Lexical competence underlying second language word association tasks:  
Examining the construct validity of response type and response time measures

Masaki Eguchi¹ 
Shungo Suzuki 
Yuichi Suzuki

Abstract
This study investigated the constructs underlying second language (L2) word association (WA) with regard to three dimensions of lexical competence—size, organization, and accessibility (Meara, 2005)—and the lexical performance of speech. One-hundred and thirteen Japanese learners of English completed a computer-delivered oral WA task along with three vocabulary tasks: a form-recall gap-filling task (size), a primed lexical decision task (organization and accessibility), and an oral cartoon narrative (lexical richness). Regression analyses explored how well these lexical competence and performance scores predicted two WA outcome variables: response profiles and response times. Form-recall vocabulary knowledge, (collocational) priming, and lexical richness explained a large amount of variance in WA response type profiles (Nagelkerke’s pseudo $R^2 = .901$). Form-recall vocabulary knowledge and lexical decision time explained 28.5% of the variance of WA response times. A three-stage model of L2 WA task performance is proposed to account for the constructs underlying WA performance.

¹ Acknowledgements
We are grateful to Studies in Second Language Acquisition reviewers as well as the journal editors Luke Plonsky and Susan Gass for their constructive feedback on earlier versions of the manuscript, and to SSLA editorial assistant, Lizz Huntley, for the assistance during the review process. Our thanks go to Takumi Uchihara and Kaori Sugita for their help on data collection and coding. We would also like to thank Kazuya Saito, George Smith, and J-SLARF members for their insightful comments. This study was funded by the EIKEN Research Assistance Program awarded to the first and second authors from the EIKEN Foundation of Japan. Accepted for publication on Feb 1st, 2021.
**Introduction**

The development of valid tools to measure L2 lexical competence has been a central issue in second language (L2) vocabulary research (Daller et al., 2007; Henriksen, 1999; Schmitt, 2014). Lexical competence can be defined as one’s ability to manipulate lexical items (e.g., Meara, 2005), and researchers have proposed various models of lexical competence (Daller et al., 2007; Meara, 2005). For instance, Meara (2005) conceptualizes it from a three-dimensional perspective: (a) *size* (i.e., the number of lexical nodes represented in the mental lexicon), (b) *organization* (i.e., the quality of lexical networks across the entire lexicon), and (c) *accessibility* (i.e., efficiency in manipulating vocabulary knowledge) of lexical knowledge. A range of measures have been developed to gauge these three dimensions at the levels of linguistic knowledge (e.g., Productive Vocabulary Levels Test; Laufer & Nation, 1999) and language use (e.g., lexical richness; Bulté & Housen, 2012; Read, 2000). Although much research has been conducted to validate the measurement tools targeting specific knowledge components of lexical items (e.g., collocations, word parts, Nation, 2013; see also Schmitt, 2014), L2 vocabulary research can be extended by developing a more global, holistic measure to better understand the three dimensions of lexical competence (Meara, 2005).

Word association (WA) is a simple, yet good, elicitation technique that allows researchers to capture lexical storage and processing (e.g., Fitzpatrick & Izura, 2011; Jiang & Zhang, 2019; for reviews see Fitzpatrick, 2012; Fitzpatrick & Thwaites, 2020). The basic procedure for (discrete) WA tasks involves asking participants to produce the first lexical item that comes to mind after hearing or reading a cue word. Over the decades, researchers have envisaged that Word Association (WA) tasks would offer useful insights into the nature of L2 lexicon and its development (Meara, 1982). The L2 WA research strand has thus explored ways in which task can fulfil its promised potential. This includes the exploration of
lexical network densities via a Monte Carlo simulation (e.g., Wilks & Meara, 2002) and the assessment of L2 global proficiency (e.g., Wolter, 2002), among others. Meanwhile, L2 WA research has identified and addressed key methodological issues associated with the application of tasks in applied linguistics, including stimulus presentation/elicitation (Fitzpatrick & Izura, 2011) and response coding (Fitzpatrick, 2006) (for a review see Fitzpatrick et al., 2015). These methodological advances now provide the basis for extending our understanding of the constructs underlying WA task performance, which is one of the essential topics identified in the literature (see Fitzpatrick & Thwaites, 2020 for such a recent call).

In responding to the particular need to understand the construct underlying L2 WA tasks, the current study scrutinized a proposal that WA tasks/measures can be “used to investigate all three of these [size, quality, and accessibility] dimensions” of lexical knowledge (Fitzpatrick, 2012, p. 4). Specifically, the current study attempts to provide additional evidence for the construct validity of L2 tasks by drawing on a three-dimensional lexical competence model (Meara, 2005) and a measurement framework of spontaneous lexical use (i.e., lexical richness; Bulté & Housen, 2012; Read, 2000). To this end, we examined the extent to which L2 WA performance measures can be predicted by a battery of lexical measures derived from a form-recall gap-filling task (size), a primed lexical decision task (organization and accessibility), and an oral cartoon narrative task (lexical richness).

The findings of the study indicate how WA task performance can reflect multiple dimensions of lexical competence at the levels of linguistic knowledge and spontaneous lexical production. Building on the current findings, we attempt to fill the aforementioned gaps in our understanding of L2 WA tasks (see Fitzpatrick & Thwaites, 2020; Meara, 1982) by proposing a theoretical model describing a three-stage process underlying WA response generation.
Literature review

Two Approaches to L2 Word Association Tasks

In a comprehensive review of L2 WA research, Fitzpatrick and Thwaites (2020) identified two important methodological approaches in the L2 WA literature—stereotypy and response-type analyses. The current study selected the latter method for data coding; however, a review of the goals and assumptions of both will be necessary to situate our work in the field. Interested readers are referred to Fitzpatrick and Thwaites (2020) for a more in-depth historical overview, synthesis, and critical analyses of the two approaches.

One motivation for using WA tasks with two analytic approaches typically comes from the belief that such measures can capture important quantitative and qualitative differences among more or less developed L2 lexicon (e.g., Zareva, 2007). In stereotypy analysis, researchers typically awarded scores for *native-like* associations (e.g., Wolter, 2002), often based on a large-scale WA norm such as the University of South Florida Association Norm (Nelson et al., 2004). Fitzpatrick and Thwaites (2020) report that stereotypy measures have typically been used as an indicator of L2 vocabulary knowledge (or L2 proficiency). Research using stereotypy (e.g., Wolter, 2002) has tended to share the assumption that the more developed a learner’s lexicon is, the more likely s/he is to produce *native-like* word associations. The findings suggest that the native-likeness of WA responses is a moderate predictor of L2 proficiency (e.g., Wolter, 2002). However, the validity of the notion of *native-likeness* as well as the reliability of the scoring method are not without problems. For example, Fitzpatrick and Thwaites (2020; see also Fitzpatrick et al., 2015) note that when considering cultural or age-group differences, native-speakers’ WA responses are not always an appropriate target for L2 learners, or a particular norm list may not match a specific population of respondents.
In response-type analysis, although L2 proficiency is one of the important variables in the design, the primary interest has been to understand the types of inter-lexical connections that are salient to learners at different proficiency levels (e.g., Meara, 1982). Methodologically, instead of assessing WA responses according to *native-likeness*, researchers have categorized cue-response pairs according to lexico-semantic relationships—typically including paradigmatic (e.g., *dog–cat*), syntagmatic (e.g., *dog–run*), and phonological/clang (e.g., *dog–log*) associations (see Söderman, 1993). The assumption has been that these response types can capture differences in the kinds and strengths of semantic/lexical connections in the mental lexicon (e.g., Zareva, 2007). Thus, a good deal of research has attempted to detect any systematic changes (or even shifts) in (dominant) response types as a function of increasing L2 proficiency. Findings suggest that (a) developmental changes in response types are more gradual than what is considered “shift” (Namei, 2004), and (b) learners’ knowledge about cue words, rather than proficiency per se, seems to account for the proportions of different response types (Söderman, 1993; Wolter, 2001). Another line of research also indicates that native speakers, whose associations are assumed to be dominated by paradigmatic responses, are sometimes found to produce a good proportion of syntagmatic (or collocation) responses, sometimes exceeding the proportion of this syntagmatic association by non-native speakers (e.g., Fitzpatrick, 2006; Nissen & Henriksen, 2006). These findings suggest that WA task performance is influenced by a range of factors.

**Conceptual and Methodological Issues in L2 WA Research**

Given the complexities involved in WA response production, subsequent studies have identified and addressed several methodological issues that greatly impact on the internal consistency and comparability of WA data across studies (see Fitzpatrick et al., 2015 for a review). Three of them are reviewed below.
The first methodological concern is the presentation of cue words. While a majority of WA research has used paper-and-pencil elicitation, more recent studies have benefited from using psycholinguistic experimentation software, which enables tight control of stimulus presentation and response time limits (Playfoot et al., 2016) and the measurement of response times (RTs) (e.g., Fitzpatrick & Izura, 2011). The measurement of RTs has extended the scope of WA research, especially mechanisms for WA response generation (e.g., Playfoot et al., 2016).

