Spring 2018

Modeling Memory: Exploring the Relationship between Word Overlap and Single Word Norms When Predicting Judgments and Recall

Nicholas Pruett Maxwell
Missouri State University, Maxwell270@live.missouristate.edu

MODELING MEMORY: EXPLORING THE RELATIONSHIP BETWEEN WORD OVERLAP AND SINGLE WORD NORMS WHEN PREDICTING JUDGMENTS AND RECALL

A Masters Thesis

Presented to

The Graduate College of

Missouri State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science, Psychology

By

Nicholas P. Maxwell

May 2018

MODELING MEMORY: EXPLORING THE RELATIONSHIP BETWEEN WORD OVERLAP AND SINGLE WORD NORMS WHEN PREDICTING JUDGMENTS AND RECALL

Psychology

Missouri State University, May 2018

Master of Science

Nicholas P. Maxwell

ABSTRACT

This study examined the interactive relationship between associative, semantic, and thematic word pair strength when predicating item relatedness judgments and cued-recall performance. In Experiment One, 112 participants were shown word pairs with varied levels of associative, semantic, and thematic overlap (measured with forward strength, cosine, and latent semantic analysis) and were asked to judge how related item pairs were before taking a cued-recall test. Experiment One had four goals. First, the judgment of associative memory task (JAM) was expanded to include three types of judgments. Next, the and interaction between database norms (FSG, COS, and LSA) was for when predicting judgments and recall. Finally, JAM slopes calculated in Hypothesis One were used to predict recall. Experiment Two sought to first replicate interaction findings from Experiment One using a new set of stimuli, and second to replicate these interactions when controlling for several single word norms. Overall, Experiment One found significant three-way interactions between the network norms when predicting judgments and recall. Experiment Two partially replicated these interactions. These results suggest that associative, semantic, and thematic memory networks form a set of interdependent memory systems used for both cognitive processes.

KEYWORDS: judgments, memory, association, semantics, thematics

This abstract is approved as to form and content

_____

Erin M. Buchanan, Ph.D.
Chairperson, Advisory Committee
Missouri State University

MODELING MEMORY: EXPLORING THE RELATIONSHIP BETWEEN WORD OVERLAP AND SINGLE WORD NORMS WHEN PREDICTING JUDGMENTS AND RECALL

By

Nicholas P. Maxwell

A Masters Thesis
Submitted to the Graduate College
Of Missouri State University
In Partial Fulfillment of the Requirements
For the Degree of Master of Science, Psychology

May 2018

Approved:

_____
Erin M. Buchanan, Ph.D.

_____
Bogdan Kostic, Ph.D.

_____
David Zimmerman, Ph.D.

_____
Julie Masterson, Ph.D.: Dean, Graduate College

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Erin M. Buchanan for chairing my committee, and for providing me with so much and guidance and direction during my time at Missouri State. Next, I want to thank Dr. Bogdan Kostic and Dr. David Zimmerman for providing me with feedback and input with related to this project. Finally, I would like to thank my family for supporting me throughout my academic career. In particular, I would like to thank my wife Juliah. Your patience and unending support throughout this journey deserves recognition.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Paired-Associate Learning

The study of cognition has rich history of exploring the role of association in human memory. One of the key findings is that elements of cognitive processing play a critical role in how well an individual retains learned information. Throughout the mid-20th century, researchers investigated this notion, particularly through the use of paired-associate learning (PAL) tasks. In this paradigm, participants are presented with a pair of items and are asked to make connections between them, so that the presentation of the first item (the cue) will in turn trigger the recall of the second item (the target). Early studies of this nature focused primarily on the effects of meaning and imagery on recall performance. For example, Smythe & Paivio (1968) found that noun imagery played a crucial role in PAL performance. Subjects were much more likely to remember word-pairs that were low in similarity if imagery between the two items was high. Subsequent studies in this area focused on the effects of mediating variables on PAL tasks as well as the effects of imagery and meaningfulness on associative learning (Richardson, 1998), with modern studies shifting their focus towards a broad range of applied topics such as how PAL is affected by aging (Hertzog, Kidder, Powell-Moman, & Dunlosky, 2002), its impacts on second language acquisition (Chow, 2014), and even evolutionary psychology (Schwartz & Brothers, 2013).

Early PAL studies routinely relied on stimuli generated from word lists that focused extensively on measures of word frequency, concreteness, meaningfulness, and imagery (Paivio, 1969). However, the word pairs in these lists were typically created due

to their apparent relatedness or frequency of occurrence together in bodies of text. While lab-generated norms appear face valid, a closer inspection shows that this method lacks a decisive method of defining the underlying relationships present between item pairs (Buchanan, 2010). Furthermore, these variables only capture psycholinguistic measurements pertaining to one individual item. PAL, by nature, is used with paired items, which requires researchers to have a reliable means of investigating concept relationships. As a result, free association norms have now become a common means of indexing the shared association strength between word pairs (Nelson, McEvoy, & Schreiber, 2004).

Associations in this context refers to the context-based connections between items that is formed by frequent co-occurrence (Nelson, McEvoy, & Dennis, 2000). Often, such associations are formed by items frequently occurring together in language. For example, the terms *peanut* and *butter* have become associated over time through their joint use to depict a particular type of food, though separately, the two concepts share very little overlap in terms of meaning. To generate free association norms, participants engage in a free association task, in which they are presented with a cue word and are asked to list the first target word that comes to mind. The probability of producing a given response to a particular cue word can then be determined by dividing the number of participants producing the desired response to the cue by the total number of responses generated (Nelson et al., 2000). This method allows researchers to calculate the forward strength (FSG) of an item pair, which is a value ranging from 0 to 1 that represents the probability of the cue item eliciting the target item. Using this technique, researchers have developed databases of associative word norms that can be used to generate stimuli with a high

degree of reliability. Many of these databases are now readily available online, with the largest one consisting of over 72,000 cue-target pairs generated from more than 5,000 cue words (Nelson et al., 2004).

Semantic Networks

Similar to association norms, semantic word norms provide researchers with another option for constructing stimuli for use in tasks requiring word-pair. These norms measure the underlying concepts represented by words and allow researchers to tap into aspects of semantic memory. Semantic memory is best described as an organized collection of our general knowledge and contains information regarding a concept's meaning (Hutchison, 2003). Models of semantic memory broadly fall into one of two categories. Connectionist models (Rogers & McClelland, 2006; Rumelhart & McClelland, 1986) portray semantic memory as a system of interconnected units representing concepts, which are linked together by a series of weighted connections representing knowledge. By triggering the input units, activation then spreads throughout the system, activating or suppressing connected units based on the weighted strength of the corresponding unit connections (Jones, Willits, & Dennis, 2015). On the other hand, distributional models of semantic memory posit that semantic representations are created through the co-occurrences of words together in a body of text and suggest that words with similar meanings will appear together in similar contexts (Riordan & Jones, 2011). Popular distributional models of semantic memory include Latent Semantic Analysis (Landauer & Dumais, 1997) and the Hyperspace Analogue to Language model (Lund & Burgess, 1996).

Feature production tasks are a common means of producing semantic word norms (Buchanan, Holmes, Teasley, & Hutchison, 2013; Vinson & Vigliocco, 2008; McRae, Cree, Seidenberg, & McNorgan,, 2005). Similar in nature to the free association tasks used to generate association norms, feature production tasks present participants with the name of a concept and participants are asked to list what they believe to the concept's most important features to be (McRae et al., 2005). Several statistical measures have been developed which measure the degree of feature overlap present between concepts. Semantic similarity between any two concepts can be measured by representing the concepts as vectors in a semantic space and calculating the cosine value (COS) between them (Maki, McKinley, & Thompson, 2004). Cosine values range from 0 (unrelated) to 1 (perfectly related). For example, the item pair *hornet – wasp* has a COS of .88. indicating a high degree of shared features between the concepts. Feature overlap can also be measured by JCN, which involves calculating the information content value for each concept and the lowest super-ordinate that is shared by each concept. This is done using and online dictionary, such as WordNet (Miller, 1995). The JCN value is then computed by summing together the difference of each concept and the lowest shared super-ordinate (Maki et al., 2004; Jiang & Conrath, 1997). The advantage of using COS values over JCN values is the limitation imposed by JCN being tied to a somewhat static database, while a semantic feature production task can be used on any concept to generate COS values. However, JCN values require less time to compute if both concepts are present in the database (Buchanan et al., 2013).

Semantic relations can be broadly described as being taxonomic or the thematic in nature. Whereas taxonomic relationships focus solely on the connections between

features and concepts within categories (e.g., *bird – pigeon*), thematic relationships center around the links between concepts and an overarching theme or scenario (e.g., *bird – nest*; Jones & Galonka, 2012). Jouravlev & McRae (2016) provide a list of 100 thematic production norms, which were generated through a task similar to feature production. In their task, participants were presented with a concept and were asked to list names of other concepts they believed to be related (as opposed to being asked to respond with important features of the item). Distributional models of semantic memory also lend themselves well to the study of thematic word relations. Because these models are text based and score word pair relations in regard to their overall context within a document, they assess both semantic and thematic knowledge. Additionally, text-based models such as LSA are able to account for both the effects of context and similarity of meaning, effectively bridging the gap between associations and semantic (Landauer, Foltz, & Laham, 1998).

Comparison of Overlap Measures

Discussion of these measures then raises the question of whether each one is truly assessing some unique concept or if they simply tap into our overall linguistic knowledge. Taken at face value, word pair associations and semantic word relations appear to be vastly different, yet the line between semantics/associations and thematics is much more blurred. While thematic word relations are indeed an aspect of semantic memory and include word co-occurrence as an integral part of their creation, themes appear to be indicative of a separate area of linguistic processing. Previous research by Maki and Buchanan (2008) appears to confirm this theory. Using clustering and factor

analysis techniques, they analyzed multiple associative, semantic, and text-based measures of associative and semantic knowledge. Their findings suggest associative measures to be separate from semantic measures. Additionally, semantic information derived from lexical measures (e.g., COS and JCN) was found to be separate from measures generated from analyses of text corpora, suggesting that text-based measures may be more representative of thematic information than purely semantic information.

While it is apparent that these word relation measures are each assessing different domains of our linguistic knowledge, care must be taken when building experimental stimuli through the use of normed databases, as many word pairs overlap on multiple types of measurements, and even the early studies of semantic priming used association word norms for stimuli creation (Lucas, 2000; Meyer, Schvaneveldt, & Ruddy, 1975; Meyer & Schvaneveldt, 1971). This observation becomes strikingly apparent when one desires the creation of word pairs related only on one dimension. One particular difficulty faced by researchers comes when attempting to separate association strength from semantic feature overlap, as highly associated items tend to be semantically related as well. Additionally, a lack of association strength between two items may not necessarily be indicative of a total lack of association, as traditional norming tasks typically do not produce a large enough set of responses to capture all possible associations between items. As such, some items with weak associations will inevitably slip through the cracks (Hutchison, 2003).

Single Word Norms

In addition to measures of word overlap, the second experiment of this study

attempted to control for several types of single word norms, which measure information pertaining to various aspects of individual words. Broadly speaking, the single word norms examined in this study can be separated into one of three categories. Base values (also referred to as lexical measures) refer to norms which capture information based on a word's structure. These measures include part of speech (POS), word frequency, and the number of syllables, morphemes, and phonemes that comprise a word. Rated values refer to age of acquisition (AOA), concreteness, imageability, valence, and familiarity. Finally, norms that provide information about the connections a word shares with others based on context will be examined. This group of single word norms includes orthographic neighborhood, phonographic neighborhood, cue and target set sizes, and cosine and feature set sizes.

First, Experiment Two sought to investigate the impact of base word norms. Chief amongst these is word frequency. Several sets of norms exist for measuring the frequency with which words occur in language, and it is important to determine which of these offers the best representation of everyday language. One of the oldest and most commonly used collections of these norms are the Kučera and Francis (1967) frequency norms. These norms consist of a set of frequency values for English words, which were generated by analyzing books, magazines, and newspapers. However, the validity of using these norms has been questioned on factors such as the properties of the sources analyzed, the size of the corpus, and the overall age of these norms. First, these norms were created solely from the analysis of written text. It is important to keep in mind that stylistically, writing tends to be more formal than everyday language and as a result, it may not be the approximation of it (Brysbaert & New, 2009). Additionally, these norms

were generated fifty years ago, meaning that these norms may not accurately reflect the current state of the English language. As such, the Kučera and Francis (1967) norms, while popular, may not be the best choice for researchers interested in gauging the effects of word frequency.

Several viable alternatives to the Kučera and Francis (1967) frequency norms now exits. One popular method is to use frequency norms obtained from the HAL corpus, which consist of approximately 131 million words (Burgess & Lund, 1997; Lund & Burgess, 1996). Other collections of frequency norms include CELEX (Baaye, Piepenbrock, & Gulikers, 1995) which is based on written text, the Zeno frequency norms (Zeno, Ivens, Millard, & Duvvuri, 1995) created from American children's textbooks, and Google Book's collection of word frequencies derived from 131 billion words which were taken from books published in the United States (see Brysbaert, Keuleers, & New (2011) for an overview and comparison of these norms). The present study uses data taken from the SUBTLEX project (Brysbaert & New, 2009), which is a collection of frequency norms derived from a corpus of approximately 51 million words. This corpus was created from movie and television subtitles. SUBTLEX norms are thought to better approximate everyday language, as lines from movies and television tend to be more reflective of everyday speech than writing samples. Additionally, the larger size of the SUBTLEX corpus contributes to validity of these norms when compared the Kučera and Francis frequency norms.

