying the identification of wordstrings as formulaic sequences, see Wray and Namba 2003, Wray 2008: chapter 9.
5 Subsequently, for the purposes of validity evaluation, the norming groups were further divided to enable both within- and between-age group analyses.
6 We also calculated 'weighted stereotypy' and 'idiosyncracy' scores for some other aspects of our study. In the former, respondents gained a score derived from the number of norms list contributors providing the same response; in the latter, respondents gained a score for every response they gave that no one else has produced.
7 The assignment to 'twin 1' or 'twin 2' was random: on the advice of the geneticists in the team, birth order was not considered a variable.

REFERENCES

Bauer, L. M. and **I. S. P. Nation.** 1993. 'Word families,' *International Journal of Lexicography* 6: 253–79.

Coxhead, A. 2000. 'A new Academic Word List,' *TESOL Quarterly* 34/2: 213–38.

Croft, W. and **D. A. Cruse.** 2004. *Cognitive Linguistics*. Cambridge University Press.

Dijkstra, T. and **W. J. B. van Heuven.** 1998. 'The BIA model and bilingual word recognition' in J. Grainger and A. M. Jacobs (eds): *Localist Connectionist Approaches to Human Cognition*. Lawrence Erlbaum, pp. 189–225.

Ellis, N. C. 1998. 'Emergentism, connectionism and language learning,' *Language Learning* 48/4: 631–64.

Entwisle, D. R. 1966. *The Word Associations of Young Children*. John Hopkins University Press.

Entwisle, D. R., D. F. Forsyth, and **R. Muuss.** 1964. 'The syntagmatic-paradigmatic shift in children's word associations,' *Journal of Verbal Learning and Verbal Behaviour* 3: 19–29.

Ervin, S. 1961. 'Changes with age in the verbal determinants of word association,' *American Journal of Psychology* 74: 361–72.

Fitzpatrick, T. 2006. 'Habits and rabbits: Word associations and the L2 lexicon,' *EUROSLA Yearbook 2006* 6: 121–45.

Fitzpatrick, T. 2007. 'Word association patterns: Unpacking the assumptions,' *International Journal of Applied Linguistics* 17: 319–31.

Fitzpatrick, T. 2009. 'Word association profiles in a first and second language: puzzles and problems' in T. Fitzpatrick and A. Barfield (eds): *Lexical Processing in Second Language Learners*. Multilingual Matters, pp. 38–52.

Fitzpatrick, T. and **C. Izura.** 2011. 'Word association in L1 and L2: An exploratory study of response types, response times and inter-language mediation,' *Studies in Second Language Acquisition* 33: 373–98.

Galton, F. 1879. 'Psychometric experiments,' *Brain* 2: 149–62.

Gewirth, L. R., A. G. Shindler, and **D. B. Hier.** 1984. 'Altered patterns of word associations in dementia and aphasia,' *Brain and Language* 21/2: 307–17.

Gollan, T. H., D. P. Salmon, and **J. L. Paxton.** 2006. 'Word association in early Alzheimer's disease,' *Brain and Language* 99: 289–303.

Gough, H. G 1976. 'Studying creativity by means of word association tests,' *Journal of Applied Psychology* 61/3: 348–53.

Harzing, A.W. 2007. Publish or Perish, available at http://www.harzing.com/pop.htm.

Henriksen, B. 2008. 'Declarative lexical knowledge' in D. Albrechtsen, K. Haastrup and B. Henriksen, *Vocabulary and Writing in a First and Second Language: Processes and Development*. Palgrave Macmillan, pp. 22–66.

Higginbotham, G. M. 2010. 'Individual learner profiles from word association tests: The effect of word frequency,' *System* 38/3: 379–90.

Hirsh, K. W. and **J. T. Tree.** 2001. 'Word association norms for two cohorts of British adults,' *Journal of Neurolinguistics* 14/1: 1–44.

Jung, C. G. 1910. 'The association method,' *The American Journal of Psychology* 21/2: 219–69.

Kent, G. H. and **A. J. Rosanoff.** 1910. 'A study of association in insanity,' *American Journal of Insanity* 67: 37–96, 317–90.

Kiss, G.R., C. Armstrong, and **R. Milroy.** 1973. *An Associative Thesaurus of English*. EP Microfilms.

Kruse, H., J. Pankhurst, and **M. Sharwood Smith.** 1987. 'A multiple word association probe in second language acquisition research,' *Studies in Second Language Acquisition* 9/2: 141–54.

Landauer, T. K., P. W. Foltz, and **D. Laham.** 1998. 'An introduction to latent semantic analysis,' *Discourse Processes* 25/2&3: 259–84.