Second, the complexities associated with response categorization involving paradigmatic, syntagmatic and clang associations (see Meara, 1982) encouraged researchers to devise a new set of WA response types which comprises Meaning-, Position-, and Form-based, and Erratic associations (see Fitzpatrick, 2006; Fitzpatrick & Izura, 2011). While the initial motivation of Fitzpatrick (2006) was to better capture how each association conforms to the vocabulary knowledge dimensions of form, meaning, and use (Nation, 2013), this categorization approach has also offered a window into the potential “activation route[s]” underlying each form of response generation (Fitzpatrick & Izura, 2011, p. 394). By investigating how quickly different response types were produced, Fitzpatrick and Izura (2011) showed that some response types, such as Meaning and Collocation (e.g., rubbish–bin, or peacock–feather), were produced significantly faster, indicating that response types may capture the qualitative differences of activation routes. They concluded that when more than one “activation trigger” was responsible, as in Meaning and Collocation, this can lead to faster RTs (Fitzpatrick & Izura, 2011, p. 392).

Third, as studies showed the impact of cue words on response production (e.g., Nissen & Henriksen, 2006), recent studies have paid more attention to cue-word selection. Researchers (e.g., Meara, 1982) have criticized the use of cue word lists that are typically used in psychological research, such as the Kent-Rosanoff list (1910), for their focus on
CONSTRUCT VALIDITY OF L2 WORD ASSOCIATION TASKS

general, high-frequency vocabulary. Responding to this call, more recent WA studies have explored a wide range of cue words, including the Academic Word List (Coxhead, 2000) to better represent mid-frequency, more abstract vocabulary (Fitzpatrick, 2006, 2007, 2009). Research is now available on how multiple lexical properties of cue words, such as word class, frequency, concreteness, and word neighbors, can affect WA behaviors (e.g., Jiang & Zhang, 2019; Nissen & Henriksen, 2006). This body of research suggests that in order to control for cue word influences, a principled selection of cue words should be a methodological prerequisite for WA research.

These methodological advances, which mainly took place during the 2000s (e.g., Fitzpatrick, 2006; Nissen & Henriksen, 2006), have afforded subsequent studies a methodological foundation to shed light on the process of bilingual lexical activation and retrieval using WA tasks (e.g., Fitzpatrick & Izura, 2011). This research thus suggests that the aforementioned methodological issues should be adequately addressed to ensure the reliability and validity of WA data (for a review, see Fitzpatrick et al., 2015). Following this line of research, the current study (a) employs psycholinguistic software to control each WA trial and measure RTs, (b) adapts a recent classification scheme that concerns meaning-, position-, form-based, and erratic responses (e.g., Fitzpatrick, 2006), and (c) proposes a stratified random sampling approach to selecting cue words based on multiple lexical properties.

Lexical Competence Underlying L2 Word Association Task Responses

Along with the research on bilingual lexical activation and processing (e.g., Fitzpatrick & Izura, 2011), another important research topic in the L2 WA literature is the extent to which WA tasks tap into L2 lexical competence (Meara, 2005; see also Schmitt, 2014 for the possibility of using WA tasks as a lexical network measure). While some
Researchers argue that more validity evidence is needed to fully establish WA tasks as a valid tool for assessing lexical competence (Cremer et al., 2010; Schmitt, 2014), promising evidence that supports the construct validity of WA tasks has been presented. For instance, at least several studies have indicated that WA performance systematically changes as a function of various aspects of L2 proficiency measured through standardized tests (e.g., TOEFL, Zareva & Wolter, 2012) and grade levels in school (e.g., Namei, 2004). In addition, WA performance has been found associative with more specific lexical measures such as vocabulary size (e.g., Fitzpatrick, 2006) and depth (e.g., Wolter, 2001).

However, very few attempts have been made to explicitly link WA performance to multiple dimensions of lexical competence—size, organization, and accessibility (Meara, 2005)—simultaneously. There is a reason why researchers may not have thoroughly investigated such potentially multiplex links. Existing research evidence suggests that WA performance, specifically response profiles, may not be strongly associated with constructs of vocabulary size (e.g., $r = .305$, Fitzpatrick, 2006), but can be affected by idiosyncratic preferences (Fitzpatrick, 2007, 2009). Fitzpatrick (2007) reported that a native speaker group demonstrated large standard deviations for the frequency of each response type, which was taken to suggest that WA behaviors of native speakers may be driven by idiosyncratic association preferences. This tendency was subsequently confirmed with an English-Welsh bilingual population, suggesting that the notion of individual preferred response patterns can be extended to L2 speakers (Fitzpatrick, 2009). Interestingly, Fitzpatrick (2009) also demonstrated that, as a function of self-rated L2 proficiency, individuals’ L2 WA response patterns resembled their L1 response patterns (although the relation was moderate; $rs = -.37$). An important implication of this is that developmental changes in L2 WA response profiles may be partially driven by individuals’ idiosyncratic L1 profiles, rather than by development in their L2 lexical competence.
The findings for individual response preferences, however, should not preclude researchers from investigating how three-dimensional lexical competence underlies L2 WA task performance. The association between the distance between L1–L2 association profiles and self-rated proficiency in Fitzpatrick (2009) was only moderate (approximately 14% of variance), leaving a large portion of the variance in L2 response patterns unaccounted for. To date, Henriksen (2008) is an important exception to the scarcity of validity evidence for the links between WA responses and multiple independent measures of lexical competence. In addressing the pseudo-longitudinal development of vocabulary knowledge by Danish-English bilingual students at three different educational levels (Grades 7, 10, and 13, respectively), Henriksen (2008) administered three lexical tasks—(a) a receptive Vocabulary Levels Test (VLT; Schmitt et al., 2001) representing size, (b) a receptive Word Connection Test representing organization, and (c) a productive WA task. Correlation analyses showed that the WA task was highly associated with VLT ($r = .546–.852$) at all grade levels, and with the Connection Task at Grade 7 ($r = .708$). Although this can be taken to indicate that the WA task can tap into multiple dimensions of lexical competence, one might argue that there are still methodological concerns for the following reasons: (a) issues pertaining to WA task design/administration such as response time limits, measurement of RTs, and cue word selection, and (b) the lack of measures capturing accessibility. Hence, methodological advancement is much needed to deepen our understanding of the construct validity of WA tasks from a multiple-dimension lexical perspective (for recent calls for research on the construct validity of WA tasks, see Fitzpatrick & Thwaites, 2020; see also Schmitt, 2014). The current study extends this line of research by following methodological recommendations and administering more comprehensive measures of lexical competence in order to probe how lexical competence relates to any part of WA performance.
Mapping WA Measures Onto Three-Dimensional Lexical Competence

Although there are various approaches to conceptualizing lexical competence, the current study draws on Meara’s (2005) framework of three-dimensional lexical competence—(a) size, (b) organization, and (c) accessibility. Meara’s framework is particularly suitable here, because it conceptualizes organization as a characteristic of the entire lexicon, and thus encourages us to focus on how lexical items interrelate in the learners’ lexicon (Meara & Wolter, 2004). Meara’s framework allows us to delineate how three dimensions of lexical competence are mapped onto WA task processes and their measures.

The first dimension, vocabulary size, concerns the number of lexical nodes that are represented in the learner’s mind (Meara, 2005; cf. the breadth dimension in Daller et al., 2007). Typically, vocabulary size is measured with tests that require some knowledge of form-meaning mapping (e.g., Productive Vocabulary Levels Test, henceforth PVLT; Laufer & Nation, 1999). Research suggests that productive knowledge is generally learned later than receptive knowledge (González-Fernández & Schmitt, 2019); thus, one would benefit from administering an assessment task tapping into productive knowledge, which would play a critical role in generating word association responses. Although size has been predominantly operationalized through receptive tasks (e.g., Wolter, 2001; Zareva, 2007), previous WA research has indicated that vocabulary size is associated with WA task performance at different levels. For instance, the larger the vocabulary size, the more likely is the speaker to know the cue words. This would result in (a) an overall increase in the number of responses produced (e.g., Zareva, 2007) and (b) changes in response-types—fewer Form-based responses and more Meaning- and/or Position-based responses (e.g., Fitzpatrick, 2006; Wolter, 2001).

Second, organization concerns the degree to which lexicon is “organised into a coherent lexical structure” (Meara, 2005, p. 274). Here, coherent lexicon can be
conceptualized in terms of the number and kinds of lexico-semantic connections among lexical items, including both semantic (e.g., Henriksen, 1999) and collocational (e.g., Wolter & Gyllstad, 2011). Accordingly, organization can be assessed through primed lexical decision tasks (e.g., Elgort, 2011; Wolter & Gyllstad, 2011), which tap into the extent to which prime words automatically activate the lexical and/or conceptual network (e.g., Lucas, 2000). Since previous studies have suggested that WA response type can reflect activation routes during response generation, such as Lexical Set and Collocation (Fitzpatrick & Izura, 2011), it can be hypothesized that the more structured a lexicon is in terms of semantic and/or collocational networks, the more likely the learners may draw on those types of connections in WA tasks. This may mean that, for example, when collocational networks are well established in one’s lexicon (as observed by greater collocational priming effects), an individual may produce a higher proportion of Position-based responses in WA tasks. However, considering the aforementioned complexity behind WA response profiles (see Fitzpatrick, 2006, 2007, 2009), it is an empirical question as to whether response profiles (particularly beyond Form-based responses) are systematically associated with organization.