In addition to word frequency, this study was also interested in testing the effects of several additional measures of lexical information that relate to the physical make-up of words. These measures include the numbers of phonemes, morphemes, and syllables

that comprise each word as well as its part of speech. The number of phonemes refers to the number of individual sounds that comprise a word (i.e., the word *cat* has three phonemes, each of which correspond to the sounds its letters make), while the term morpheme refers to the number of sound units that contain meaning. *Drive* contains one morpheme, while *driver* contains two. Morphemes typically consist of root words and their affixes. Additionally, word length (as measured by the number of individual characters a word consists of) and the number of syllables a word contains were also investigated, as previous research has suggested that the number of syllables a word contains may play a role in processing time. In general, longer words require longer processing time (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), and shorter words tend to be more easily remembered (Cowan, Baddeley, Elliott, & Norris, 2003).

Next, this study examined the effects of norms that measure word properties rated by participants. The first of these is age of acquisition, which is a measure of the average age at which a word learned. This norm is generated by presenting participants with a word and having them estimate the age (in years) at which they believe that they would have learned it (Kuperman et al., 2012). Age of acquisition ratings have been found to be predictive of recall; for example, Dewhurst, Hitch, & Barry (1998) found that recall was higher for lately acquired words. Also, of interest are measures of a word's valence, which refers to its intrinsic pleasantness or perceived positiveness (Bradley & Lang, 1999). Valence ratings are important across multiple psycholinguistic research settings, including research on emotion, the impact of emotion on lexical processing and memory, estimating the sentiments of larger passages of text, and estimating the emotional value of new words based on valence ratings of semantically similar words (Warriner, Kuperman,

& Brysbaert, 2013). The next of these rated measures is concreteness, which refers to the degree that a word relates to a perceptible, tangible object (Brysbaert, Warriner, & Kuperman, 2013). Similar to concreteness, imageability is best described as being a measure of a word's ability to generate a mental image (Stadthagen-Gonzalez & Davis, 2006). Both imageability and concreteness have been linked to recall, as items rated higher in these areas tend to be more easily remembered (Nelson & Scheiber, 1992). Finally, familiarity norms can be described as an application of word frequency, as these norms measure the frequency of exposure to a particular word (Stadthagen-Gonzalez & Davis, 2006).

The final group of norms being investigated were those which provide information based on connections with neighboring words. Phonographic neighborhood refers to the number of words that can be created by changing one sound in a word (i.e., *cat* to *kite*). Similarly, orthographic neighborhood refers to the number of words that can be created by changing a single letter in a word, such as changing *cat* to *bat*. (Adelman & Brown, 2007; Peereman & Content, 1997). Previous findings have suggested that the frequency of a target word relative to that of its orthographic neighbors has an effect on recall, increasing the likelihood of recall for that word (Carreiras, Perea, & Grainger, 1997). Additionally, both of these measures have been found to affect processing speed for items (Buchanan et al., 2013; Adelman & Brown, 2007; Coltheart, Davelaar, Jonnasson, & Besner 1977). Next, two single word norms directly related item associations were examined. These norms measure the number of associates a word shares cue or target connections with. Cue set size (QSS) refers to the number of cue words that a target word is connected to, while target set size (TSS) is a count of the

number of target words a cue word is connected to (Schreiber & Nelson, 1998). Previous research has shown evidence for a cue set size effect in which cue words that are linked to a larger number of associates (target words) are less likely to be recalled than cue words linked to fewer target words (Nelson, Schreiber, & Xu, 1999). As such, feature list sizes and cosine set sizes will be calculated for norms taken from the Buchanan et al. (2013) semantic feature norm set.

Application to Judgment Studies

Traditional judgment of learning (JOL) tasks can be viewed as an application of the PAL paradigm; participants are given pairs of items and are asked to judge how accurately they would be able to correctly match the target with the cue on a recall task. Judgments are typically made out of 100, with a participant response of 100 indicating full confidence in recall ability. In their 2005 study, Koriat and Bjork examined overconfidence in JOLs by manipulating associative overlap (measured in FSG) between word-pairs and found that subjects were more likely to overestimate recall for pairs with little or no associative relatedness. Additionally, this study found that when accounting for associative direction, subjects were more likely to overestimate recall for pairs that were high in backwards strength (BSG) but low in FSG. To account for this finding, the authors suggested that JOLs may rely more heavily on overlap between cue and target with the direction of the associative relationship being of secondary importance. Take for example the pair *feather – bird*, which has a FSG of .051 and a BSG of .359. However, this item pair also has a cosine of .272 (suggesting low to moderate feature overlap) and an LSA score of .517 (suggesting moderate thematic overlap). As such, some of the

overconfidence in JOLs may be attributed to more than just item associations. Paired items may also be connected together by similar themes or share certain features, both of which could potentially result in inflated JOLs.

JOL tasks can be manipulated to investigate perceptions of word overlap by having participants judge how related they believe the stimuli to be (Maki. 2007a; Maki, 2007b). The judgment ratings obtained from this task can then be compared to the normed databases to create a similar accuracy function or correlation as is created in JOL studies. When presented with the item pair, participants are asked to estimate the number of people out of 100 who would provide the target word when shown only the cue (Maki, 2007a), which mimics how the association word norms are created through free association tasks. Maki (2007a) investigated such judgments within the context of associative memory by having participants rate how much associative overlap was shared between items and found that responses greatly overestimated the actual overlap strength for pairs that were weak associates, while underestimating strong associates. This finding replicates the Koriat and Bjork (2005) findings for judgments on associative memory, rather than on learning.

The judgment of associative memory function (JAM) is created by plotting the judged values by the word pair's normed associative strength and calculating a fit line, which characteristically has a high intercept (representing an overconfidence bias) and a shallow slope (indicating low sensitivity to changes in relatedness strength). Figure 1 illustrates this function. Overall, the JAM function has been found to be highly reliable and generalized well across multiple variations of the study, with item characteristics such as word frequency, cue set size (QSS), and semantic similarity all having a minimal

influence on it (Maki, 2007b). Furthermore, an applied meta-analysis of more than ten studies of JAM indicated that bias and sensitivity are nearly unchangeable, often hovering between 40-60 points for the intercept and .20-.40 for the slope (Valentine & Buchanan, 2013). Additionally, Valentine & Buchanan (2013) extended this research to include judgments of semantic memory with the same results.

OVERVIEW OF EXPERIMENTS

The present study combines PAL and JAM to examine item recall within the context of items judgments, while extending Maki's JAM task to include additional judgment tasks corresponding to semantic and thematic memory. Relationship strengths between word pairs were manipulated across each of the three types of memory being investigated. Instead of focusing solely on one variable or trying to create stimuli that represented only one form of relatedness, a range of item relatedness for each variable was included to explore potential interactions.

Specifically, this research was conceptualized within the framework of a three-tiered view of the interconnections between these memory systems as it relates to processing concept information. The three-tiered view was inspired by models of reading and naming, particularly the triangle models presented by Seidenberg and McClelland (1989) and Plaut, D. C., McClelland, Seidenberg, & Patterson (1996). These models explored the nature of reading as bidirectional relations between semantics, orthography, and phonology. One goal of this research was to examine if the semantic, associative, and thematic systems are interactive for judgment and recall processes, much like the proposed interactive nature of phonology, orthographics, and semantics for reading and naming processes. Potentially, association, semantic, and thematic facets of word relations each provide a unique component that can be judged and used for memory, thus, suggesting three separate networks of independent information. This view seems unlikely, in that research indicates that there is often overlap in the information provided by each measure of word-pair relatedness. Instead, dynamic attractor networks, as

proposed by Hopfield (1982) and McLeod, Shallice, & Plaut (2000) may better represent the interplay between these representations of concepts, as these models posit a similar feedback relationship between concepts in a network. Using these models as a theoretical framework, this study sought to understand how these three types of word-pair information may interact when judgment and recall processes were applied to concept networks

Experiment One

Experiment One examined how different levels of associative overlap (measured with FSG), semantic overlap (measured with COS), and thematic overlap (measured with LSA) affect cognitive tasks such as short-term item recall and judgments of item relatedness. Four hypotheses were tested in Experiment One.

Hypothesis One. First, this study aimed to expand previous findings from Valentine & Buchanan (2013), Buchanan (2010), and Maki (2007a; 2007b) to include three types of judgments of memory in one experiment, while replicating JAM bias and sensitivity findings. The three databases norms for association, semantics, and thematics were used to predict each type of judgment and overall average slope and intercept values were calculated for each participant. It is expected that mean slope and intercept values for each type of judgment will be significantly different from zero and within the range of previous findings.

Hypothesis Two. Given the amount of overlap present between these variables, it is expected that an interaction will exist between the database norms when predicting judgments and controlling for judgment type. Multilevel modeling was used to examine

this interaction between associative, semantic, and thematic database norms in relation to participant judgments.

Hypothesis Three. The analyses were then extended to recall as the dependent variable of interest. A multilevel logistic regression was used to examine the interaction between the three database norms when predicting recall, while controlling for judgment type and rating. As with judgments, it is expected that this interaction will be significant and that judgment ratings will positively predict recall (i.e., words rated as more related will be remembered better).

Hypothesis Four. The final hypothesis tested whether judgment slopes obtained from Hypothesis One were predictive of recall. Whereas Hypothesis Three examined the direct relationship between word relatedness and recall, this hypothesis explored whether participant sensitivity to word relatedness was a predictor of recall. This analysis used a multilevel logistic regression to control for multiple judgment slope conditions.

Experiment Two

Experiment Two sought to replicate interaction findings from Experiment One with a new set of stimuli, while also expanding the analysis to control for norms measuring single word information for the item pairs used. As with the previous experiment, multilevel models were used to explore the relationships between variables. The extended analysis introduced the different types of single word norms into the analysis in a series of steps, based upon the neighborhood they belong to. Finally, single word norms were generated for the stimuli used in Experiment One. This set of stimuli as then combined with the stimuli used in Experiment Two, and judgment and recall

interaction findings were tested for using the combined stimuli set. The end goal was to determine which neighborhood of single word norms has the greatest overall impact on recall and judgment making and to further assess the impact of network connections after controlling for the various neighborhoods of single word information.

METHOD

Participants

Approval for this project was obtained from the Missouri State University Institutional Review Board (Study number IRB-FY2017-533; approved March 22, 2017; renewed February 9, 2018). First, a power analysis was conducted using the *simr* package in *R* (Green & MacLeod, 2016). This package uses simulations to calculate power for mixed linear models created from the *lme4* and *nlme* packages in *R* (Bates, Mächler, Bolker, & Walker, 2015; Pinheiro, Bates, D., Debroy, Sarkar, & R Core Team, 2017). The results of this analysis suggested a minimum of 35 participants would be required to detect an effect at 80% power. However, because power often tends to be underestimated, participant recruitment was extended within the confines of available funding (Brysbaert & Stevens, 2018). Thus, 112 participants were recruited to take part in Experiment One, and an additional 221 were recruited for Experiment Two, leading to 333 total participants in the combined data set. Participants were recruited from Amazon's Mechanical Turk, which is a website that allows individuals to host projects and connects them with a large pool of respondents who complete them for small amounts of money (Buhrmester, Kwang, & Gosling, 2011). Participant responses were screened for a basic understanding of the study's instructions. Common reasons for rejecting responses included participants entering related words when numerical judgment responses were required, responding with numerical ratings during the cued recall task, or participants responding to the cue words during recall with phrases or sentences instead of individual words. Participants who completed Experiment One correctly were compensated $1.00

for their participation, and those who completed Experiment Two correctly were compensated $2.00.

Materials

The stimuli used in Experiment One were 63 words pairs of varying associative, semantic, and thematic relatedness which were created from the Buchanan et al. (2013) word norm database and website. Associative relatedness was measured with forward strength (FSG), which is the probability that a cue word will elicit a desired target word (Nelson et al., 2004). This variable ranges from zero to one, with zero being indicative of no association between pairs, while a rating of one indicates that participants would always give the target word in response to the cue. Semantic relatedness was measured with cosine (COS), which is a measure of semantic feature overlap (Buchanan et al., 2013; Vinson & Vigliocco, 2008; McRae et al., 2005). This variable ranges from zero to one wherein zero indicates no shared semantic features between concepts, and higher numbers indicate more shared features between concepts. Finally, thematic relatedness was calculated with Latent Semantic Analysis (LSA), which generates a score based upon the co-occurrences of words within a document (Landauer et al., 1998; Landauer & Dumais, 1997) LSA values also range from zero to one, with zero indicating no co-occurrence and higher values representing higher co-occurrence. These values were chosen to represent these categories based on face validity and previous research on how word pair variables overlap (Maki & Buchanan, 2008).

Experiment two followed this same design and used an additional 63 word-pairs which were created in the same manor using the Buchanan et al. (2013) norms. Single

word norm information was also obtained for each cue and target item. Word frequency was collected from the SUBTLEX project (Brysbaert & New, 2009). Part of speech, word length, and the number of morphemes, phonemes, and syllables of each item were derived from the Buchanan et al. (2013) word norms (originally collected as part of the English Lexicon Project, Balota, Yap, Hutchison, Cortese, Kessler, Loftis, Treiman, 2007) For items with multiple parts of speech (for example, *drink* can refer to both a beverage and the act of drinking a beverage), the most commonly used tense of the word was used. Following the design of Buchanan et al. (2013), this was determined using Google's define feature. Concreteness, cue set size (QSS), and target set size (TSS) were taken from the South Florida Free Association Norms (Nelson et al., 2004). Feature set size (FSS, i.e., the number of features listed as part of the definition of a concept) and cosine set size (COSC, i.e., the number of semantically related words above a cosine of zero) were calculated from Buchanan et al. 2013. Imageability and familiarity norms were taken from the Toglia and colleagues set of semantic word norms (Toglia, 2009; Toglia & Battig, 1978). Age of acquisition (AOA) ratings were pulled from the Kuperman et al. (2012) database. Finally, valence ratings for all items were obtained from the Warriner et al. (2013) norms.

Because information about single word norms was collected during the data creation process, one limitation is that the item pairs created were constrained to only those items which appeared across all word norm databases used. To control for this, single word information was collected post-hoc for the stimuli used in Experiment One for items appearing in the various databases. The two datasets were then merged to create

a combined dataset, which was used for the single word analyses in Experiment Two. Tables 1 and 2 display descriptive statistics for this combined dataset.