Marslen-Wilson, W. D. 1987. 'Functional parallelism in spoken word-recognition,' *Cognition* 25/1–2: 71–102.

Meara, P. 1978. 'Learners' word associations in French,' *Interlanguage Studies Bulletin* 3: 192–211.

Meara, P. 1983. 'Word associations in a foreign language: a report on the Birkbeck vocabulary project,' *Nottingham Linguistic Circular* 11/2: 29–38.

Meara, P. 2009. *Connected Words*. John Benjamins.

Merten, T. 1992. 'Wortassoziation und Schizophrenie - eine empirische Studie,' *Nervenarzt* 63: 401–8.

Merten, T. 1993. 'Word association responses and psychoticism,' *Personality and Individual Differences* 14: 837–9.

Merten, T. and **I. Fischer.** 1999. 'Creativity, personality and word association responses: associative behaviour in forty supposedly creative persons,' *Personality and Individual Differences* 27: 933–42.

Miller, E. N. and **L. J. Chapman.** 1983. 'Continued word association in hypothetically psychosis-prone college students,' *Journal of Abnormal Psychology* 92/4: 468–78.

Miller, G. A. 1995. 'WordNet: a lexical database for english,' *Communications of the ACM* 38/11: 39–41.

Mollet, E., A. Wray, and **T. Fitzpatrick.** 2011. 'Accessing second-order collocation through lexical cooccurrence networks' in T. Herbst, S. Faulhaber, and P. Uhrig (eds): *The Phraseological View of Language*. Mouton de Gruyter, pp. 87–122.

Mollet, E., A. Wray, T. Fitzpatrick, N. R. Wray, and **M. J. Wright.** 2010. 'Choosing the best tools for comparative analyses of texts,' *International Journal of Corpus Linguistics* 15: 429–73.

Namei, S. 2004. 'Bilingual lexical development: a Persian-Swedish word association study,' *International Journal of Applied Linguistics* 14/3: 363–88.

Nation, I. S. P. 2001. *Learning Vocabulary in Another Language*. Cambridge University Press.

Nelson, D. L., C. L. McEvoy, and **T. A. Schreiber.** 1998. 'The University of South Florida word association, rhyme, and word fragment norms' http://www.usf.edu/FreeAssociation/.

Nelson, K. 1977. 'The syntagmatic-paradigmatic shift revisited: A review of research and theory,' *Psychological Bulletin* 84: 93–116.

Palermo, D. S. and **J. J. Jenkins.** 1964. *Word Association Norms: Grade School through College*. University of Minnesota Press.

Postman L. J. and **G. Keppel** (eds). 1970. *Norms of Word Association*. Academic Press.

Rosenzweig, M. R. 1970. 'International Kent-Rosanoff word association norms emphasizing those of French male and female students and French workmen' in L. J. Postman and G. Keppel (eds): *Norms of Word Association*. Academic Press, pp. 95–176.

Russell, W. A. and **J. J. Jenkins.** 1970. 'The complete Minnesota norms for responses to 100 words from the Kent-Rosanoff Word Association Test,' *Norms of Word Association*. Academic Press.

Sachdev, P. S., A. Lammel, J. N. Troller, T. Lee, M. J. Wright, D. Ames, W. Wen, N. J. Martin, H. Brodaty, and **P. R. Schofield.** 2009. 'A comprehensive neuropsychiatric study of elderly twins: the older Australian twins study,' *Twin Res Hum Genet* 12: 573–82.

Schmitt, N. 2010. *Researching Vocabulary*. Palgrave Macmillan.

Söderman, T. 1993. 'Word associations of foreign language learners and native speakers—different response types and their relevance to lexical development' in B. Hammarberg (ed.): *Problems, Process and Product in Language Learning*. Abo.

Sommer, R. 1901. *Diagnostik der Geisteskrankheiten*. Urban und Schwarzenberg.

Wolter, B. 2002. 'Assessing proficiency through word associations: is there still hope?,' *System* 30: 315–29

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press.

Wray, A. 2008. *Formulaic Language: Pushing the Boundaries*. Oxford University Press.

Wray, A. and **K. Namba.** 2003. 'Formulaic language in a Japanese-English bilingual child: a practical approach to data analysis,' *Japanese Journal for Multilingulism and Multiculturalism* 9: 24–51.

Wright, M. J. and **N. G. Martin.** 2004. 'The Brisbane Adolescent Twin Study: outline of study methods and research projects,' *Australian Journal of Psychology* 52: 65–78.

Zareva, A. and **B. Wolter.** 2012. 'The 'promise' of three methods of word association analysis to L2 lexical research,' *Second Language Research* 28/1: 41–67.