Third, accessibility concerns the speed or fluency of lexical processing (Meara, 2005; cf. the fluency dimension in Daller et al., 2007). One way to operationalize accessibility would be to use Lexical Decision Time (LDT) in lexical decision tasks, reflecting the time it takes to recognize and access presented (non-)words in one’s mental lexicon. Theoretically speaking, accessibility should (at least partially) be associated with the speed of lexical activation and retrieval in WA tasks. Prior research has shown that increases in L2 proficiency allow faster L2 WA response production, thus narrowing the differences in within-speaker L1-L2 WA RTs (Fitzpatrick & Izura, 2011).

The aforementioned lexical measures (PVLT and Primed Lexical Decision tasks) are highly controlled in nature, with a set of target vocabulary items being pre-selected for
assessment (i.e., selective measures; Read, 2000). On the other hand, L2 vocabulary research has witnessed wider application of free written and spoken production tasks to assess lexical competence (i.e., lexical richness; Read, 2000; see also Bulté & Housen, 2012). Recent developments in corpus-based tools have enabled the automatic assessment of a wide range of lexical features in such contexts, including but not limited to frequency, concreteness, and multiword units (see Kyle et al., 2017). In terms of elicitation methods, WA tasks may be compatible with these free production measures, because both approaches require the spontaneous elicitation of lexical items (Fitzpatrick, 2012; Kormos, 2006; Playfoot et al., 2016). Indeed, it is proposed that the WA tasks may be situated at the middle point of the spectrum of lexical assessment tools from controlled to free production ones (see Dóczi & Kormos, 2016). The present study attempts to provide empirical evidence for a potential linkage among the three types of lexical measurement (i.e., controlled task, WA task, and free production tasks).
The current study

Motivated by the links hypothesized between WA performance and three dimensions of lexical competence and lexical richness, the current study aimed to examine these links in a comprehensive manner. To this end, WA performance was examined in terms of response type profiles (e.g., Fitzpatrick, 2006, 2009) and RT (e.g., Fitzpatrick & Izura, 2011). We hypothesized that these WA measurements would be associated with the three dimensions of lexical competence concerning size, organization, and accessibility (Meara, 2005) as well as lexical richness (e.g., Read, 2000). In this study, we measured each component of lexical competence and lexical richness through a form-recall gap-filling task (form-meaning mappings), a primed lexical decision task (lexical decision time and semantic and collocational priming), and an oral cartoon narrative task (spontaneous vocabulary use). The following research questions (RQs) guided our study:

1. To what extent are WA task response type profiles predicted by (a) knowledge of form-meaning mappings, (b) semantic and collocational priming, (c) lexical decision time, and (d) lexical richness?

2. To what extent are WA response times (RTs) predicted by (a) knowledge of form-meaning mappings, (b) semantic and collocational priming, (c) lexical decision time, and (d) lexical richness?

External lexical tasks (a)–(c) were chosen to reflect the model of lexical competence (Meara, 2005). We chose a mixture of receptive and productive types of measurement tasks to reflect the fact that both types of knowledge are necessary to complete a WA task—receptive knowledge for cue word recognition and activation of the lexical network, and productive knowledge for response generation. To tap into (a) form-meaning mappings (i.e., size), a variant of a form-recall vocabulary test, Productive Vocabulary Levels Test (PVLT; Laufer &
Nation, 1999), was chosen. The more correct answers that participants were able to provide on PVLT, the larger their vocabulary size. A controlled production task was deemed appropriate as research suggests that productive knowledge may be more difficult (González-Fernández & Schmitt, 2019), and the construct is considered to capture the degree to which learners are able to retrieve a lemma from the lexicon during WA tasks. Primed lexical decision was used to derive two types of measures: (b) semantic and collocational priming effects and (c) lexical decision times. Semantic and collocational priming effects were intended to measure organization, since they tap into the lexical network in terms of either semantics (e.g., Elgort, 2011) or collocations (e.g., Wolter & Gyllstad, 2011). Lexical decision time was intended to tap into accessibility through its response latency (i.e., how fast a learner can process cue words). Finally, an oral cartoon narrative was elicited to measure (d) diversity and sophistication aspects of lexical richness (see Bulté & Housen, 2012). Specifically, lexical diversity, the variety of lexical items produced in a text (see Read, 2000), essentially reflects the number of vocabulary items that are available to the learner during language production (e.g., Bulté & Housen, 2012). Lexical sophistication, on the other hand, can indicate qualitative aspects of lexical production in terms of, for example, frequency, concreteness, and association strengths of multiword units (e.g., Kyle et al., 2017). We decided to measure these using spontaneous cartoon narratives for the following reasons. First, as an oral WA task was employed to measure the speed of word association, we decided to adopt oral modality for eliciting lexical production as well. Second, in order to minimize the influence of speech content on lexical use, we employed a closed speaking task which predefined the content of speech.
Method

Participants

A total of 122 Japanese learners of English were recruited from a private university in Japan. The final sample consisted of 113 participants, after removing nine due to procedural problems with experimental materials ($n = 3$) or attrition ($n = 6$). The participants’ English proficiency levels ranged from B1 to C1 of the Common European Framework of Reference (CEFR) estimated from self-report scores of in-house placement and/or standardized English proficiency tests such as TOEFL iBT (Papageorgiou et al., 2015).

Word Association Task

In order to select a balanced set of cue words, we took a stratified sampling approach. This was done in two steps. First, we constructed a large pool of cue words (5,060 lemmas in total) and then computed ten lexical properties for each candidate word. A cluster analysis was run on the candidates using lexical properties to create subgroups within this cue-word pool. Second, using this cluster membership, we selected a final set of 96 cue words.

Constructing the cue word pool. We based our initial cue-word pool on a large database previously constructed in the domain of cognitive science: University of South Florida Free Association Norms (Nelson et al., 2004). We conducted a cluster analysis on this initial pool (Euclidian distance, Ward method) using R (R development Core Team, 2014). A total of ten lexical properties were included which measured the distributional, associational, semantic, and formal aspects of the words—(a) SUBTLEXus frequency (Brysbaert & New, 2009), (b) set size, (c) primary strength (proportions of the strongest associations) (Nelson et al., 2004), (d) polysemy (Balota et al., 2007), (e) semantic diversity (Hoffman et al., 2013), (f) concreteness (Brysbaert et al., 2014), (g) age of acquisition (Kuperman et al., 2012), (h) number of letters, (i) orthographic neighbors (Balota et al., 2007), and (j) character bigram
frequency (Balota et al., 2007). These features are described in Online Appendix S1; for a cluster dendrogram see Online Appendix S2.

Selecting cue words. A stratified sample of 96 cue words was drawn up in the following manner. First, to include cue words which were familiar to our participants, all infrequent words—operationalized as 5,000 word-level (i.e., lemma count) or above in New JACET 8000 (JACET, 2016)—were discarded. Second, cue words were randomly sampled from four clusters. Following prior research (Nissen & Henriksen, 2006; Zareva & Wolter, 2012), equal numbers of nouns, verbs and adjectives were sampled, resulting in a total of 96 cue words (32 nouns, 32 verbs, and 32 adjectives crossed with 24 cues from each of the four clusters). Online Appendices S3 and S4 present descriptive statistics for the lexical features of the selected cue words, and a list of the 96 cue words.

Procedure. The WA task was administered using Superlab 4.5 (Cedrus Corporation, 2011) on a Macintosh computer. The design of the task and procedure were adapted from Fitzpatrick and Izura (2011). Each trial included three phases, and each phase corresponded to a different slide on the computer screen: fixation, cue word presentation, and response (see Fig. 1). First, participants saw a fixation cross in the middle of the screen and heard a 1,000ms beep. Immediately following this, a cue word appeared on the next screen. Participants were instructed to respond orally with the first lexical item that came into their mind as quickly as possible. They were allowed to produce multiword responses (e.g., matter–no matter what). The response deadline was set to 7,000ms in order to minimize strategic responses by participants (Playfoot et al., 2016). Once a verbal response was recognized, participants were presented with a third screen, where they typed their response in an empty textbox without any time limit. Concurrently, the whole session was audio-recorded and the recordings were used for RT measures (see below).