Each experiment arranged stimuli pairs into three item blocks, with each block consisting of 21 word-pairs. Blocks were structured to have seven words of low COS (0 - 0.33), medium COS (.34 - .66), and high COS (.67 - 1). COS was chosen due to limitations with the size of the available data across all norm sets. However, the result of this selection process was that values for the remaining network norms (FSG and LSA) were contingent upon the COS strengths of the selected stimuli. To counter this, stimuli were selected at random based on the different COS groupings so as to cover a broader range of FSG, LSA, and single word norm values. Table 3 shows stimuli information for word pair norms from Experiment One, and Table 4 displays this information from Experiment Two. The studies were built online using Qualtrics, and each experiment used three surveys that were created to counter-balance the order in which judgment blocks appeared. Each word pair appeared counter-balanced across each judgment block, and stimuli were randomized within blocks. This process of counter-balancing resulted in each stimuli pair receiving a judgment for each of the three types of memory being investigated.

Procedure

Both experiments followed the same procedure, with each one divided into three phases. In the first section, participants were presented with word pairs and were asked to make judgments of how related they believed the words in each pair to be. This judgment phase consisted of three blocks of 21 word-pairs which corresponded to one of three

types of word pair relationships: associative, semantic, or thematic. Each block was preceded by a set of instructions explaining one of the three types of relationships, and participants were provided with examples which illustrated the type of relationship to be judged. The associative block began by explaining associative memory and the role of free association tasks. Participants were provided with examples of both strong and weak associates. For example, *lost* and *found* were presented as an example of a strongly associated pair, while *article* was paired with *newspaper*, *the*, and *clothing* to illustrate that words can have many weak associates. The semantic judgment block provided participants with a brief overview of how words are related by meaning and showed examples of concepts with both high and low feature overlap. *Tortoise* and *turtle* were provided as an example of two concepts with significant overlap. Other examples were then provided to illustrate concepts with little or no overlap. For the thematic judgments, participants were provided with an explanation of thematic relatedness. *Tree* is explained to be related to *leaf*, *fruit*, and *branch*, but not *computer*. Participants were then given three concepts (*lost*, *old*, *article*) and were asked to generate words that they felt were thematically related to each concept. Complete instructions for each judgment condition are available in the appendix.

Judgment instructions for each block were contingent on the type of judgment being elicited. For example, instructions in the associative block asked participants to estimate how many college students out of 100 would respond to the cue word with the given target, while instructions for the semantic judgments asked participants to indicate the percent of features shared between two concepts. All judgment instructions were modeled after Buchanan (2010) and Valentine & Buchanan (2013).

Participants then rated the relatedness of the word pairs based on the set of instructions they received. In accordance with previous work on JOLs and JAM, item judgments were made using a scale of zero to 100, with zero indicating no relationship, and 100 indicating a perfect relationship. Participants typed their responses into the survey. After finishing the first block, participants then completed the remaining judgment blocks in the same manner. Each subsequent judgment block changed the type of judgment being made. Three versions of the study were created, with counter-balanced the order in which judgment blocks appeared. Participants were randomly assigned to a survey version. This resulted in each word-pair receiving judgments for each of the three types of relationships. This study design was used for both experiments.

After completing the judgment blocks, participants were presented with a short distractor task to account for recency effects. In the section, participants were presented with a randomized list of the 50 U.S. states and were asked to arrange them in alphabetical order. This task was timed to last two minutes. Once time had elapsed, participants automatically progressed to the final section, which consisted of a cued-recall task. In this section, participants were presented with each of the 63 cue words from the judgment phase and were asked to complete each word-pair by responding with the correct target word. This task was not timed, and participants were informed that they would incur no penalties for guessing. This task presented stimuli in a randomized order.

RESULTS

Experiment One

Data Processing and Descriptive Statistics. First, the recall portion of the study was coded as zero for incorrect responses, one for correct responses, and NA for participants who did not complete the recall section (i.e., all or nearly all responses were blank). Additionally, all word responses to judgment items were deleted and set as missing data. The final dataset was created by splitting the initial data file into six sections (one for each of the three experimental blocks and their corresponding recall scores). Each section was individually melted using the *reshape* package in *R* (Wickham, 2007) and was written as a csv file. The six output files were then combined to create the final dataset. With 112 participants, the dataset in long format contained 7,056 rows of data (i.e., 112 participants * 63 judgments). One incorrect judgment data point which was outside the range of the scale (> 100) was corrected to NA. Missing data points for judgments or recall were then excluded from the analysis, which included word responses to judgment items (i.e., responding with *cat* instead of a numerical rating). These types of responses excluded participants from receiving Amazon Mechanical Turk payment. In total, 787 data points were excluded from this analysis (188 judgment only, 279 recall only, and 320 across both judgments and recall), leading to a final *N* of 105 participants and 6,269 observations. Recall and judgment scores were then screened for outliers using Mahalanobis distance at $p < .001$, and no outliers were found (Tabachnick & Fidell,

2007). To screen for multicollinearity, correlations between judgment items, COS, LSA, and FSG were examined, and *r* for all correlations was found to be < .50.

The mean judgment of memory for the associative condition (*M* = 58.74, *SD* = 30.28) was lower than the semantic (*M* = 66.98, *SD* = 28.31) and thematic (*M* = 71.96, *SD* = 27.80) judgment conditions. Recall averaged over 60% for all three conditions: associative *M* = 63.40, *SD* = 48.18; semantic *M* = 68.02, *SD* = 46.65; thematic *M* = 64.89, *SD* = 47.74.

Hypothesis One. The first hypothesis sought to replicate bias and sensitivity findings from previous research while also expanding the JAM function to include judgments based on three types of memory. FSG, COS, and LSA were used to predict each type of judgment. Judgment values were divided by 100, so as to place them on the same scale as the database norms. Slopes and intercepts were then calculated for each participant's ratings for each of the three judgment conditions, as long as they contained at least nine data points out of the 21 that were possible. Single sample *t*-tests were then conducted to test whether slope and intercept values were significantly different from zero. The results of these tests are reported in Table 5. Slopes were then compared to the JAM function, which is characterized by high intercepts and shallow slopes. Because the scaling of the data, to replicate this function, intercepts should range from .40 to .60, and slopes should be in the range of .20 to .40. Intercepts for associative, semantic, and thematic judgments were each significant, and all fell within or near the expected range. Thematic judgments had the highest intercept at .656, while associative judgments had the lowest intercept at .511.

The JAM slope was successfully replicated for FSG in the associative judgment condition, with FSG significantly predicting association, although the slope was slightly higher than expected at .491. COS and LSA did not significantly predict association. For semantic judgments, each of the three database norms were significant predictors. However, JAM slopes were not replicated for this judgment type, as FSG had the highest slope at .118, followed by LSA at .085, and COS at .059. These findings were mirrored for thematic judgments, as each database norm was a significant predictor, yet slopes for each predictor fell below the range of the expected JAM slopes. Again, FSG had the highest slope, this time just out of range at .192, followed closely by LSA at .188, Interestingly, COS slopes were found to be negative for this judgment condition. Overall, although JAM slopes were not perfectly replicated within each judgment type, the high intercepts and shallow slopes present in all three conditions are still indicative of overconfidence and insensitivity in participant judgments.

Hypothesis Two. As a result of the overlap between variables in Hypothesis One, the goal of the second hypothesis was to test for an interaction between the three database norms when predicting participant judgment ratings. First, database norms were mean centered to control for multicollinearity. The *nlme* package and *lme* function were used to calculate these analyses (Pinheiro et al., 2017). A maximum likelihood multilevel model was used to test the interaction between FSG, COS, and LSA when predicting judgment ratings while controlling for judgment type, with participant number being used as the random intercept factor. Multilevel models were used to retain all data points (rather than averaging over items and conditions, while controlling for correlated error due to participants, as the models are advantageous for multiway repeated measures designs

(Gelman, 2006). This analysis resulted in a significant three-way interaction between FSG, COS, and LSA ($b = 3.324$, $p < .001$), which was then examined through simple slopes analysis. Table 6 shows values for main effects, two-way, and three-way interactions.

To investigate this interaction, simple slopes were calculated for low, average, and high levels of COS. This variable was chosen because manipulating COS made it possible to track changes across FSG and LSA. Significant two-way interactions were found between FSG and LSA at both low COS ($b = -1.492$, $p < .001$), average COS ($b = .569$, $p < .001$), and high COS ($b = .355$, $p = .013$). A second level was then added to the analysis in which simple slopes were created for each level of LSA, allowing us to assess the effects of LSA at different levels of COS on FSG. When both COS and LSA were low, FSG significantly predicted judgment ratings ($b = .663$, $p < .001$). At low COS and average LSA, FSG decreased but still significantly predicted judgment ratings ($b = .375$, $p < .001$). However, when COS was low and LSA was high, FSG was not a significant predictor ($b = .087$, $p = .079$). A similar set of results was found at the average COS level. When COS was average and LSA was LOW, FSG was a significant predictor, ($b = .381$, $p < .001$). As LSA increased at average COS levels, FSG decreased in strength: average COS, average LSA FSG ($b = .355$, $p = .013$) and average COS, high LSA FSG ($b = .161$, $p < .001$). This finding suggests that at low COS, LSA and FSG create a seesaw effect in which increasing levels of thematics is counterbalanced by decreasing importance of association when predicting judgments. FSG was not a significant predictor when COS was high and LSA was low ($b = .099$, $p = .088$). At high COS and average LSA, FSG significantly predicted judgment ratings ($b = .167$, $p < .001$), and

finally when both COS and LSA were high, FSG increased and was a significant predictor of judgment ratings (b = .236, p < .001). Thus, at high levels of COS, FSG and LSA are complementary when predicting recall, increasing together as COS increases. Figure 2 displays the three-way interaction wherein the top row of figures indicates the seesaw effect, as LSA increases FSG decreases in strength. The bottom row indicates the complementary effect where increases in LSA occur with increases in FSG predictor strength.

Hypothesis Three. Given the results of Hypothesis Two, this next hypothesis sought to extend the analysis to participant recall scores. A multilevel logistic regression was used with the *lme4* package and *glmer()* function (Pinheiro et al., 2017), testing the interaction between FSG, COS, and LSA when predicting participant recall. As with the previous hypothesis, type of judgment was controlled for, as well as covaried judgment ratings. Participants were used as a random intercept factor. Judged values were a significant predictor of recall, (b = .686, p < .001) where increases in judged strength predicted increases in recall. A significant three-way interaction was detected between FSG, COS, and LSA (b = 24.572, p < .001). See Table 7 for main effects, two-way, and three-way interaction values.

The moderation process from Hypothesis Two was then repeated, with simple slopes first calculated at low, average, and high levels of COS. This set of analyses resulted in significant two-way interactions between LSA and FSG at low COS (b = -7.845, p < .001) and high COS (b = 5.811, p = .009). No significant two-way interaction was found at average COS (b = -1.017, p = .493). Next, simple slopes were then calculated for low, average, and high levels of LSA at the low and high levels of COS, so

as to assess how FSG effects recall at varying levels of both COS and LSA. When both COS and LSA were low, FSG was a significant predictor of recall ($b = 4.116$, $p < .001$). At low COS and average LSA, FSG decreased from both low levels, but was still a significant predictor ($b = 2.601$, $p < .001$), and finally, low COS and high LSA, FSG was the weakest predictor of the three ($b = 1.086$, $p = .030$). Figure 3 displays the three-way interaction. As with Hypothesis Two, LSA and FSG counterbalanced one another at low COS, wherein the increasing levels of thematics led to a decrease in the importance of association in predicting recall. At high COS and low LSA, FSG was a significant predictor ($b = 2.447$, $p = 0.003$). When COS was high and LSA was average, FSG increased as a predictor and remained significant ($b = 3.569$, $p < .001$). This finding repeated when both COS and LSA were high, with FSG increasing as a predictor of recall ($b = 4.692$, $p < .001$). Therefore, at high levels of COS, LSA and FSG are complementary predictors of recall, increasing together and extending the findings of Hypothesis Two to participant recall. The top left figure indicates the counterbalancing effect of recall of LSA and FSG, while the top right figure shows no differences in simple slopes for average levels of cosine. The bottom left figure indicates the complementary effects where LSA and FSG increase together as predictors of recall at high COS levels.

Hypothesis Four. The final hypothesis in Experiment One investigated whether the judgment slopes and intercepts obtained in Hypothesis One would be predictive of recall ability. Whereas Hypothesis Three indicated that word relatedness was directly related to recall performance, this hypothesis instead looked at whether or not participants' sensitivity and bias to word relatedness could be used a predictor of recall (Maki, 2007b). This analysis was conducted with a multilevel logistic regression, as

described in Hypothesis Three where each database slope and intercept were used as predictors of recall using participant as a random intercept factor. These analyses were separated by judgment type, so that each set of judgment slopes and intercepts were used to predict recall. The separation controlled for the number of variables in the equation, as all slopes and intercepts would have resulted in overfitting. These values were obtained from Hypothesis One where each participant's individual slopes and intercepts were calculated for associative, semantic, and thematic judgment conditions. Table 8 displays the regression coefficients and statistics. In the associative condition, FSG slope significantly predicted recall (b = .898, p = .008), while COS slope (b = .314, p = .568) and LSA slope (b = .501, p = .279) were non-significant. In the semantic condition, COS slope (b = 2.039, p < .001) and LSA slope (b = 1.061, p = .020) were both found to be significant predictors of recall. FSG slope was non-significant in this condition (b = .381, p = .187). Finally, no predictors were significant in the thematic condition, though LSA slope was found to be the strongest (b = .896, p = .090)

Exploratory Analysis. Finally, an analysis was conducted to test whether interaction findings from Hypotheses Two and Three were influenced by either practice effects from completing multiple judgment blocks in succession or by interference from the different types of judgment instructions (i.e., completing the judgment task for a block using the previous block's set of instructions). To investigate this potential order effect, a new set of multilevel models were created which tested for interaction findings between database norms when predicting judgments and recall only corresponding to data from the first judgment block in the study. The models used were identical to ones used in Hypotheses Two and Three in every other way. Overall, significant three-way

interactions were found between COS, FSG, and LSA when predicting judgments (b = 4.040, $p < .001$) and recall (b = 22.685, p = .028). Figure 4 displays interaction findings for judgments, and Figure 5 displays findings for recall.