[Insert Figure 1 here]
Response type coding. In this study, a response type measure was used to profile the participants’ response patterns (e.g., Fitzpatrick, 2006). We adapted a basic three-type distinction of meaning, position, and form, following a series of coding schemes proposed by Fitzpatrick and colleagues (e.g., Fitzpatrick, 2006, 2007; Fitzpatrick et al., 2015; Fitzpatrick & Izura, 2011). Before coding our data, all individual responses were assembled into a large list of cue-response combinations, as recommended by Fitzpatrick et al. (2015). Any identifiable misspellings were corrected. Any duplicates of cue-response combinations were also deleted to ensure that coding decisions were affected by neither the frequency of combinations nor patterns of individual associations, thereby boosting reliability.

Since the granularity of previous coding schemes varied widely, depending on the research focus (Fitzpatrick, 2006, 2007; Fitzpatrick et al., 2015; Fitzpatrick & Izura, 2011), a two-phase calibration procedure was conducted to construct a scheme that suited our research purposes. First, the first and second authors discussed any clarity issues related to previous coding schemes and modified relevant parts of the scheme as necessary. The second phase of calibration was conducted with an independent coder familiar with the L2 WA literature. The scheme and flowchart were explained by the first author, and 10 per cent of the sample data ($k = 300$) were coded independently. Subsequently, inter-coder agreement was evaluated using Fleiss’ $K$, a multinomial alternative to Cohen’s Kappa coefficient (Zapf et al., 2016).
This resulted in an acceptable agreement rate (Fleiss’ K = .669 [bootstrapped CI = .606–.731]).

The current coding scheme has nine categories, with one response category being No Response (see Table 1; see Online Appendix S5 for a detailed description). One important difference from previous coding schemes (e.g., Fitzpatrick, 2006; Fitzpatrick et al., 2015) concerned the use of external resources in the coding (see Mollin, 2009 for an attempt to use corpus-based collocation strength for scoring). For any meaning-based responses to be categorized in a Lexical Set, the cue-response combinations need to be listed as synonyms, antonyms, hypernyms, or hyponyms in WordNet (Fellbaum, 1998). For the Collocation category, cue-response combinations needed to be listed in at least one of the three collocation dictionaries that are often used in L2 vocabulary research—Longman Collocations Dictionary and Thesaurus (Mayor, 2013), BBI Dictionary of English Word Combinations (Benson et al., 1997), and LTP Dictionary of Selected Collocations (Hill & Morgan, 1997); otherwise, they were counted as Free Combination. From a substantive point of view, distinctions may thus be drawn between two sets of categories (i.e., Lexical Set versus Other Conceptual, and Collocation versus Free Combination) in terms of the tightness of association (cf. Fitzpatrick, 2006).

For the analysis of WA response type measures, we removed participants who produced the same type of response at a rate of 60 per cent or above (≥ 57 responses), since these participants may have employed a certain response strategy during the task. Two participants were excluded from the main analysis based on this criterion: one who produced 88 erratic responses (i.e., repetitions); and one who produced 72 form-based responses. Since we used multinomial count regression (see Statistical Analysis), the conventional +/− 3 SD criterion for univariate outliers was not applied.
Table 1
The Nine WA Response-Type Categories in This Study

<table>
<thead>
<tr>
<th>Category used in analysis</th>
<th>Sub-categories</th>
<th>Examples</th>
</tr>
</thead>
</table>
| Lexical Set               | Synonym        | • brilliant → smart  
|                           |                | • disappear → vanish |
|                           | Antonym        | • modern → antique  
|                           |                | • simple → complicated |
|                           | Hierarchical   | • capture → release (sister term)  
|                           |                | • succeed → try (entailment)  
|                           |                | • exchange → trade (hyponym and troponym) |
| Other Conceptual          | N/A            | • exercise → foot  
|                           |                | • area → zoo  
|                           |                | • darkness → forest |
| Collocation               | N/A            | • modern → architecture |
|                           |                | • skip → class  
|                           |                | • brush → up |
| Free Combination          | N/A            | • religious → problem  
|                           |                | • obtain → goods  
|                           |                | • spicy → noodle |
| Form                      | Affix manipulation | • bend → bent |
|                           |                | • feed → fed  
|                           |                | • scream → screaming |
|                           | Similar in form only | • matter → butter  
|                           |                | • dine → dinosaur |
|                           |                | • bend → vendor  
|                           | Two-step association | • delicious → feat (via feast) |
|                           |                | • win → loose (via lose) |
| Erratic                   | Erratic Repetition | • daily → map  
|                           |                | • proud → aprood (Non-English word) |
| No Response               | N/A            |         |
| Form and Meaning          | N/A            | • cow → cowgirl  
| Position and Meaning      | N/A            | • release → fish |

Note. Categories were adapted from previous coding schemes by Fitzpatrick and her colleagues (Fitzpatrick, 2006, 2007; Fitzpatrick et al., 2015; Fitzpatrick & Izura, 2011).
Figure 2 presents boxplots of frequency counts for each response type per participant. On average, Other Conceptual was the most frequent category (25.8 responses), followed by Collocation (19.4 responses), Lexical Set (15.8 responses), and so forth.

Figure 2. Distribution of word associations by response category ($n = 107$).

**Response time cleaning.** In order to evaluate the speed of word association, we measured the response times (RTs) of audio recordings using Praat software (Boersma & Weenink, 2018), instead of the voice detection function of the experimental software. This approach was taken because a previous study reported that 20 per cent of L2 WA responses were lost due to malfunctions of experimental software, noise, and participant hesitation (e.g., “uhh”, “err”) (Fitzpatrick & Izura, 2011). Specifically, we measured the interval between termination of the 1,000ms beep sound (see above) and the onset of each oral response. In the
analyses of RT data (RQ2), two participants were removed due to technical problems with the audio-recording. When trimming WA RT data, any trials that were below 300ms or above 3 SD of each participant’s mean were removed from the analysis (a total of 51 RTs [0.005%] were removed from the RT analysis).
Productive Vocabulary Levels Test

Three sections (2K, 3K, 5K word family levels) of the Productive Vocabulary Levels Test (Laufer & Nation, 1999) were administered to measure the form-recall vocabulary knowledge (i.e., size) of the participants.

Procedure. Participants were tested using a paper format. No time limitation was imposed on participants’ responses in order to elicit the upper limit of their vocabulary knowledge. To ensure that they understood the test format, participants were provided with instructions as well as an example.

Scoring. Following de Jong et al. (2012), PVLT was scored based on knowledge of lemma and derivation. Spelling errors were not penalized when they were identifiable. Errors in inflection (e.g., third person singular -s) were not penalized, considering the target construct. Subsequently, the number of correct responses was tallied across the three levels to create a single measure reflecting vocabulary size.

Primed Lexical Decision Task

The lexical decision task was designed to assess priming effects and lexical decision time. A total of 240 prime-target pairs were created (see Appendix S6). The list consists of five categories: (a) 20 semantically-related trials (e.g., vary–differ), (b) 20 collocationally-related trials (e.g., surrounding–environment), (c) 40 unrelated critical trials (e.g., volcano–children), (d) 60 unrelated filler trials (e.g., castle–student), and (e) 100 non-word trials (e.g., pull–varn). Two measures of the priming effect—semantic and collocational priming—were derived from RT differences between related trials and an unrelated critical trial; that is, (c)–(a) and (c)–(b), respectively. Overall, average lexical decision times were derived from the RT of (c). The last two categories, (d) and (e), were included in order to prevent strategic responses. In constructing nonword targets, WordGen (Duyck et al., 2004) were used, and its
output were manually checked to conform to English phonotactics. In order to achieve greater coherence across the lexical tasks, real words were sampled from the words remaining in the WA cue-word pool mentioned above (For more details of the stimulus construction, see Online Appendix S6; for a list of related prime–target pairs see Online Appendix S7).

Procedure. The experimental session was run using DMDX on a Windows 10 computer (Refresh rate = 16.67ms). Participants’ responses were recorded with the keyboard response to this computer. Figure 3 illustrates the procedure. First, participants were presented with a fixation cross for 2,000ms. Second, following the presentation of a prime word for 300ms, the target was presented in bold letters for 3,000ms and responses were made. Participants were told to respond with the “j” key if they saw an English word and the “f” key if they saw a non-word. The session began with ten practice items (five real-word, five non-word items; no semantically- or collocationally-related trials were included in the practice session).

Figure 3. Procedure for a single primed lexical decision Trial.

Data coding. The primed lexical decision task yielded two sets of measures: Lexical Decision Time (LDT) and two priming effect measures (Semantic and Collocation Priming).
Prior to coding the data, we adopted a set of participant exclusion criteria. We first set the minimum accuracy rate to 75%. No participant had to be removed based on this criterion. Second, any single trials whose RTs were below 300ms or above 3 SD of mean RTs were removed. This resulted in the loss of 0.5% of the data (77 trials). Third, any prime–target pairs that failed to reach an accuracy rate of 75% were removed, which led to an additional loss of 9 items. These losses did not significantly affect the contrasts across the prime–target conditions.