The simple slopes analyses conducted in previous hypotheses were then repeated. Simple slopes were calculated for low, average, and high COS for both the judgment model and the recall model. Overall, this set of analyses yielded similar results to those found in Hypotheses Two and Three. As found previously, LSA and FSG counterbalanced one another when semantics were low, wherein the increasing levels of thematics in turn led to a decrease in the importance of association in predicting judgments and recall. This trend reversed with increases in semantics. At high levels of semantics, LSA and FSG complimented one another, increasing together. The replication of these findings from Hypotheses Two and Three suggested that the multiple judgment instructions used in the previous hypotheses did not have an adverse effect on the reliability of participant judgments or recall scores obtained in the full experiment.

Experiment Two

Whereas Experiment One primarily focused on the effects of associative, semantic, and thematic database norms in the prediction of judgments and recall, Experiment Two focused on the effects of single word norms. Experiment Two first sought to replicate findings from Experiment One when using a novel set of stimuli pairs, and second, to test whether the interaction would replicate when controlling for single word norms and assess which single word norms were most predictive of these cognitive processes.

Data Processing and Descriptive Statistics. Data processing for Experiment Two followed the same procedure used in Experiment One. Recall was coded as zero for incorrect responses, one for correct responses, and NA for participants who left either all or the majority of recall responses incomplete. All word responses to judgment items were deleted and set to missing data, as well as numerical rating responses on the cued-recall task. The final dataset was created by splitting the initial data file into six sections (corresponding to each of three experimental blocks and their respective recall sections) and individually melting each section using the *reshape* package in *R* (Wickham, 2007). Melted files were then written as csv files and combined to create the final dataset.

In long format, the dataset for Experiment Two contained 13,923 rows of data (221 participants * 63 judgments). Data screening followed the same process used in Experiment One. Nine judgment data points were set to NA as they fell outside the range of the scale (> 100). Missing data points for judgments and recall were then excluded. As before, this also included word responses to judgments and numerical responses to recall. Participants whose data was excluded because they failed to follow instructions did not receive payment on Amazon's Mechanical Turk. 1,472 data points were excluded from the final analysis (833 from judgment only, 393 from recall only, and 246 across both), leading to a total of 12,451 observations from 211 participants in the final dataset. Recall and judgment scores were then screened for outliers using Mahalanobis distance at $p <$ .001, and five outliers were detected. Thus, after removing outliers, 12,446 data points remained in the final data set. Finally, multicollinearity was screened for by checking the correlations between network norms and single word norms. Because of high correlations between the various lexical measures representing word length (number of characters,

syllables, morphemes, and phonemes, $r > .75$), only number of individual characters was included in the analysis to represent word length.

As found in Experiment One, the mean judgment of memory in the associative condition ($M = 59.67$, $SD = 30.28$) was lower than in the semantic ($M = 63.33$, $SD = 30.63$) and thematic ($M = 68.97$, $SD = 28.25$) judgment conditions. Additionally, recall averaged lower than Experiment One, being at or slightly below 60% for all three conditions: associative $M = 58.05$, $SD = 49.35$; semantic $M = 60.52$, $SD = 48.85$; thematic $M = 58.51$, $SD = 49.28$.

When examining the effects of single word norms, the initial dataset for Experiment Two was combined with the dataset used in Experiment One, which had been updated to contain information corresponding to each of the single word norms being investigated. This combined dataset contained a total of 18,713 data points collected across 316 participants (after excluding participants in data screening). This dataset was used for the analyses investigating the effects of single word norms.

Replication of Interaction Findings. First, analyses were conducted to test whether interactions between database norms would replicate with the new stimuli set. These analyses mimicked the design from Hypotheses Two and Three from the first experiment. All database norms were mean centered.

Judgments. The *nlme* package in *R* was used to create a maximum likelihood multilevel model to test for an interaction between FSG, COS, and LSA when predicting judgment scores (Pinheiro et al., 2017). Although the interaction was not significant, the main effects of FSG and LSA were still significant. Table 9 displays main effects and interactions. Consistent with previous findings, FSG was the strongest predictor of

judgments (b = .422, p < .001). Although the interaction was not significant, simple slopes were still calculated to assess the underlying relationship between FSG and LSA at each level of COS to see if it displayed a pattern similar to that found in Experiment One. Figure 6 displays this relationship. FSG became weaker with each increase of LSA strength at each of the three levels of semantic overlap; thus, only the competitive relationship between the two database norms replicated for this analysis.

Recall. The *lme4* package was then used to create a multilevel logistic regression (Pinheiro et al., 2017), which tested whether the interaction found between the database norms when predicting recall would replicate with the new stimuli set. Participants were used as a random intercept factor, and judgment scores and type of judgment being made were controlled for. Overall, a significant three-way interaction was detected between FSG, COS, and LSA (b = -22.572, p < .001). This was a partial replication, as this interaction was in the opposite direction as the one found in Experiment One. Table 10 reports main effects, two-way, and three-way interaction values. As with the previous Experiment, simple slopes were then calculated for low, average, and high levels of LSA at the low and high levels of COS, so as to assess how FSG affected recall at varying levels of both COS and LSA. In line with findings from the previous experiment, these analyses yielded significant two-way interactions between LSA and FSG at low COS (b = 5.590, p = .013) and high COS (b = -7.514, p < .001), with no significant two-way interaction being found at average COS (b = -.962, p = .489). Staying consistent with the process used in Experiment One, a second set of simple slopes were then calculated for low, average, and high levels of LSA at the low and high levels of COS, so as to assess how FSG affected recall at varying levels of both COS and LSA. In contrast to previous

findings, when both COS and LSA were low, FSG did not predict recall (b = .087, p = .881). At low COS and average LSA, FSG increased in strength, and became a significant predictor (b = 1.213, p < .001). Finally, at low COS and high LSA, FSG increased further as a predictor (b = 2.339, p < .001). The observed interaction followed a trend opposite of that in Experiment One. Instead of the competitive relationship observed previously for low COS, LSA and FSG were complimentary and increased together. As COS increased FSG and LSA became competitive, which was the opposite of Experiment One. As such, at high COS and low LSA, FSG was a significant predictor (b = 3.900, p < .001). FSG weakened when LSA increased to average levels, (b = 1.606, p < .001), and continued to weaken further when both COS and LSA were high, with FSG decreasing further as a predictor of recall (b = .872, p < .001). Figure 7 displays simple slopes graphs for the three-way interaction when predicting recall. The bottom left figure indicates the counterbalancing effect of high COS levels of LSA and FSG, while the top left figure displays the complementary effects where LSA and FSG increased together as predictors of recall at low COS levels.

Extension to Single Word Norms. The final group of analyses examined the effects of single word norms on recall and judgments and whether interaction findings from Experiment One would replicate after controlling for single word norms. These analyses were conducted using an expanded dataset which combined data collected across both experiments.

Single word norms were placed into one of three categories. Frequency (measured with SUBTLEX) and word length were used as measures of lexical information. Age of acquisition, valence, familiarity, concreteness, and imageability were classified as rated

properties. Orthographic and phonographic neighborhoods, cue and target set sizes, feature set size, and cosine connectedness were grouped together as neighborhood connections.

Because of the large number of predictor variables being examined, stepwise regressions were initially performed on each category of single word norms to select the best predictors within each category. Stepwise regression enables researchers to select the best combination of independent variables for predicting the dependent variable; thus, some predictor variables may be dropped and not incorporated into the final model (Tabachnick & Fidell, 2007). Two models were created per category of single word norms, each corresponding to one of the dependent variables being investigated. Stepwise analyses were conducted using the *MASS* package in *R* (Venables & Ripley, 2002). Table 11 shows the final set of single word predictor variables retained from the stepwise analyses. When predicting judgments, the majority of the variables were retained across all models with the exception of part of speech for cue and targets, cosine connectedness for cue items, orthographic neighborhood for cue items, and phonographic neighborhood for target items. When predicting recall, cue and target part of speech, imageability for target items, concreteness for cue items, and cosine connectedness for cue and target items were excluded.

Judgments. Next, multilevel modeling was used to investigate whether interaction findings from Experiment One would replicate after controlling for each of the single word norms selected via stepwise analyses. This analysis was conducted hierarchically, with single word norms entered in to the model through a series of steps. Each step corresponded to one of the categories of single word norms, with each model using

judgment scores as the dependent variable of interest and controlling for the type of judgment being made. Marginal and Conditional R^2 values (R^2_m and R^2_c respectively) were calculated at each step of the judgement model using the *MuMIn* package (Barton, 2018). Marginal R^2 describes the proportion of variance that is explained solely by the fixed factors in the model, while the conditional R^2 value is used to describe the proportion of the variance that can be explained by both fixed and random factors (Lefcheck, 2013).

Model one examined the lexical properties of words (R^2_m = .027, R^2_c = .194). The second model added rated word properties words (R^2_m = .054, R^2_c = .220), and the third model added in neighborhood connections (R^2_m = .068, R^2_c = .232). Network norms and the three-way interaction between them were entered into the analysis in the fourth and final model (R^2_m = .118, R^2_c = .283).

Table 12 displays main effects and interaction findings for all variables in the step they were entered to control for table size. The main investigation focused on the fourth and final step of the model with the network interaction. The main effects of each individual single word norm are not discussed, however, of notable interest is the way in which several single word predictors tended to balance out across cue and target items. This finding occurred when either the cue or target version of a particular single word norm predictor showed a positive relationship, while the other displayed a negative relationship. Several cue-target predictor pairs displaying this trend were found at each step of the model. Pairs following this trend included frequency (cue b = .014, p < .001; target b = -.032, p < .001), age of acquisition (cue b = .015, p < .001; target b = -.014, p < .001), and feature set size (cue b = .001, p < .001, target b = -.001, p < .001). Therefore,

even though it appeared that many features related to single words were significant predictors of judgments, the related cue and target information often canceled each other out in strength. Consistent with previous judgment models, FSG was found in the final step to be the strongest overall predictor of judgments (b = .391, p < .001). The three-way interaction between network norms was not significant (b = .558, p = .099). To explore potential differences in effects, simple slopes were calculated using the same process as before. Figure 8 displays these findings. FSG and LSA strength were competitive at all levels of COS, with increases in thematic strength decreasing the overall predictiveness of association strength. These results matched the replication portion of this experiment, indicating that FSG and LSA competition findings still hold, even after controlling for other concept information that is activated when reading in the lexical network.

Recall. Finally, the previous set of analyses was repeated using recall as the dependent variable. A multilevel logistic regression was used, and the hierarchical design used to investigate judgments was mimicked. In addition to controlling for the type of judgments being elicited, these models also controlled for participant judgment ratings. Model steps corresponded to those used for investigating judgments. Marginal and conditional R^2 values were calculated using the *piecewiseSEM* package in *R* (Lefcheck, 2016). Lexical properties were entered into the first step (R^2_m = .026, R^2_c = .282), step two added rated word properties words (R^2_m = .052, R^2_c = .331), step three added in neighborhood connections (R^2_m = .062, R^2_c = .340), and the step four added network norms and the three-way interaction between (R^2_m = .082, R^2_c = .363)

As with the judgment analysis, several single word norms appeared to balance out one another across cue and target items. For example, frequency (cue b = -.258, p < .001;

target b = .082, p = .006), length (cue b = .138; p < .001, target b = -.047. p < .001), and feature set size. (cue b = -.012, p < .001; target b = .015, p < .001) all displayed relationships of this nature. When examining the fourth step, FSG was the strongest overall predictor of recall (b = 1.866, p < .001), and a significant three-way interaction was detected between FSG, COS, and LSA. See Table 13 for a complete list of main effects and interaction findings.

Finally, simple slopes were calculated using the same process utilized in the previous analyses to examine the three-way interaction between network norms when predicting recall. Replicating findings from the first section of Experiment Two, FSG and LSA were competitive at high COS and complimentary at low COS. Once again, this stands as a partial replication of findings from Experiment One. As with the initial replication model that did not include single word norms, the interaction present in this model is in the opposite direction as the one found in Experiment One. However, as seen with judgments, the interactive effects continued to be found even when controlling for other lexical variables. In the final section of Experiment Two, FSG and LSA were competitive at high COS and complimentary at low COS. Once again, this stands as a partial replication of findings from Experiment One. Figure 9 illustrates these findings.

DISCUSSION

Experiment One Summary

Experiment One investigated the relationship between associative, semantic, and thematic word relations and their effect on participant judgments and recall performance through the testing of four hypotheses. In Hypothesis One, bias and sensitivity findings first proposed by Maki (2007a) were successfully replicated in the associative condition, with slope and intercept values falling within the expected range. While these findings were not fully replicated when extending the analysis to include semantic and thematic judgments (as the slopes in these conditions did not fall within the appropriate range), participants still displayed high intercepts and shallow slopes, suggesting some degree of overconfidence in judgment making and an insensitivity to changes in strength between pairs.

When looking at the frequency that each predictor was the strongest in making these judgments, FSG was the strongest predictor for both the associative and semantic conditions, while LSA was the best predictor in the thematic condition. In each of the three conditions, COS was the weakest predictor, even when participants were asked to make semantic judgments. This finding suggests that associative relationships seem to take precedence over semantic relationships when judging pair relatedness, regardless of what type of judgment is being elicited. Additionally, this finding may be taken as further evidence of a separation between associative information and semantic information, in which associative information is always processed, while semantic information may be suppressed due to task demands (Buchanan, 2010; Hutchison & Bosco, 2007).