Subsequent to this process, mean RTs for each prime–target condition were calculated, and the mean RTs for (c) unrelated critical trials were used as LDT. Two priming effects were computed for each of the (a) semantically-related and (b) collocationally-related trials using the following formula: Mean RTs for (c) unrelated critical trials – Mean RTs for (a) semantically-related or (b) collocationally-related trials, respectively.

One possible challenge when producing two separate types of measures (LDT and priming effects) from a single task is that the priming effect measures (calculated as [c]–[a] and [c]–[b], respectively) could covary with the LDT (RT of [c]). To address this issue statistically, these two priming effect measures were orthogonalized from (c) LDT, and the residuals of each regression model were used as the final priming measures. Thus, any effects of the two priming variables would be equivalent to the effect of priming after statistically controlling for LDT.

**Spontaneous Speech Task**

We administered an eight-frame cartoon narrative speech describing the story of a man and a woman agreeing on a plan to go to the beach (Suzuki, 2018; see Online Appendix S8).
**Procedure.** Participants were first presented with the cartoon and given two minutes of planning time. No time limit was set for the oral narration. Note-taking was not allowed during preparation or narration.

**Data coding.** The speech data were transcribed, and all hesitations (e.g., lexicalized fillers, false starts, repetitions) were deleted for lexical analyses. To represent the multi-faceted construct of lexical richness (Eguchi & Kyle, 2020; Kyle et al., 2017; Read, 2000), three measures of lexical richness (lexical diversity, lexical sophistication, and phraseological sophistication) were derived from the transcribed speech data.

The Measure of Textual Lexical Diversity Original All Words (henceforth MTLD) was selected to index lexical diversity. MLTD is defined as “the mean length of sequential word strings in a text that maintain a given TTR [Type-Token Ratio] value” (McCarthy & Jarvis, 2010, p. 384). MTLD has been demonstrated to be one of the few measures robust to particularly short texts (i.e., 50 words; Zenker & Kyle, 2021). For lexical sophistication, we selected Brysbaert’s concreteness for content words. This measure has been found to better reflect advanced word use in spoken mode, compared to frequency-based indices (see Eguchi & Kyle, 2020). For phraseological sophistication, we selected averaged bigram Mutual Information (MI), which is defined as the mean MI score of learner-produced bigrams (i.e., two-word contiguous sequences) that occur in the reference corpus, here the Corpus Of Contemporary American (COCA; Davies, 2009) spoken section (Kyle et al., 2017). These indices were computed using the Tool for Automatic Analysis of Lexical Diversity 1.3.1 (see Kyle et al., 2020) (MTLD) and the Tool for Automatic Analysis of Lexical Sophistication 2.8.1 (Kyle et al., 2017) (Brysbaert’s concreteness and Bigram MI).

**Procedure**

Participants were tested individually in two experimental sessions (each of them lasted for 60–90 mins). In the first session, participants completed the WA task followed by
the primed lexical decision task, with a break between the two. Participants also completed a background questionnaire. During the second session, participants completed the speaking task followed by PVLT. The order of the sessions was counterbalanced across participants. The two sessions were also separated by at least one week to minimize potential effects across tasks.

**Statistical Analysis**

The first research question of the study asks whether WA response profiles are predicted by concurrent vocabulary measures: (a) PVLT, (b) lexical decision time (LDT), (c) two priming effects, and (d) three lexical richness indices. A persistent obstacle to WA research lies in the difficulty of dealing with the categorical nature of response profiles (e.g., Fitzpatrick, 2006; Fitzpatrick & Izura, 2011), which violates the assumptions of linear regression. In order to deal with the categorical outcome variable, multinomial count regression analysis was conducted through the mblogit() function in the mclogit package 0.8.5.1 (Elff, 2020) using R statistical software 3.6.0 (R development Core Team, 2014). This model takes a series of count data and tests whether and how the expected proportions of each category, compared against a predetermined reference category, would change (i.e., be predicted) by adding a series of predictors. Thus, it can be considered a multivariate extension of logistic regression (because count data are converted to proportion data) calculated against a reference category.² For the purposes of presentation, Other Conceptual (e.g., area–zoo) was selected as the reference category, since it was the most frequent category and was found to be relatively unaffected by the predictors.

For model selection, we used the best-subset regression method with the dredge() function in MuMIn package 1.43.6 (Bartoń, 2019). Best-subset regression can be considered an alternative to stepwise regression which runs all possible models with the predictors and subsequently ranks them according to model fit indices, such as the Bayesian Information
Criterion (BIC), and the sample-size-corrected Akaike Information Criterion (AICc). Plots for predicted probabilities were produced using ggplot2 package 3.2.1 in R (Wickham, 2016).

The second research question pertained to RT data from the WA task. Using the same vocabulary measures as RQ1, linear regression analyses were conducted with the best-subset method. The best model was checked for assumptions of linear residuals and multicollinearity by tolerance values.

Assumptions for the main analyses were checked as follows. All the measures (but WA response types) were confirmed to be normally distributed (see Appendix S9 for descriptive statistics). Subsequently, univariate outliers (± 3 SD) were removed for predictor variables, resulting in 107 participants for RQ1 (Response type measure) and 105 participants for RQ2 (RT measure). In the multinomial regression model, all the predictors were standardized using the z-score formula. The R code and the statistical output in an html format can be found in the Online Appendix; the original rmarkdown file and anonymized dataset can be found on IRIS database for those who would like to reproduce the analysis (https://www.iris-database.org/iris/app/home/index ).
Results

Lexical Competence Underlying L2 WA Response Profiles

RQ1 investigated the contribution of L2 lexical competence, which was defined as size, organization, accessibility and lexical richness, to WA response profiles. The result of the best-subset regression model indicated that WA response profiles were best predicted by Productive Vocabulary Levels Test (PVLT), Collocation Priming, the Measure of Textual Lexical Diversity (MTLD), and Concreteness (Nagelkerke’s pseudo \( \text{adj } R^2 = .901 \); for details of best-subset regression see Online Appendix S10).

Table 2 presents the LogOdds, SEs and \( p \)-values of the final model, where SEs were adjusted for the dispersion parameter of the data (see Online Appendix S11 for further details). The intercepts tested whether the selected category was significantly different from the reference category (Other Conceptual; e.g., area–zoo) when all predictors were at their means. The slopes indicated whether the effects of predictor variables were statistically different from those of Other Conceptual (expressed on a LogOdds scale). For instance, the intercept for the Lexical Set (e.g., brilliant–smart, capture–release) category was significantly lower than Other Conceptual (.603 times; \( p < .001 \)), and the slope for PVLT on Lexical Set was 1.150 times steeper than Other Conceptual.
Table 2
Abbreviated Summary of the Best Multinomial Count Regression

<table>
<thead>
<tr>
<th></th>
<th>LogOdds</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical Set</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-.505</td>
<td>.055</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>PVLT</td>
<td>.140</td>
<td>.063</td>
<td>.028</td>
</tr>
<tr>
<td>Collocation Priming</td>
<td>.058</td>
<td>.056</td>
<td>.301</td>
</tr>
<tr>
<td>MTLD</td>
<td>-.062</td>
<td>.059</td>
<td>.290</td>
</tr>
<tr>
<td>Concreteness</td>
<td>-.007</td>
<td>.060</td>
<td>.908</td>
</tr>
<tr>
<td><strong>Collocation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-.299</td>
<td>.052</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>PVLT</td>
<td>.064</td>
<td>.059</td>
<td>.283</td>
</tr>
<tr>
<td>Collocation Priming</td>
<td>-.134</td>
<td>.054</td>
<td>.013</td>
</tr>
<tr>
<td>MTLD</td>
<td>.030</td>
<td>.055</td>
<td>.588</td>
</tr>
<tr>
<td>Concreteness</td>
<td>-.137</td>
<td>.056</td>
<td>.016</td>
</tr>
<tr>
<td><strong>Free combination</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-.745</td>
<td>.060</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>PVLT</td>
<td>.015</td>
<td>.067</td>
<td>.823</td>
</tr>
<tr>
<td>Collocation Priming</td>
<td>-.163</td>
<td>.062</td>
<td>.009</td>
</tr>
<tr>
<td>MTLD</td>
<td>.044</td>
<td>.062</td>
<td>.477</td>
</tr>
<tr>
<td>Concreteness</td>
<td>-.033</td>
<td>.065</td>
<td>.614</td>
</tr>
<tr>
<td><strong>Form</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-.945</td>
<td>.064</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>PVLT</td>
<td>-.113</td>
<td>.072</td>
<td>.115</td>
</tr>
<tr>
<td>Collocation Priming</td>
<td>.000</td>
<td>.065</td>
<td>.995</td>
</tr>
<tr>
<td>MTLD</td>
<td>-.151</td>
<td>.068</td>
<td>.027</td>
</tr>
<tr>
<td>Concreteness</td>
<td>-.067</td>
<td>.070</td>
<td>.340</td>
</tr>
<tr>
<td><strong>Erratic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.924</td>
<td>.097</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>PVLT</td>
<td>-.311</td>
<td>.098</td>
<td>.002</td>
</tr>
<tr>
<td>Collocation Priming</td>
<td>-.208</td>
<td>.095</td>
<td>.029</td>
</tr>
<tr>
<td>MTLD</td>
<td>.107</td>
<td>.092</td>
<td>.244</td>
</tr>
<tr>
<td>Concreteness</td>
<td>.089</td>
<td>.100</td>
<td>.370</td>
</tr>
<tr>
<td><strong>Form and meaning</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-3.164</td>
<td>.169</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>PVLT</td>
<td>.025</td>
<td>.193</td>
<td>.897</td>
</tr>
<tr>
<td>Collocation Priming</td>
<td>.060</td>
<td>.170</td>
<td>.725</td>
</tr>
<tr>
<td>MTLD</td>
<td>-.123</td>
<td>.180</td>
<td>.494</td>
</tr>
<tr>
<td>Concreteness</td>
<td>-.058</td>
<td>.183</td>
<td>.751</td>
</tr>
<tr>
<td><strong>Position and meaning</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-5.454</td>
<td>.536</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>PVLT</td>
<td>-.149</td>
<td>.565</td>
<td>.793</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Collocation Priming</td>
<td>.367</td>
<td>.469</td>
<td>.435</td>
</tr>
<tr>
<td>MTLD</td>
<td>-.060</td>
<td>.517</td>
<td>.907</td>
</tr>
<tr>
<td>Concreteness</td>
<td>-.102</td>
<td>.547</td>
<td>.852</td>
</tr>
</tbody>
</table>

*Note.* The reference category was Other Conceptual for the entire model; Dispersion parameter was estimated to be 3.15; Standard Errors and p-values, and Confidence Intervals were adjusted for overdispersion through the `dispersion()` function in the `mclogit` package (Elff, 2020).

In order to paint a more complete picture of the relationships, Figures 4–7 show plots of predicted probabilities using smoothed lines and 95% confidence intervals represented as ribbons around the lines. These plots used marginal means, describing changes in outcomes when other predictors are held at their means. Figure 4 illustrates the effect of PVLT on the WA response profile. The probability of Other Conceptual response (the reference category) only slightly decreased (non-significant considering the plotted CI) as a function of PVLT score. For many of the remaining response types, CIs were overlapping when PVLT scores were below -2SD; however, three distinct patterns were identified as a function of PVLT: (a) increase, (b) constant, and (c) decrease. First, the increased response types included Lexical Set and Collocation (e.g., *modern–architecture*), the two categories capturing tight associations (cf. Fitzpatrick, 2006). Compared to Other Conceptual (the reference category), these two categories yielded positive trends (1.150 times for Lexical Set \( p = .028 \) and 1.066 for Collocation \( p = .283 \)). Second, the proportion of Free combination (i.e., position-based associations absent from collocation dictionaries consulted; hence considered to be loose associations) appeared to be relatively constant across PVLT scores. Third, Form and Erratic categories showed negative trends compared to Other Conceptual (.893 time \( p = .115 \) and .733 time \( p = .002 \))—the higher the PVLT score, the less likely that Form and Erratic responses were produced in word association.
Figure 4. The marginal effect of Productive Vocabulary Levels Test (PVLT) on the word association response profile.

Figure 5 presents the effect of Collocation Priming on the response profile, when other predictors were held at their means. Interestingly, collocational priming appeared to capture a trade-off between Meaning- versus Position-based responses. This contrasted with the fact that PVLT differentiated response profiles according to the tightness of associations (compare Figs 4 and 5). Specifically, the reference category—Other Conceptual—increased from around 25% to over 30% as a function of increased priming effects. The slope for Lexical Set (1.060 times \( p = .301 \)) did not differ significantly from this, suggesting that the two Meaning-based categories increased. In contrast, two Position-based categories—Collocation (.874 times \( p = .013 \)) and Free combination (.850 times \( p = .009 \))—decreased, as did Erratic (.812 times \( p = .029 \)).
Figure 5. The marginal effect of Collocation Priming (controlled for lexical decision time) on the word association response profile.

Figure 6 illustrates the effect of MTLD on the WA response profile. Other Conceptual showed a slight increasing trend. The other categories were then compared with this. The most noticeable pattern was the significant negative slope of Form (.860 time; $p = .002$), suggesting that participants with higher MTLD scores were less likely to produce form-based responses in the WA task. Concerning other response types, Collocation and Free Combination showed slight (non-significant) upward trends, while Lexical Set tended to decrease with higher MTLD values (.940 time; $p = .290$).

[Insert Figure 6 here.]
Finally, Figure 7 presents the effect of concreteness (plotted on a reversed scale for ease of interpretation) on the response profile. When concreteness was high (on the left of Fig. 7), participants appeared to produce similar amounts of Lexical Set, Collocation, and Free Combination (CIs were overlapping at around 15 to 20%). As we move along the concreteness continuum, one noticeable pattern emerges for Collocation (1.15 times = exp[.137]; \( p = .016 \)), which increased by almost 10%. The remaining response types were either constant or decreasing. Thus, it appears that Collocation responses replaced the other response types as a function of abstract (i.e., sophisticated) words in spontaneous speech.
Figure 7. The marginal effect of concreteness (scale reversed) on the word association response profile.

**Lexical Competence Underlying L2 WA Response Time**

RQ2 investigated whether three-dimensional lexical competence and lexical performance in spontaneous speech predict the RT of WA responses. The final model included only PVLT and Lexical Decision Time (LDT) (see Online Appendix S12 for the results of best-subset regression modelling). Considering its theoretical compatibility, LDT was entered into the model first (Table 3). Overall, the model explained 28.5% of the variance in WA RTs. This indicates that participants with a larger vocabulary size and a shorter LDT tend to respond more quickly during WA tasks (see Figs 8 and 9 for scatter plots).
### Table 3

**Results of a Linear Multiple Regression for Word Association Response Time**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Adj. $R^2$ change</th>
<th>Adj. $R^2$</th>
<th>Estimate</th>
<th>SE</th>
<th>$\beta$</th>
<th>t-value</th>
<th>$p$</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td></td>
<td>2.897</td>
<td>.444</td>
<td>–</td>
<td>6.518</td>
<td>&lt; .001</td>
<td>.807</td>
<td></td>
</tr>
<tr>
<td>Lexical Decision Time</td>
<td>.159</td>
<td>.159</td>
<td>.001</td>
<td>.000</td>
<td>.232</td>
<td>2.510</td>
<td>0.014</td>
<td>.807</td>
</tr>
<tr>
<td>PVLT</td>
<td>.285</td>
<td>.126</td>
<td>-.035</td>
<td>.008</td>
<td>-.404</td>
<td>-4.452</td>
<td>&lt; .001</td>
<td>.807</td>
</tr>
</tbody>
</table>
Figure 8. The relationships between LDT and WA RTs.

Figure 9. The relationships between PVLT and WA RTs.
Discussion

Responding to a recent call to examining the construct underlying L2 WA tasks (see Fitzpatrick & Thwaites, 2020), the current study investigated the extent to which L2 WA response profiles and response times were predicted by a battery of lexical measures: (a) Productive Vocabulary Levels Test (PVLT), (b) semantic priming, (c) collocational priming, (d) lexical decision time (LDT), (e) Measure of Textual Lexical Diversity (MTLD), (f) concreteness, and (g) bigram Mutual Information. The results of multinomial count regression (RQ1) indicated that four out of seven measures (PVLT, collocational priming, MTLD, concreteness) predicted a large proportion of WA response profiles (Nagelkerke’s pseudo $R^2 = .901$). Further, the results of linear regression (RQ2) demonstrated that two of the measures (lexical decision time and PVLT) explained 28.5% of the variance in WA RTs. Overall, the evidence provided here lends support to the hypothesis that WA task performance reflects multiple dimensions of lexical competence and performance (Fitzpatrick, 2012).

Word Association Response Profile Reflects Lexical Competence—Size and Organization

With regard to the response profiles (RQ1), the final model suggested that the profiles were associated with three broad sub-dimensions of lexical competence and performance—vocabulary size (PVLT), organization (collocational priming), and lexical richness (MTLD and Concreteness). The primary predictor was vocabulary size, accounting for the increases in Lexical Set and Collocation (tight associations; cf. Fitzpatrick, 2006) irrespective of the superordinate response categories of Meaning- or Position-based association. It also predicted the decrease in Forms and Erratic responses. Given that we operationalized vocabulary size through a controlled form-recall test (PVLT; Laufer & Nation, 1999), the construct of size
tapped into in this study would not only reflect familiarity to the formal aspect of words, but also represent the stability of overall form-recall knowledge. Accordingly, the current findings lend more direct support to the hypothesis that size (the number of form-meaning mappings) may allow increased access to responses that have semantically and/or collocationally tight relationships with the cue words—Lexical Set and Collocation. Moreover, the decrease in Forms and Erratic responses may indicate that semantic and collocational lexical associations with the cue words tended to be prioritized over form and erratic associations (see also Jiang & Zhang, 2019). This supports our hypothesized link between vocabulary size and WA response profile, as well as previous findings (e.g., Fitzpatrick, 2006; Wolter, 2001).