Hypothesis Two examined the three-way interaction between FSG, COS, and LSA when predicting participant judgments. At low semantic overlap, a seesaw effect was found in which increases in thematic strength led to decreases in associative predictiveness. This finding was then replicated in Hypothesis Three when extending the analysis to predict recall. By limiting the semantic relationships between pairs, an increased importance is placed on the role of associations and thematics when making judgments or retrieving pairs. In such cases, increasing the amount of thematic overlap between pairs results in thematic relationships taking precedent over associative relationships. However, when semantic overlap was high, a complementary relationship was found in which increases in thematic strength in turn led to increases in the strength of FSG as a predictor. This result suggests that at high semantic overlap, associations and thematic relations build upon one another. Because thematics is tied to both semantic overlap and item associations, the presence of strong thematic relationships between pairs during conditions of high semantic overlap boosts the predictive ability of associative word norms. Again, this complementary effect was found when examining both recall and judgments.

Finally, Hypothesis Four used the judgment slopes and intercepts calculated in Hypothesis One to investigate if participants' bias and sensitivity to word relatedness could be used to predict recall. For the associative condition, the FSG slope significantly predicted recall. In the semantic condition, recall was significantly predicted by both the COS and LSA slopes. However, for the thematic condition, although the LSA slope was the strongest, no predictors were significant. One explanation for this finding is that thematic relationships between item pairs act as a blend between associations and

semantics. As such, LSA faces increased competition from the associative and semantic database norms when predicting recall in this manner.

Experiment Two Summary

Experiment Two aimed to replicate interaction findings from the first experiment, first when using a novel set of stimuli and then when controlling for single word norms. First, when attempting to replicate the original interactions using the new set of stimuli, the three-way interaction was not significant when predicting participant judgments. Although the three-way interaction was not significant, a simple slopes analysis showed that FSG and LSA strengths were competitive with another at each level of semantic overlap. When extending this initial replication to predict recall, a significant three-way interaction was detected between the network norms. However, this interaction was in the opposite direction as the original findings from the first experiment, as FSG and LSA strength were found to be complimentary at low levels of semantics and became increasingly competitive at higher levels.

Similar trends were then found when attempting to replicate these interactions while controlling for the single word norms. No significant three-way interaction was detected when predicting judgment scores. Simple slopes analyses showed that increasing thematic overlap between pairs decreased the predictiveness of FSG at all levels of semantic overlap (i.e., competition at all levels). When recall was examined as the dependent variable of interest, the three-way interaction between network norms was significant. Again, the direction of this interaction was opposite to that found in Experiment One, which was consistent with the previous interaction. Simple slopes

analyses revealed associative and thematic overlap were complimentary to one another at low levels of semantic overlap and became increasingly competitive as semantic overlap increased. Overall, this set of replication analyses were only partially successful, which result may be due to several limitations with the available normed databases used to select the stimuli. This is discussed in further detail at the end of this section.

General Discussion

Overall, these findings shed some light on the degree to which the processing of associative, semantic, and thematic information impacts retrieval and judgment making tasks and the interactive relationship that exists between these three types of lexical information. While previous research has shown that memory networks are divided into separate systems which handle storage and processing for meaning and association, the presence of these interactions suggests that connections exist between these networks, linking them to one another. One interpretation is that these memory systems may form a three-tiered, interconnected system. First, information enters the semantic memory network, which processes features of concepts and provides a means of categorizing items based on the similarity of their features. Next, the associative network adds information for items based on contexts generated by reading or speech. Finally, the thematic network pulls in information from both the semantic and associative networks to create a mental representation of both the item and its place world relative to other concepts.

This study did not explore the timing of information input from each of these systems, but it may be similar to a dual-route model of reading and naming, in that each

runs in parallel when contributing to the judgment and recall process (Coltheart, Curtis, Atkins, & Haller, 1993).Viewing this model purely through the lens of semantic memory, it draws comparison to dynamic attractor models (Jones et al., 2015; McLeod et al., 2000; Hopfield, 1982). One of the defining features of dynamic attractor models is that they allow for some type of bidirectionally or feedback between connections in the network. In the study of semantic memory, these models are useful for taking into account multiple restraints such as links between semantics and the orthography of the concept in question.

This study extends this notion as a means of framing how these three memory systems are connected. The underlying meaning of a concept is linked with both information pertaining to its co-occurrences in everyday language and information relating to the general contexts in which it typically appears. How then does this hypothesis lend itself towards the broader context of psycholinguistic research? One application of this hypothesis may be models of word recognition. One popular model is Seidenberg and McClelland (1989) "triangle model", and several variations of this model have been proposed and tested (see Harley, 2008 for a review). This model recognizes speech and reading based upon the orthography, phonology, and meaning of words. Each of these three word properties are linked in such a way that orthography is linked to phonology, phonology is linked with meaning, and meaning is linked to orthography (forming a triangle). The pathways between word properties are bidirectional, allowing for feedback between connections. Clearly, these facets are important to consider, as this study indicated that many of the phonological and orthographic variables were significantly related to judgments and recall. The bidirectional pathways may explain why cue and targets have balanced contributions to judgments and recall, as each

contributes a small component to the final output from the participant. As both cue and target are activated in memory, the networks for these concepts are also activated, and each appears to be correspondingly weighted, potentially indicating that the focus of attention was spread across cue and target and these were weighted evenly.

Whereas the original version of this model focused almost exclusively on the link between orthography and phonology, Harm and Seidenberg (2004) developed a version which included a focus on semantics, with word meaning being based on input from the orthography and phonology components of the model. The results from this study indicated that associations and thematics should also be defined more clearly, rather than all incorporated into a semantic network (Maki & Buchanan, 2008). Future studies could examine how these networks and connections separate, to further distinguish how they are structured in memory. This set of experiments indicated that the relation between these values, when activated by judgment and memory processes, was often competitive. This finding may indicate separate networks that compete for attention when completing cognitive tasks. However, these findings may also support a race style model, as often described when studying reading. Each separate connection may be activated in parallel, but the weight given to each component will depend on the strength of activation of competing information. Ultimately, further studies will be needed to explore the interconnections between the semantic, thematic, and associative networks.

Limitations

The results of this study should be considered with the following limitations in mind. First, in Experiment One, pairs were randomly selected based on their cosine

values, with 21 low, medium, and high pairs each being selected. This method was selected to get a range of values across FSG and LSA. However, the set of norms used to generate the stimuli pairs contained a disproportionate number of pairs low in FSG and LSA strength compared to medium or high strength pairs. For example, the database contained a total 356 high COS pairs, of which 326 were had low FSG, 26 were medium, and only four pairs were in high both COS and FSG. LSA followed a similar trend, with only two pairs in the entire dataset being high on all three network norms. Because of this limitation, COS was equally represented at all three levels of overlap strength, but pairs were much more likely to have weaker associative or thematic relationships.

In addition to the limitations above, the stimuli selected for Experiment Two also had to be included in several unconnected databases of single word norms, which severely limited which words could be selected. For example, Experiment Two contained 56 word pairs with weak associative relationships, six with moderate associative overlap, and only one pair with high associative overlap. To help control for this, the single word norm analyses in Experiment Two used a combined data set where single word norms were gathered for the stimuli used in Experiment One, although this dataset contained several NAs for each single word norm predictor. While mean judgment and recall scores remained fairly stable across both experiments, some of the discrepancy between interaction findings (in particular the change in the direction of the interaction when predicting recall) may be remedied by using a more balanced set of stimuli. As such future studies should focus on creating overlap between current normed databases, as well as larger, more comprehensive collections of word norms for use in these types of studies.

REFERENCES

Adelman, J. S., & Brown, G. D. A. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review, 14*(3), 455–459.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (CD-ROM). Philadelphia.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445–459. doi: 10.3758/BF03193014.

Barton, K. (2018). MuMin: multi-model inference. *R* package version 1.40.4. https://cran.r-project.org/web/packages/MuMIn/index.html.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bradley, M. M., & Lang, P. J. (1999). *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings* (No. C-1). The Center for Research in Psychophysiology, University of Florida.

Brysbaert, M., Keuleers, E., & New, B. (2011). Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, *2*, 1–27. doi: 10.3389/fpsyg.2011.00027.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. doi: 10.3758/BRM.41.4.977.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A Tutorial. *Journal of Cognition, 1*(1), 1–20. doi: 10.5334/joc.10.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911.

Buchanan, E. M. (2010). Access into memory: Differences in judgments and priming for semantic and associative memory. *Journal of Scientific Psychology*, *March*, 1–8.

Buchanan, E. M., Holmes, J. L., Teasley, M. L., & Hutchison, K. A. (2013). English semantic word-pair norms and a searchable Web portal for experimental stimulus

creation. *Behavior Research Methods*, *45*(3), 746–757. doi: 10.3758/s13428-012-0284-z.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk. *Perspectives on Psychological Science*, *6*(1), 3–5. doi: 10.1177/1745691610393980.

Burgess, C., & Lund, K. (1997). Representing abstract words and emotional connotation in a high-dimensional memory space. *In Proceedings of the cognitive science* society (pp.61–66). Psychology Press.

Carreiras, M., Perea, M., & Grainger, J. (1997). Effects of the orthographic neighborhood in visual word recognition: Cross-task comparisons. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 857–871. doi: 10.1037/0278-7393.23.4.857.

Chow, B. W.-Y. (2014). The differential roles of paired associate learning in Chinese and English word reading abilities in bilingual children. *Reading and Writing*, *27*(9), 1657–1672. doi: 10.1007/s11145-014-9514-3.

Coltheart, M., Davelaar, E., Jonnasson, T., & Besner, D. (1977). Access to the internal lexicon. *Attention and Performance VI*, 535–555.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*(4), 589–608.

Cowan, N., Baddeley, A. D., Elliott, E. M., & Norris, J. (2003). List composition and the word length effect in immediate recall: A comparison of localist and globalist assumptions. *Psychonomic Bulletin & Review*, 10(1), 74–79. doi: 10.3758/BF03196469.

Dewhurst, S. A., Hitch, G. J., & Barry, C. (1998). Separate effects of word frequency and age of acquisition in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(2), 284–298. doi: 10.1037/0278-7393.24.2.284.

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, *48*(3), 432–435. doi: 10.1198/004017005000000661.

Green, P., & MacLeod, C. J. (2016). SIMR: An *R* package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498.

Harley, T. (2008). *The psychology of language: From data to theory* (3rd ed.) New York, NY: Psychology Press.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*(3), 662-720. doi: 10.1037/0033-295X.111.3.662.

Hertzog, C., Kidder, D. P., Powell-Moman, A., & Dunlosky, J. (2002). Aging and monitoring associative learning: Is monitoring accuracy spared or impaired? *Psychology and Aging*, *17*(2), 209–225. doi: 10.1037/0882-7974.17.2.209.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558. doi: 10.1073/pnas.79.8.2554.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review*, *10*(4), 785–813. doi: 10.3758/BF03196544.

Hutchison, K. A., & Bosco, F. A. (2007). Congruency effects in the letter search task: Semantic activation in the absence of priming. *Memory & Cognition*, *35*(3), 514-525. doi: 10.3758/BF03193291.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of International Conference Research on Computational Linguistics*, 19–33. doi: 10.1.1.269.3598.

Jones, L. L., & Golonka, S. (2012). Different influences on lexical priming for integrative, thematic, and taxonomic relations. *Frontiers in Human Neuroscience*, *6*, 1–17. doi: 10.3389/fnhum.2012.00205.

Jones, M. N., Willits, J., & Dennis, S. (2015). Models of semantic memory. In A. T. Townsend & Jerome R. Busemeyer (Eds.), *Oxford handbook of mathematical and computational psychology* (pp. 232–254). Oxford University Press.

Jouravlev, O., & McRae, K. (2016). Thematic relatedness production norms for 100 object concepts. *Behavior Research Methods*, *48*(4), 1349–1357. doi: 10.3758/s13428-015-0679-8.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 187–194. doi: 10.1037/0278-7393.31.2.187.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day English*. Providence, RI: Brown University Press.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44(*4), 978–990. doi: 10.3758/s13428-012-0210-4.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. doi: 10.1037//0033-295X.104.2.211.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2), 259–284. doi: 10.1080/01638539809545028.

Lefcheck, J. (2013, March 13). R^2 for linear mixed effects models [Blog post]. Retrieved from https://jonlefcheck.net/2013/03/13/r2-for-linear-mixed-effects-models/.

Lefcheck, J. (2016). Package 'piecewiseSEM'. *R* package version 1.2.1. https://cran.r-project.org/web/packages/piecewiseSEM/piecewiseSEM.pdf.

Lucas, M. (2000). Semantic priming without association: a meta-analytic review. *Psychonomic Bulletin & Review*, *7*(4), 618–630. doi: 10.3758/BF03212999.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. doi: 10.3758/BF03204766.

Maki, W. S. (2007a). Judgments of associative memory. *Cognitive Psychology*, *54*(4), 319–353. doi: 10.1016/j.cogpsych.2006.08.002.

Maki, W. S. (2007b). Separating bias and sensitivity in judgments of associative memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 231–237. doi: 10.1037/0278-7393.33.1.231.

Maki, W. S., & Buchanan, E. M. (2008). Latent structure in measures of associative, semantic, and thematic knowledge. *Psychonomic Bulletin & Review*, *15*(3), 598–603. doi: 10.3758/PBR.15.3.598.

Maki, W. S., McKinley, L. N., & Thompson, A. G. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). *Behavior Research Methods, Instruments, & Computers*, *36*(3), 421–431. doi: 10.3758/BF03195590.

McLeod, P., Shallice, T., & Plaut, D. C. (2000). Attractor dynamics in word recognition: Converging evidence from errors by normal subjects, dyslexic patients, and a connectionist model. *Cognition 74*(1), 91–114. doi: 10.1016/S0010-0277(99)00067-0.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559. doi: 10.3758/BRM.40.1.183.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words:668 Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. doi: 10.1037/h0031564.

Meyer, D. E., Schvaneveldt, R. W., & Ruddy, M. G. (1975). Loci of contextual effects on visual word-recognition. In P. M. A. Rabbitt (Ed.), *Attention and performance v*. London, UK: Academic Press.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition, 28*(6), 887–899. doi: 10.3758/BF03209337.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 3*6(3), 402–407. doi: 10.3758/BF03195588.

Nelson, D. L., Schreiber, T. A., & Xu, J. (1999). Cue set size effects: sampling activated associates or cross-target interference? *Memory & Cognition, 27*(3), 465–477. doi: 10.3758/BF03211541.

Paivio, A. (1969). Mental imagery in associative learning and memory. Psychological Review, *76*(3), 241–263. doi: 10.1037/h0021465.

Paivio, A. (1971). *Imagery and Verbal Processes*. Oxford: Holt, Rinehart, & Winston.

Peereman, R., & Content, A. (1997). Orthographic and phonological neighborhoods in naming: Not all neighbors are equally influential in orthographic space. *Journal of Memory and Language, 37*(3), 382–410.

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D., & R Core Team. (2017). nlme: Linear and Nonlinear Mixed Effects Models. Retrieved from https://cran.r-project.org/package=nlme68.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*(1), 56–115.

Richardson, J. T. E. (1998). The availability and effectiveness of reported mediators in associative learning: A historical review and an experimental investigation. *Psychonomic Bulletin & Review, 5*(4), 597–614. doi: 10.3758/BF03208837.

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation.

Topics in Cognitive Science, 3(2), 303–345. doi: 10.1111/j.1756-8765.2010.01111.x.

Rogers, T. T., & McClelland, J. L. (2006). *Semantic cognition*. Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986*). Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1*. Cambridge, MA: MIT Press.

Schwartz, B. L., & Brothers, B. R. (2013). Survival processing does not improve paired-associate learning. In B. L. Schwartz, M. L. Howe, M. P. Toglia, & H. Otgaar (Eds.), *What is adaptive about adaptive memory?* (pp. 159–171). Oxford University Press.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.

Smythe, P. C., & Paivio, A. (1968). A comparison of the effectiveness of word imagery and meaningfulness in paired-associate learning of nouns. *Psychonomic Science, 10*(2), 49–50. doi: 10.3758/BF03331401.

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods, 38*(4), 598–605. doi: 10.3758/BF03193891.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). New York, NY: Allyn & Bacon.

Toglia, M. P. (2009). Withstanding the test of time: The 1978 semantic word norms. *Behavior Research Methods, 41*(2), 531–533.

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Hillside, NJ: Earlbaum.

Valentine, K. D., & Buchanan, E. M. (2013). JAM-boree: An application of observation oriented modelling to judgements of associative memory. *Journal of Cognitive Psychology, 25*(4), 400–422. doi: 10.1080/20445911.2013.775120.

Venables, W. N. & Ripley, B. D. (2002) *Modern applied statistics with S* (4th ed.). New York, NY: Springer.

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods, 40*(1), 183–190. doi: 10.3758/BRM.40.1.183.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*(12).

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educators' word frequency guide.* Brewster, NY: Touchstone Applied Science.

Table 1. Summary Statistics of Single Word Norms for Experiment Two Cue Items.

Variable	Citation	Mean (SD)	Minimum	Maximum
QSS	Nelson et al., 2004	14.75 (4.49)	4.00	27.00
Concreteness	Nelson et al., 2004	5.27 (1.09)	1.98	7.00
SUBTLEX	Brysbaert & New, 2009	3.14 (.77)	1.34	5.35
Length	Buchanan et al., 2013	5.07 (1.43)	3.00	10.00
Ortho N	Buchanan et al., 2013	6.44 (5.79)	0.00	20.00
Phono N	Buchanan et al., 2013	16.55 (14.38)	0.00	51.00
Phonemes	Buchanan et al., 2013	4.11 (1.32)	2.00	9.00
Syllables	Buchanan et al., 2013	1.43 (.67)	1.00	4.00
Morphemes	Buchanan et al., 2013	1.06 (.24)	1.00	2.00
AOA	Kuperman et al., 2012	5.50 (1.75)	2.47	11.05
Valence	Warriner et al., 2013	5.69 (1.17)	1.91	7.89
Imageability	Toglia & Battig, 1978	5.41 (.76)	3.02	6.61
Familiarity	Toglia & Battig, 1978	6.18 (.29)	5.30	6.79
FSS	Buchanan et al., 2013	15.04 (10.46)	5.00	57.00
COSC	Buchanan et al., 2013	81.94 (73.59)	1.00	347.00

Note: QSS: Cue Set Size, Ortho N: Orthographic Neighborhood Size, Phone N: Phonographic Neighborhood Size, AOA: Age of Acquisition, FSS: Feature Set Size, COSC: Cosine Connectedness.

Table 2. Summary Statistics of Single Word Norms for Experiment Two Target Items.

Variable	Citation	Mean (SD)	Minimum	Maximum
TSS	Nelson et al., 2004	14.79 (5.06)	5.00	29.00
Concreteness	Nelson et al., 2004	5.34 (1.05)	1.28	7.00
SUBTLEX	Brysbaert & New, 2009	3.34 (.68)	1.59	5.36
Length	Buchanan et al., 2013	4.81 (1.68)	2.00	10.00
Ortho N	Buchanan et al., 2013	8.10 (7.47)	0.00	29.00
Phono N	Buchanan et al., 2013	19.16 (15.93)	0.00	59.00
Phonemes	Buchanan et al., 2013	3.86 (1.50)	1.00	10.00
Syllables	Buchanan et al., 2013	1.35 (.65)	1.00	4.00
Morphemes	Buchanan et al., 2013	1.06 (.23)	1.00	2.00
AOA	Kuperman et al., 2012	4.92 (1.66)	2.47	11.63
Valence	Warriner et al., 2013	5.81 (1.13)	1.95	7.89
Imageability	Toglia & Battig, 1978	5.46 (.75)	2.95	6.45
Familiarity	Toglia & Battig, 1978	6.28 (.29)	5.19	6.85
FSS	Buchanan et al., 2013	16.58 (12.95)	5.00	57.00
COSC	Buchanan et al., 2013	91.28 (89.90)	2.00	462.00

Note: TSS: Target Set Size, Ortho N: Orthographic Neighborhood Size, Phone N: Phonographic Neighborhood Size, AOA: Age of Acquisition, FSS: Feature Set Size, COSC: Cosine Connectedness.

Table 3. Summary Statistics for Experiment One Network Norms.

Variable	Citation	Mean (SD)	Minimum	Maximum
FSG	Nelson et al., 2004	.15 (.19)	.01	.75
COS	Maki et al., 2004	.44 (.28)	.00	.88
LSA	Landauer & Dumais, 1997	.36 0.19)	.03	.90

Note: COS: Cosine, FSG: Forward Strength, LSA: Latent Semantic Analysis
After viewing the examples at the start of the block, participants completed the

Table 4. Summary Statistics for Experiment Two Network Norms.

Variable	Citation	Mean (SD)	Minimum	Maximum
FSG	Nelson et al., 2004	.13 (.19)	.01	.83
COS	Maki et al., 2004	042 (.29)	.00	.84
LSA	Landauer & Dumais, 1997	.38 (.20)	.05	.88

Note: COS: Cosine, FSG: Forward Strength, LSA: Latent Semantic Analysis

Table 5. Summary Statistics for Experiment One Hypothesis One.

Variable	Mean (SD)	*t* (df)	*p*
A Intercept	.511 (.245)	20.864 (99)	< .001
A COS	-.030 (.284)	-1.071 (99)	.287
A FSG	.491 (.379)	12.946 (99)	< .001
A LSA	.035 (.317)	1.109 (99)	.270
S Intercept	.587 (.188)	31.530 (101)	< .001
S COS	.059 (.243)	2.459 (101)	.016
S FSG	.118 (.382)	3.128 (101)	.002
S LSA	.085 (.304)	2.816 (101)	.006
T Intercept	.656 (.186)	35.475 (100)	< .001
T COS	-.081 (.239)	-3.405 (100)	< .001
T FSG	.192 (.306)	6.290 (100)	< .001
T LSA	.188 (.265)	7.111 (100)	< .001

Note: A: Associative judgments, S: Semantic judgments, T: Thematic judgments

Table 6. MLM statistics for Experiment One Hypothesis Two.

Variable	*b*	*SE*	*t*	*p*
Intercept	.603	.014	43.287	< .001
COS	-.103	.017	-6.081	< .001
LSA	.090	.022	4.196	< .001
FSG	.271	.029	9.420	< .001
COS:LSA	-.141	.085	-1.650	.099
COS:FSG	-.374	.111	-3.364	< .001
LSA:FSG	-.569	.131	-4.336	< .001
COS:LSA:FSG	3.324	.490	6.791	< .001

Note: Database norms were mean centered.

Table 7. MLM statistics for Experiment One Hypothesis Three.

Variable	*b*	*SE*	*z*	*p*
Intercept	.301	.138	2.188	.029
COS	.594	.179	3.320	< .001
LSA	-.350	.204	-1.714	.087
FSG	3.085	.302	10.205	< .001
COS:LSA	2.098	.837	2.506	.012
COS:FSG	1.742	1.306	1.334	.182
LSA:FSG	-1.017	1.484	-0.685	.493
COS:LSA:FSG	24.572	6.048	4.063	< .001

Note: Database norms were mean centered.

Table 8. MLM Statistics for Hypothesis Four

Variable	*b* (*SE*)	*z*	*p*
(Intercept)	-.432 (0.439)	-.983	.326
A Intercept	1.514 (0.604)	2.507	.012
A COS	.314 (0.550)	.572	.568
A FSG	.898 (0.337)	2.667	.008
A LSA	.501 (0.463)	1.081	.279
(Intercept)	-.827 (0.463)	-1.787	.074
S Intercept	2.292 (0.681)	3.363	<0.001
S COS	2.039 (0.518)	3.939	< .001
S FSG	.381 (0.289)	1.319	.187
S LSA	1.061 (0.455)	2.335	.020
(Intercept)	.060 (0.599)	.101	.920
T Intercept	1.028 (0.756)	1.360	.174
T COS	.792 (0.566)	1.401	.161
T FSG	-.394 (0.441)	-.894	.371
T LSA	.896 (0.529)	1.694	.090

Note: A: Associative judgments, S: Semantic judgments, T: Thematic judgments

Table 9. MLM Statistics for Judgment Replication

Variable	*b*	*SE*	*t*	*p*
Intercept	.615	.009	66.43	< .001
COS	.011	.011	1.054	.293
LSA	.132	.018	7.386	< .001
FSG	.422	.020	20.622	< .001
COS:LSA	-.359	.059	-6.033	< .001
COS:FSG	-.171	.059	-1.968	.049
LSA:FSG	-.456	.153	-2.972	.003
COS:LSA:FSG	.193	.410	.471	.638

Note: Database norms were mean centered.

Table 10. MLM Statistics for Recall Replication

Variable	*b*	*SE*	*z*	*p*
Intercept	.303	.107	2.825	.005
COS	.633	.099	6.421	< .001
LSA	.681	.163	4.180	< .001
FSG	1.780	.198	9.081	< .001
COS:LSA	3.084	.537	5.748	< .001
COS:FSG	2.011	.833	5.414	.016
LSA:FSG	-.962	1.391	-.691	.489
COS:LSA:FSG	-22.464	3.671	-6.119	< .001

Note: Database norms were mean centered.

Table 11. Single Word IVs Retained after Stepwise Analyses

Step	Judgment Models	Recall Models
One	Length 1	Length 1
One	Length 2	Length 2
One	SUBTLEX 1	SUBTLEX 1
One	SUBTLEX 2	SUBTLEX 2
Two	AOA 1	AOA 1
Two	AOA 2	AOA 2
Two	Familiarity 1	Familiarity 1
Two	Familiarity 2	Familiarity 2
Two	Valence 1	Valence 1
Two	Valence 2	Valence 2
Two	Imageability 1	Imageability 1
Two	Imageability 2	Imageability 2
Two	Concreteness 1	Concreteness 2
Two	Concreteness 2	------
Three	QSS	QSS
Three	TSS	TSS
Three	FSS 1	FSS 1
Three	FSS 2	FSS 2
Three	Ortho 2	Ortho 1
Three	Phono 1	Ortho 2
Three	COSC 2	Phono 1
Three	------	Phono 2

Note: 1 = Cue item, 2 = Target item

Table 12. MLM Statistics for Hierarchical Judgment Model

Step	IV	*b* (SE)	*t*	*p*
One	SUBTLEX 1	.014 (.003)	4.308	< .001
	SUBTLEX 2	-.032 (.003)	-9.219	< .001
	Length 1	-.005 (.001)	-2.790	.005
	Length 2	-.001 (.001)	-.819	.413
Two	AOA 1	.015 (.002)	7.416	< .001
	AOA 2	-.014 (.002)	-6.348	< .001
	Familiarity 1	.075 (.013)	5.694	< .001
	Familiarity 2	-.091 (.011)	-8.160	< .001
	Valence 1	-.001 (.002)	-.072	.9425
	Valence 2	-.022 (.002)	-10.257	< .001
	Imageability 1	.053 (.005)	10.256	< .001
	Imageability 2	-.074 (.005)	-13.440	< .001
	Concreteness 1	-.015 (.004)	-4.144	< .001
	Concreteness 2	.045 (.004)	10.128	< .001
Three	QSS	.005 (.001)	-8.910	< .001
	TSS	-.002 (.001)	-3.676	.160
	FSS 1	-.001 (.001)	-4.339	< .001
	FSS 2	.001 (.001)	4.884	< .001
	Ortho N 2	.001 (.001)	7.403	< .001
	Phono N 1	-.001 (.001)	-7.789	< .001
	COSC 2	-.001 (.001)	-3.288	.010
Four	FSG	.391 (.021)	18.186	< .001
	LSA	.123 (.106)	7.919	< .001
	COS	.027 (.011)	2.429	.015
	COS:FSG	-.100 (.091)	-1.091	.276
	COS:LSA	-.367 (.058)	-6.379	< .001
	LSA:FSG	-.393 (.106)	-3.691	< .001
	COS:FSG:LSA	.558 (.338)	1.648	.099

Note: 1 = Cue item, 2 = Target Item. FSG, COS, and LSA have been mean centered. Statistics are reported for the step in which the variable was entered into the model.