Regarding organization (as indexed by collocational priming), a contrasting picture emerged (Fig. 5). Particularly with regard to Meaning- and Position-based categories, while vocabulary size had to do with the increases in response types capturing tight associations (i.e., Lexical Set and Collocation), collocational priming did not seem to account for WA response profiles in the same way. Instead, collocational priming was associated with a more global distinction between Meaning- and Position-based responses irrespective of tightness, namely, increases in two Meaning-based responses and decreases in Position-based ones. Since collocational priming effects would capture learners’ sensitivity or the availability of a collocational inter-lexical link between lexical entries, we hypothesized that learners with high collocational priming scores would be ready to produce more Collocation and Free Combination responses in the WA task. However, our regression model showed the opposite. There are two potentially complementary explanations of this finding.

The first possibility is that, in the WA task, when speakers possess rich inter-lexical networks (as measured by collocational priming), and when both meaning- and position-based responses are available in the lexicon for such enhanced networks, meaning-based
responses might be prioritized in WA tasks. This might have resulted from the fact that the cognitive processes underlying WA response types and the priming task might be different. While an inter-lexical network to be primed (e.g., formal, semantic, etc.) can be specified by researchers in the design of priming tasks, WA tasks do not impose such constraints. Hence, it seems plausible to argue that, in the WA task, even though both semantic/ conceptual- and collocational/ co-occurrence inter-lexical connections are available in their mental lexicon, learners with elaborate lexical organization may have tended to produce Meaning-based responses at the expense of Position-based responses. This may be in line with meaning-primary explanation of WA tasks (e.g., Zareva & Wolter, 2012).

At this point, one may notice that the first scenario, the meaning-primary explanation of WA tasks, may run counter to the present findings concerning the increases in Collocation as a function of PVLT. It may also go against previous findings demonstrating collocation (or syntagmatic) responses as a frequent response type by proficient users of the target language (e.g., Fitzpatrick, 2006; Nissen & Henriksen, 2006). Recall that our statistical model (and Figs 4–7) presupposes that the significant effect of one predictor is conditional on the means of all the remaining predictors in the model (i.e., marginal means). Accordingly, the observed patterns should be interpreted in a multidimensional space: when vocabulary size was fixed at the same level, our participants’ response profiles still varied along the collocational priming continuum—those who preferred Meaning-based responses or Position-based responses, respectively (see Fig. 5). Put differently, the strengths of collocational priming effects accounted for specific individual tendencies to prioritize Meaning-based responses over Position-based ones, which were hidden behind the averaged increasing trends in both Lexical Set and Collocation as a function of vocabulary size. Given this interpretation, the present findings may shed light on potential sources of individual response preferences (Fitzpatrick, 2009). That is, while the previous study listed some potential sources of
response preferences, such as verbal intelligence, creativity, personality, etc. (Fitzpatrick, 2009), it could be that such individual orientations to Meaning- or Position-based associations, reported in Fitzpatrick (2009), partly arise from the multi-dimensional nature of lexical competence (Meara, 2005). The current interpretation may also provide some insights into inconsistent findings in WA research; that is, the inconsistencies across studies might stem from comparability due to the fact that these studies used limited numbers of external measures that tap into very different constructs of L2 skills (e.g., vocabulary size, TOEFL total scores, C-test scores, etc.). Thus, our results would underscore the importance of carefully considering the specific lexical constructs each external measure would represent in the multidimensional space (e.g., Meara, 2005; Schmitt, 2014) in order to better understand the constructs underlying L2 WA task performance.

The second scenario which explains why our collocation priming measure was associated with the increase in meaning-based responses has to do with the interpretation of collocational priming effects. Specifically, it is possible that our collocational priming measure may not have been a pure collocational measure, but rather one that involved the activation of both lexical and conceptual networks. This was because prime–target pairs can be conceptually related (e.g., fish–swim) beyond the tendency for textual co-occurrences. If so, the reason why collocational priming was a significant predictor of WA tasks may be due to additive priming effects, or associative boost, stemming from the two loci (collocational and conceptual) (see Lucas, 2000). In any case, our findings clearly indicate that WA response profiles are systematically shaped not only by vocabulary size, but also by the organization aspect of lexical competence (Meara, 2005). More precise relationships should be clarified in future studies.

**Word Association Response Profile Reflects Spontaneous Lexical Production**
As with lexical richness indices, the best-fit model included two predictor measures—MTLD and Concreteness. The significant associations between spontaneous lexical production and WA responses suggest that some cognitive processing mechanisms may be shared across WA tasks and free speech production tasks (Dóczi & Kormos, 2016). In order to produce a lexically diverse oral narrative, particularly given the predefined content of the cartoon prompt, one needs to first effectively activate a range of lexical entries relevant to their intended message (here, the story described in the cartoon) (Kormos, 2006; Levelt et al., 1999). Failure in spontaneous lexical activation would manifest as greater lexical repetition, thus lowering lexical diversity (McCarthy & Jarvis, 2010). It can thus be speculated that the spontaneous nature of WA required a process of searching for response words according to the conceptual specification and co-occurrence information of the given cue words.

Our findings showed that a low lexical diversity score was associated with a relatively high proportion of Form-based responses (around 15%; the median of Form-based responses was around 8%; Figs 2 and 6). As the speakers produced more lexically diverse speech, Form-based responses declined steadily, while Other conceptual, Free-combination and Collocation responses showed (slight) positive trends. This may indicate that speakers increasingly drew on a range of inter-lexical links during WA tasks, although the changes in each response type were subtle, possibly because the responses were dispersed across multiple response types.

With regard to Concreteness, our model (Fig. 7) indicated that the variance associated with this index explained the increases in Collocation responses. Abstract words are found to be generally difficult to retrieve from the lexicon and acquired later due to their referents’ low perceptual saliences (Crossley et al., 2016). Given that WA response types can reflect activation routes (Fitzpatrick & Izura, 2011), those who can draw on such abstract lexical items in spontaneous production (i.e., lexical sophistication) may have developed activation
CONSTRUCT VALIDITY OF L2 WORD ASSOCIATION TASKS

routes involving lexical co-occurrence information in the target language. This may be consistent with previous findings that collocation (or syntagmatic) responses are the hallmark of a proficient user of the target language (e.g., Fitzpatrick, 2006; Nissen & Henriksen, 2006)).

To supplement these interpretations, it is also worth noting that although Bigram MI was not included in the final regression model, it tended to be associated with increases in Free Combination when PVLT was held constant (see Online Appendix S13). Thus, it appears that all the three lexical richness indices were, more or less, associated with increases in Position-based responses. These consistent relations between lexical use measures and Position-based responses in WA might align with the fact that, unlike PVLT and priming tasks, the spontaneous production of speech requires the construction of phrases and clauses, which would involve the retrieval of the syntactic properties of lexical items and/or collocational knowledge of the target language, including formulaic sequences (Kormos, 2006). In other words, the variance in Position-based responses in the WA task, significantly explained by lexical richness measures, may also indicate that Position-based responses in WA tasks tap into collocational knowledge and/or co-occurrence information of lexical items (i.e., knowledge represented in lemma; Kormos, 2006; Levelt et al., 1999).

WA Response Times Reflect Vocabulary Size and Accessibility

The results of a linear regression (RQ2) indicated that 28.5% of the variance of RT in the WA task was explained by two dimensions of L2 lexical competence: accessibility (15%; Fig. 8) and vocabulary size (12.6%; Fig. 9). This is in line with previous findings from Fitzpatrick and Izura (2011), who demonstrated a strong association between WA RTs and vocabulary size. The current study extended their findings by showing that RTs are associated with not only vocabulary size, but also with accessibility.
Two predictors can explain the underlying cognitive processes that may affect RTs in L2 WA tasks. First, accessibility (as indexed by Lexical Decision Time) would reflect the time course of recognizing and accessing the lexical representation of a cue word. The faster that speakers can access lexical representations, the shorter are their overall WA RTs. With regard to vocabulary size, the larger the lexicon, the more likely that a learner will have complete lexical entries (Wolter, 2001). This will ease the demand on cue word identification and response encoding, respectively, leading to faster RTs overall. On the other hand, our findings suggest that priming measures or lexical richness would not be reflected in WA RT measures. This may mean that RTs are capable of measuring a relatively specific sub-dimension of lexical competence, namely, accessibility (Meara, 2005).