Table 13. MLM Statistics for Hierarchical Recall Model

Step	IV	*b* (SE)	*z*	*p*
One	SUBTLEX 1	-.257 (.280)	-9.234	< .001
	SUBTLEX 2	.082 (.030)	2.761	.006
	Length 1	.138 (.015)	9.225	< .001
	Length 2	-.047 (.012)	-3.866	< .001
Two	AOA 1	-.023 (.018)	-1.331	.183
	AOA 2	-.087 (.019)	-4.474	< .001
	Familiarity 1	-.587 (.118)	-4.975	< .001
	Familiarity 2	-.283 (.098)	-2.873	.004
	Valence 1	-.143 (.020)	-7.269	< .001
	Valence 2	-.012 (.019)	-.619	.536
	Imageability 1	.086 (.032)	2.697	.007
	Concreteness 2	-.131 (.027)	-4.838	< .001
Three	QSS	.001 (.003)	.271	.786
	TSS	-.015 (.004)	-3.401	< .001
	FSS 1	-.012 (.002)	-5.494	< .001
	FSS 2	.015 (.002)	7.305	< .001
	Ortho N 1	-.017 (.006)	-3.023	.003
	Ortho N 2	-.007 (.004)	-1.579	.114
	Phono N 1	-.001 (.002)	-.117	.907
	Phono N 2	-.007 (.002)	-3.087	.002
Four	FSG	1.866 (.210)	8.880	< .001
	LSA	.867 (.146)	4.710	< .001
	COS	0.278 (.102)	2.713	.007
	COS:FSG	-1.014 (.905)	-1.120	.263
	COS:LSA	3.779 (.524)	7.205	< .001
	LSA:FSG	-1.862 (1.010)	-1.844	.065
	COS:FSG:LSA	-8.808 (3.161)	-2.786	.005

Note: 1 = Cue item, 2 = Target Item. FSG, COS, and LSA have been mean centered. Statistics are reported for the step in which the variable was entered into the model.

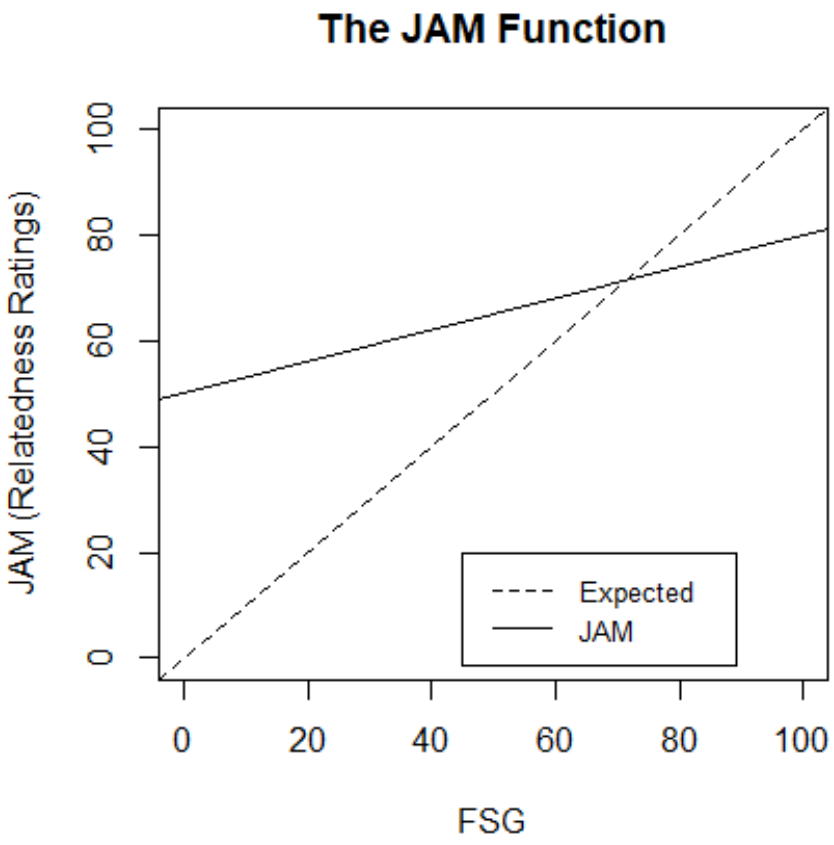

Figure 1. JAM slope findings from Maki (2007a). JAM is characterized by a high intercept (between 40 and 60) and a shallow slope (between .20 and .40). The dashed line shows expected results if judgment ratings are perfectly calibrated with association norms.

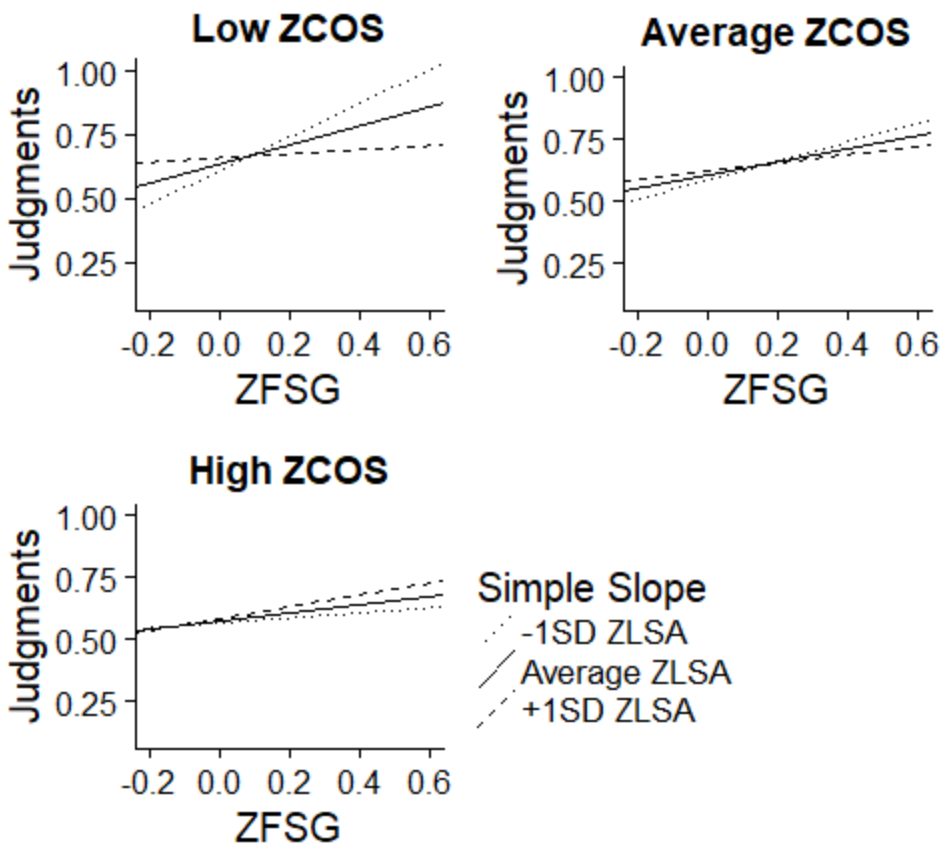

Figure 2. Simple slopes graph displaying the slope of FSG when predicting participant judgments at low, average, and high LSA split by low, average, and high COS. All variables were mean centered.

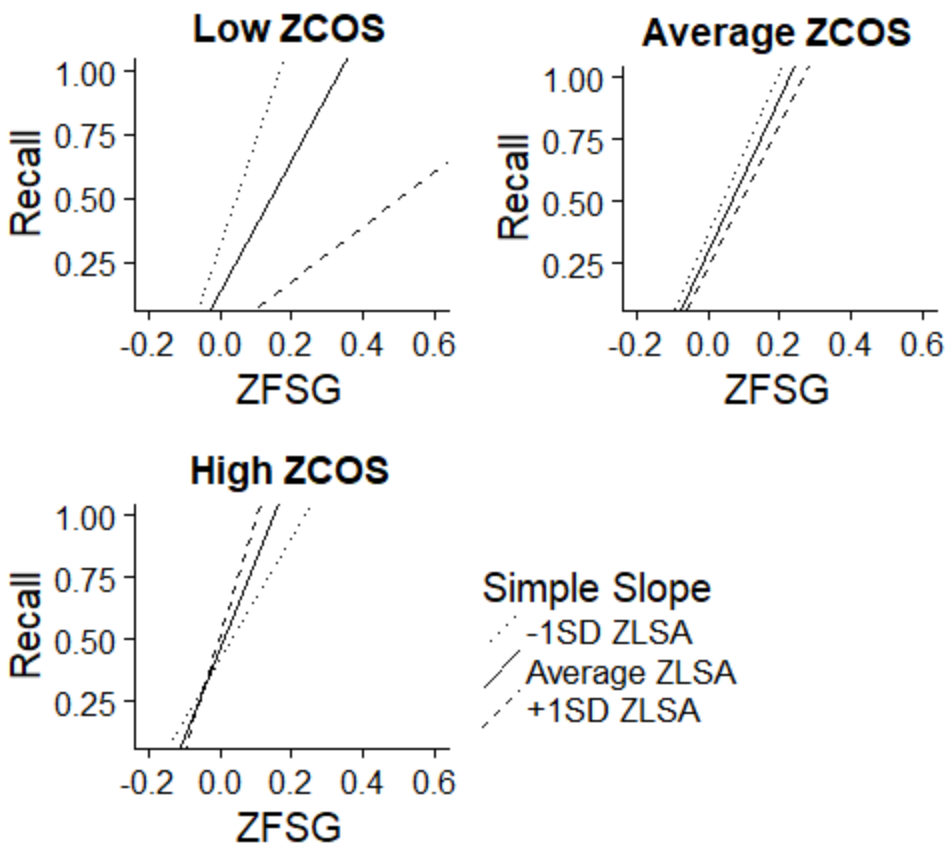

Figure 3. Simple slopes graph displaying the slope of FSG when predicting participant judgments at low, average, and high LSA split by low, average, and high COS. All variables were mean centered.

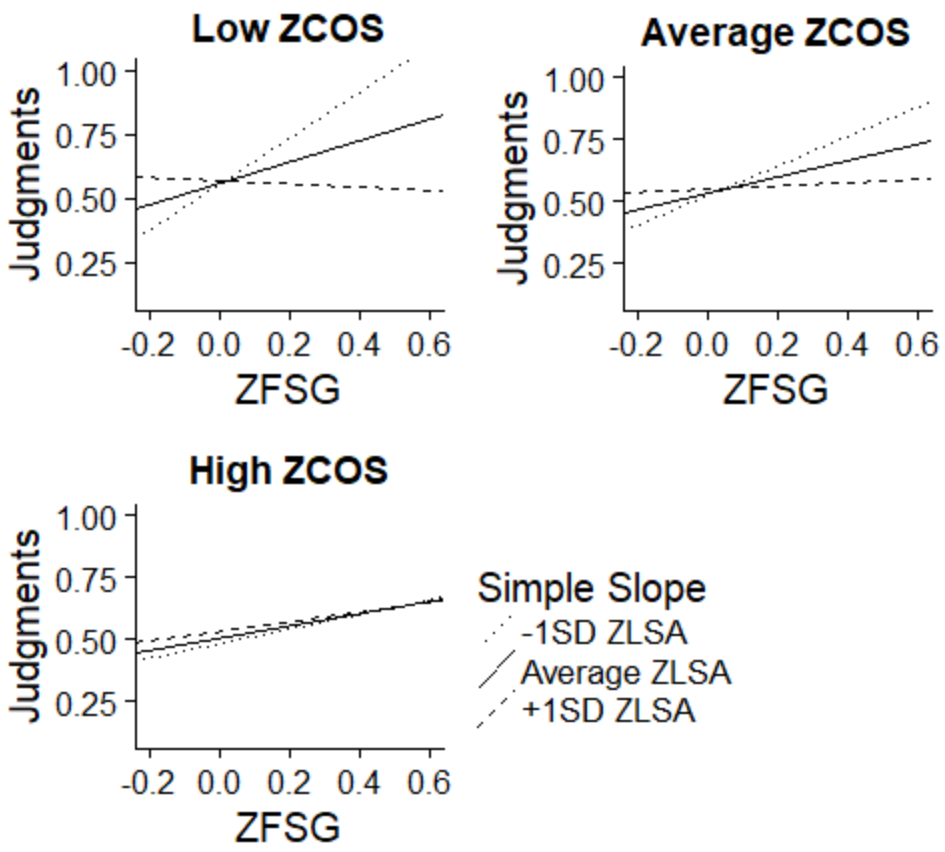

Figure 4. Simple slopes graph displaying the slope of FSG when predicting participant judgments based on block one performance at low, average, and high LSA split by low, average, and high COS. All variables were mean centered.