A Theoretical Model of L2 Word Association

Previously, apparent links between WA tasks and lexical competence were at times questioned (e.g., Cremer et al., 2010). Nonetheless, our regression models demonstrated that three-dimensional lexical competence systematically explained WA performance. One critical issue remained regarding the construct validity of WA tasks/measures, however, is the lack of a theoretical model that gives an account of loci in L2 WA performance (see Fitzpatrick & Thwaites, 2020; Meara, 1982). We propose a tentative theoretical model for WA task performance, integrating the current findings with the extant bodies of previous work on WA tasks (e.g., Fitzpatrick & Izura, 2011), bilingual word recognition mechanisms (e.g., Dijkstra & van Heuven, 2002), and speech production mechanisms (Kormos, 2006; Levelt et al., 1999). As shown in Figure 10, cognitive processes during a WA task essentially consists of three stages: (1) cue word recognition, (2) inter-lexical link activation, and (3) response encoding and articulation. In each stage, we assume that relevant pieces of information in the mental lexicon are accessed (lexeme, lemma, and lexical concept; Levelt et
al., 1999) and that activation generally follows the principle of activation spreading in human cognition (Collins & Loftus, 1975).

Figure 10. A proposed theoretical model of word association.

A WA trial begins with a (visual) presentation of a cue word to the participants—the “cue word recognition phase” (e.g., matter at bottom left in Fig. 10). This visual cue activates a set of orthographic candidates in the lexicon in the course of word recognition processes, such as mattress and cutter (i.e., parallel activation of orthographic representations; Dijkstra & van Heuven, 2002). It is assumed that orthographic candidates can incrementally activate other components of their lexical representation, including phonology, lemma, and concepts. Concurrently, a lexical entry consistent with the orthographic input will be identified (Dijkstra & van Heuven, 2002). As supported by the present findings, vocabulary size and accessibility then contribute to the efficiency of cue word recognition (see also Wolter, 2001).
As fragments of cue word lexical entry—concept, lemma, phonology, and/or orthography—are activated, these nodes are responsible for initiating the second stage, “inter-lexical link activation.” Although the dependency of the time course of this process on cue word recognition is yet to be revealed, the model tentatively assumes that the fragment of activation at each level of cue-word lexical entry can spread further to related lexical items via inter-lexical links (see dotted lines in Fig. 10, which represent these inter-lexical links). Taking the previous example (i.e., matter as a cue), the conceptual node, MATTER, would activate other lexical concepts related in terms of lexical semantics (e.g., synonyms, hypernymy) and/or potential thematic/semantic roles (e.g., predicate-theme, agent-predicate, etc.) (Collins & Loftus, 1975). Thus, activation via lexical concepts can result in either Meaning-based (e.g., PROBLEM, ISSUE) or Position-based (e.g., DEAL WITH) responses. From the lemma node, co-occurrence (or syntactic) information of the cue word may be activated, leading to Position-based responses such as matter—doesn’t matter. The inter-lexical links posited at the lexeme (either orthography or phonology) can trigger responses such as matter—shatter, which are subsequently judged to be Form-based responses. In sum, the second stage results in the temporal activation of a set of lexical links via multiple activation routes in the entire mental lexicon, with varying strengths (all the connected nodes, dotted lines, in Fig. 10). The findings of this study demonstrate that different external measures highlighted distinct aspects of response profiles—vocabulary size for tight Meaning and Position associations (i.e., Lexical sets and Collocation), collocational priming for the trade-off between Meaning versus Position-based preferences, and lexical richness for Position-based responses. These findings indicate that the response type categorization employed in the current study captured important qualitative differences in WA behaviors. Given this interpretation, the findings lend support to the process posited at the inter-lexical activation stage of the proposed model, as well as to the idea that the resulting response types
(Fig. 10 on the right) may reflect such distinct “activation route[s]” (Fitzpatrick & Izura, 2011, p. 394).

Finally, among the activated links, a response item is eventually selected. A phonological representation of the selected item is activated and then prepared for overt production (Levelt et al., 1999). Following the principle of activation spreading (Collins & Loftus, 1975), the model assumes that the higher the level of activation of a given lexical link is, the more likely that the link and attached lexical item will be selected as a response. However, previous literature also indicated possible deliberate manipulations of responses (see Playfoot et al., 2016) and/or preferred response patterns (e.g., Fitzpatrick, 2007, 2009) in WA tasks. Thus, we propose that both lexical competence and task-specific factors, such as participants’ perceived goals of the task (see Fitzpatrick & Izura, 2011, p. 376), will ultimately shape response profiles (see Fig 10, on the right). While lexical competence would be responsible for the activation spreading across the lexicon, the perceived goals of the task may concurrently invoke cognitive processes, such as “perspective taking” (Levelt et al., 1999, p. 8), “self-monitoring” (Levelt et al., 1999, p. 33) or “decision-making heuristics” (Playfoot et al., 2016, p. 610; see also Dijkstra & van Heuven, 2002). This process would allow speakers to choose to attend to specific aspects of conceptual and/or lexical information, take control over the re-initiation of the lexical search process and/or inhibit highly activated lexical information.

Once the response item is selected, the success rate and speed of response encoding would depend on the completeness of lexical information of the response item (as supported by the significant impact of vocabulary size on RTs; see Fig. 9). In sum, this psycholinguistic model for WA task is tentatively presented, and it is hoped that the model will stimulate further validation research on WA tasks based on this model.
Conclusions and future directions

Although word association has been used to investigate the L2 mental lexicon, surprisingly little attention has been paid to the specific constructs of lexical competence that are reflected in WA tasks (Schmitt, 2014). To address this issue, the current study predicted two kinds of WA measures from a range of measures of lexical competence and lexical richness. The results indicated that WA response profiles reflect vocabulary size, organization, and spontaneous lexical production, while WA RTs are associated with vocabulary size and accessibility. These findings suggest that WA performance reflects multiple aspects of lexical competence.

There are several limitations and an important agenda to set for further research on WA tasks. First, we have chosen only response profiles and the response time of WA performance for the sake of comparability with previous studies (e.g., Fitzpatrick, 2006; Fitzpatrick et al., 2015; Fitzpatrick & Izura, 2011; Nissen & Henriksen, 2006; Wolter, 2001). Future research can extend the scope by using other sets of measures, such as the size of associative domains (Zareva, 2007) and the quality of collocational associations (e.g., Mollin, 2009; Zareva & Wolter, 2012). Second, although efforts were made to cover three-dimensional lexical competence (size, organization, and accessibility; Meara, 2005), we had to be selective regarding which versions of receptive vs productive measures to use to represent each dimension. For instance, some important dimensions of lexical competence (e.g., productive knowledge of accessibility) could not be examined. Future research should extend the current findings by tapping into both receptive and productive lexical knowledge through productive tasks such as primed picture naming tasks as well as primed lexical decision tasks. Third, our priming task may have biased participants toward yes responses because of unbalanced yes/no trials. Also, our collocational prime-target pairs did not rule out potential conceptual activations. These characteristics of priming tasks in the current study
might have under- or over-estimated the contribution of priming measures to WA performance. Third, we have used only a single oral task to gauge lexical use in speech production. As different task design features may influence lexical richness performance or language performance in general (de Jong et al., 2012; Read, 2000), it would be helpful to replicate the current findings with different types of speech tasks. Finally, our proposed model of WA task performance does not directly speak to developmental perspectives on the L2 lexicon. Despite these limitations, our findings contribute to the understanding of the construct underlying L2 WA task performance. Taking this study as a stepping stone, more research is needed to refine and validate different types of WA measures so that they can be used to reveal the complex nature of vocabulary development and L2 lexical competence (Meara, 2005; Schmitt, 2014).
NOTE

1 Another prominent model is Daller et al. (2007)’s lexical space model, which comprises (a) breadth, (b) depth, and (c) fluency. While (a) breadth and (c) fluency dimensions do not differ significantly from (a) size and (c) accessibility in Meara (2005), (b) depth is conceptualized as the quality of knowledge about individual words, often operationalized using Nation’s (2013) framework of vocabulary knowledge.

2 The choice of the reference category is arbitrary, and does not affect the overall fit of the model (i.e., deviance). However, it will affect how each parameter is calculated and, potentially, interpreted. This drawback can be overcome by producing a set of expected probabilities for each dependent category as a function of the independent variable. Plots of these predicted probabilities and their CIs will offer a complete picture of how predictor(s) impact on the probabilities of multinomial responses.

3 With regard to inter-lexical link activation in the lemma stratum, the proposed model refrains from taking any theoretical stance on whether this process happens via holistically stored chunk nodes or entrenched memory traces between multiple independent lemmas resulting from repeated usage experiences.
References


Bartoń, K. (2019). *MuMIn: Multi-Model Inference* (1.43.6) [Computer software]. https://cran.r-project.org/web/packages/MuMIn/index.html


https://doi.org/10.1080/15434303.2020.1844205


https://doi.org/10.1017/S0140525X99001776

https://doi.org/10.3758/BF03212999


https://ggplot2.tidyverse.org


https://doi.org/10.1017/S0272263101001024


https://doi.org/10.1093/applin/amr011

https://doi.org/10.1177/0267658307076543