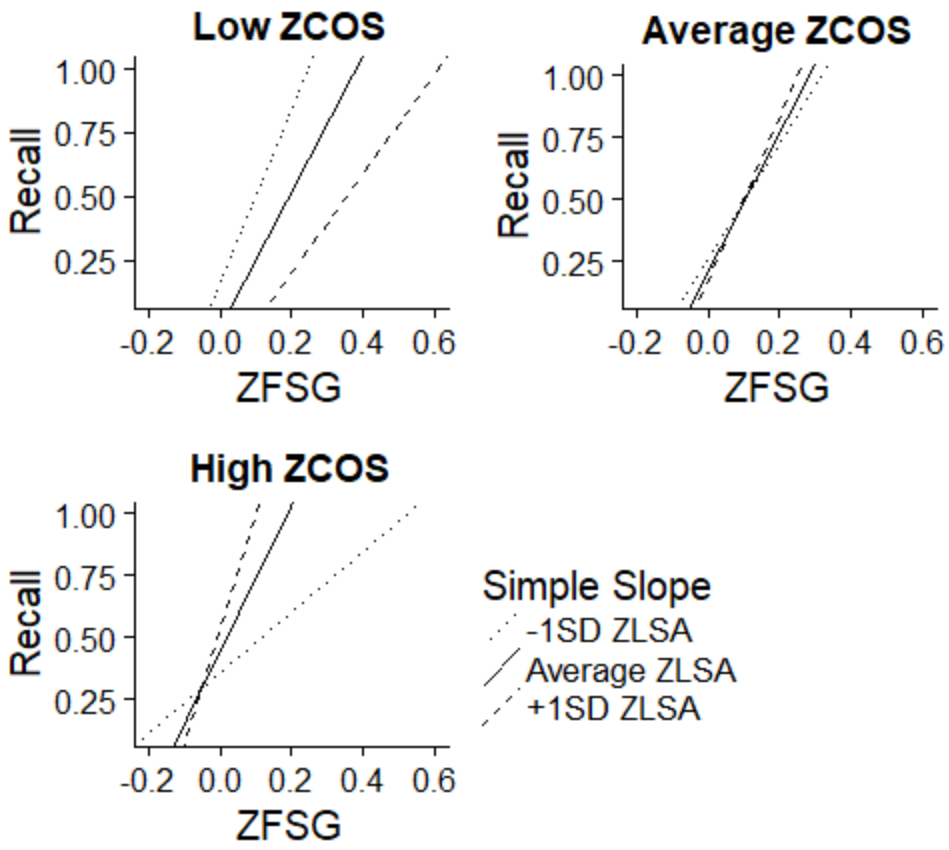

Figure 5. Simple slopes graph displaying the slope of FSG when predicting participant recall based on block one performance at low, average, and high LSA split by low, average, and high COS. All variables were mean centered.

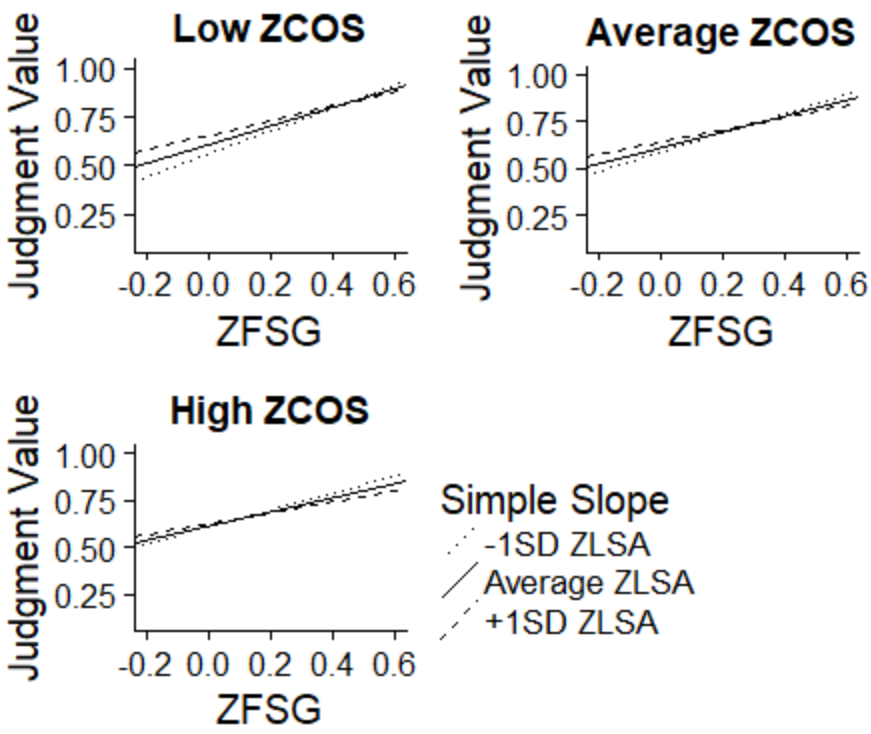

Figure 6. Simple slopes graph displaying the slope of FSG when predicting participant judgments at low, average, and high LSA split by low, average, and high COS. All variables were mean centered.

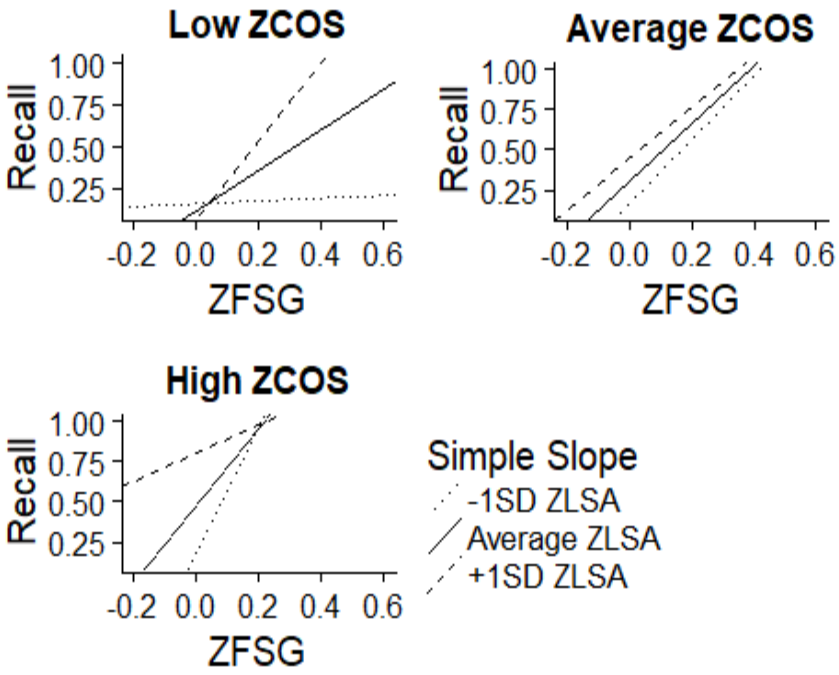

Figure 7. Simple slopes graph displaying the slope of FSG when predicting participant recall at low, average, and high LSA split by low, average, and high COS. All variables were mean centered.

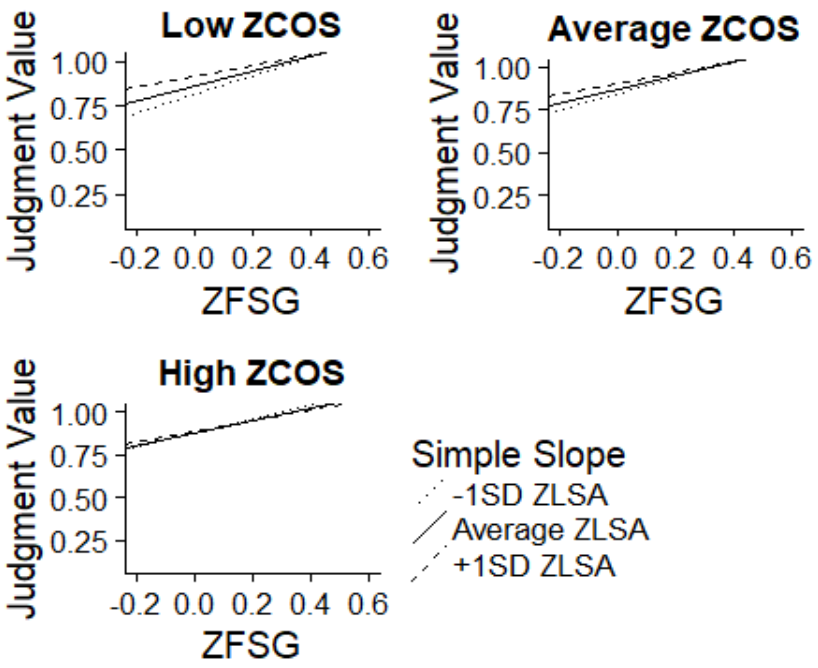

Figure 8. Simple slopes graph displaying the slope of FSG when predicting participant judgments at low, average, and high LSA split by low, average, and high COS while also controlling for single word norms. FSG, LSA, and COS have been mean centered.

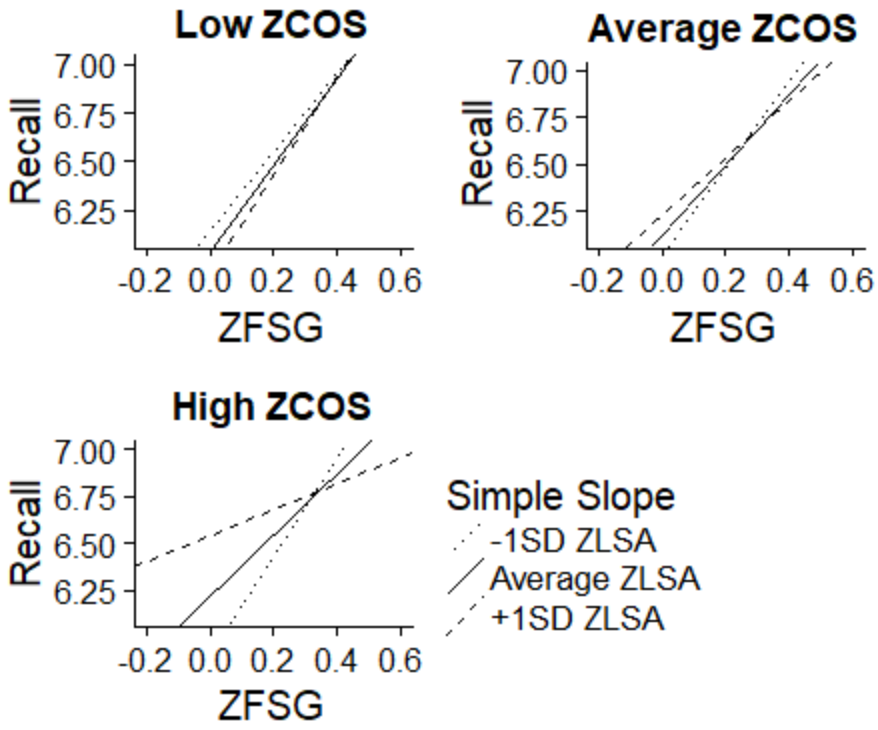

Figure 9. Simple slopes graph displaying the slope of FSG when predicting participant recall at low, average, and high LSA split by low, average, and high COS while also controlling for single word norms. FSG, LSA, and COS have been mean centered.

APPENDIX

Instructions for associative judgments:

This experiment is concerned with the structure of human associative memory. This knowledge is structured in some way and the mental structure is thought to come about through a process of associative learning. For example, consider the word (and concept of) DOG. We often see the word DOG appear in the same context as the word CAT. "It's raining cats and dogs." "I have two dogs, but my neighbor has a cat." And so on. By experiencing the words CAT and DOG together many times, we develop an association (a mental connection) between them. With lots of this kind of associative learning experience during our lives, we develop a very large and very complex associative memory.

Psychologists are interested in understanding the structure of associative memory and have several ways of investigating it. One method of investigating associative memory is known as a test of "free association." In free association tests, participants like you are given a series of words and are asked to respond to each word by writing the first word that pops into mind.

Here is an example of a free association test.

In the space provided, please write the first word that comes to your mind in response to each of the following words.

LOST

OLD

ARTICLE

Participants have many responses to the free association task, and here are some examples of the ones that people normally write:

LOST - FOUND, LOST has one very strong associate.

OLD - NEW or YOUNG, OLD has two equally strong associates.

ARTICLE - NEWSPAPER or THE or CLOTHING and many others, ARTICLE has many weak associates.

Instructions for semantic judgments:

This block is concerned with the how humans perceive similarities between the characteristics which define words. Consider the following words (and concepts) TORTOISE, TURTLE, SNAIL, and BANNER. We know that a TORTOISE is a reptile with an exoskeleton and a hard shell. If we compare the word TORTOISE with the word TURTLE, we find that they share a majority of the same features. Therefore, their definitions or characteristics overlap greatly.

When compared, the words TORTOISE and SNAIL are found to have characteristics that are fairly similar, but there are almost as many dissimilar traits as there are similar. For example, both animals are reptiles and have exoskeleton shells. However a tortoise is a reptile and a snail is a mollusk. This means that the characteristics only moderately overlap. Although TORTOISE and SNAIL share many of the same taxonomy traits, the psychological concepts shared by TORTOISE and SNAIL differ from thw taxonomic traits in the sense that the psychological traits separate the two concepts more than taxonomy would.

If we compare the word TORTOISE with the word BANNER we find that there is very little which these two objects have in common.TORTOISE and BANNER are so dissimilar because the concepts of what make a TORTOISE and what make a BANNER hardly overlap at all. For example they are both physical objects, but a TORTOISE is animate and a BANNER is inanimate.

Here is an example of a semantic characteristic task.

In the space provided, please fill in one or two properties of the concept that you can think of to which the word refers. Examples of different types of properties would be: physical properties, such as internal and external parts, and how it looks, sounds, smells, feels, or tastes; functional properties, such as what it is used for; where, when and by whom it is used; things that the concept is related to, such as the category that it belongs in; and other facts, such as how it behaves, or where it comes from.

LOST

OLD

ARTICLE

Participants have many responses to the semantic characterization task, and here are some examples of the ones that people normally write:

LOST - missing, not found, gone

OLD - age, not young

ARTICLE - has words, is part of a paper, is a word

Instructions for thematic judgments:

This block is concerned with the thematic relationships between word pairs. Words that are thematically related are connected by a related concept and may often occur near each other in language. For example, the word TREE is thematically related to LEAF, FRUIT, BRANCH, and FOREST because they all appear in text together due to related meaning. TREE and COMPUTER would not be thematically related because they would not be in the same writing together.

Here is an example of a theme creation test.

In the space provided, please write down words that you think are thematically related to the provided word.

LOST	
OLD	
ARTICLE	

Participants have many responses to the thematic creation task, and here are some examples of the ones that people normally write:

LOST - missing, children, kids, found, objects

OLD - age, antiques, young, people

ARTICLE - newspaper, scandal, print, press