Modeling Memory: Exploring the Relationship Between Word Overlap and Single Word Norms when Predicting Relatedness Judgments and Retrieval

Nicholas P. Maxwell[1] & Erin M. Buchanan[1]

[1] Missouri State University

Author Note

Nicholas P. Maxwell is a graduate student at Missouri State University. Erin M. Buchanan is an Associate Professor of Psychology at Missouri State University.

Correspondence concerning this article should be addressed to Nicholas P. Maxwell, 901 S. National Ave, Springfield, MO, 65897. E-mail: maxwell270@live.missouristate.edu

Abstract

This study examined the interactive relationship between semantic, thematic, and associative word pair strength in the prediction of item relatedness judgments and cued-recall performance. Previously, we found significant three-way interactions between associative, semantic, thematic word overlap when predicting participant judgment strength and recall performance (Maxwell & Buchanan, 2018), expanding upon previous work by Maki (2007a). In this study, we first seek to replicate findings from the original study using a novel stimuli set. Second, this study will further explore the nature of the structure of memory, by investigating the effects of single concept information (i.e., word frequency, concreteness, etc.) on relatedness judgments and recall accuracy. We hypothesize that associative, semantic, and thematic memory networks are interactive in their relationship to judgments and recall, even after controlling for base rates of single concept information, implying a set of interdependent memory systems used for both cognitive processes.

Keywords: judgments, memory, association, semantics, thematics

Modeling Memory: Exploring the Relationship Between Word Overlap and Single Word Norms when Predicting Relatedness Judgments and Retrieval

Previous research conducted on Judgments of Associative Memory (JAM) has found that these judgments tend to be stable and highly generalizable across varying contexts (Maki, 2007a, 2007b; Valentine & Buchanan, 2013). The JAM task can be viewed as a manipulation of the traditional Judgment of Learning task (JOL). In a JOL task, participants are presented with cue-target word pairs and are asked to make a judgment (typically, on a scale of zero to 100) of how accurately they would be able to respond with the proper target word based on the presentation of a particular cue word (Dunlosky & Nelson, 1994; Nelson & Dunlosky, 1991). JAM tasks expand upon this concept by changing the focus of the judgments performed by participants. When presented with the item pair, such as *cheese-mouse*, participants are asked to judge the number of people out of 100 who would respond with the pair's target word if they were only shown the cue word (Maki, 2007a).

This process mimics the creation of associative words norms (i.e., forward strength; D. L. Nelson, McEvoy, & Schreiber, 2004). As such, these judgments can be viewed as the participants' approximations of how associatively related they perceive the paired items to be. The JAM function can then be created by plotting participants' judgments against the word's normed associative strength and calculating a line of best fit. This fit line typically displays a high intercept (bias) and a shallow slope (sensitivity), meaning that participants are biased towards overestimating the associative relatedness between word pairs, and show difficulties differentiating between different amounts of item relatedness (Maki, 2007a). These results are often found in JOL research (Koriat & Bjork, 2005), and they are highly stable across various contexts and instructional manipulations in JAM tasks (Valentine & Buchanan, 2013).

Building upon this research, we initially explored recall accuracy within the context of word pair judgments, while also expanding the JAM task to incorporate judgments of semantic and thematic memory. In the pilot study, 63 word-pairs of varying associative, semantic, and thematic overlap were created and arranged into three blocks, consisting of 21 word-pairs each. Associative overlap was measured with forward strength (FSG; D. L. Nelson et al., 2004), semantic overlap was measured with cosine (COS; Buchanan, Holmes, Teasley, & Hutchison, 2013; McRae, Cree, Seidenberg, & McNorgan, 2005; Vinson & Vigliocco, 2008), and thematic relatedness between pairs was measured with latent semantic analysis (LSA; Landauer & Dumais, 1997; Landauer, Foltz, Laham, Folt, & Laham, 1998). These word pairs were then judged by 112 participants who were recruited from Amazon's Mechanical Turk. Stimuli were arranged into three blocks based, each preceeded by a set of instructions explaining either an associative, semantic, or thematic relationship between words. Three versions of the study were created, counterbalancing the order in which judgment instructions and stimuli blocks appeared. Thus, each participant made one set of judgments corresponding to each type of memory, and each word pair recieved each type of judgment.

After completing the judgment phase, participants then completed a cued recall task in which they were presented with the cue word from each of the previously presented word pairs and were asked to complete each pair with the missing target (Maxwell & Buchanan, 2018). Significant three-way interactions were found between database norms when predicting judgments and recall. When semantic overlap was low, thematic and associative strength were competitive, with increases in thematic overlap decreasing the strength of associative overlap as a predictor. However, this trend saw a reversal when semantic overlap was high, with thematic and associative strength complimenting one another. Overall, our findings from this study indicated the degree to which the processing of associative, semantic, and thematic information impacts retrieval and judgment making, while also displaying the interactive relationship that exists between these three types of information.

The proposed study seeks to expand upon this work by extending the original analysis to include multiple single word norms. These norms provide information about different "neighborhoods" of concept information. Broadly speaking, they can be separated into one of three categories. Base values refer to norms which capture information based on a word's structure. These include part of speech, word frequency, and the number of syllables, morphemes, and phonemes that comprise a word. Rated values refer to age of acquisition, concreteness, imageability, valence, and familiarity. Finally, we seek to examine norms that provide information about the connections a word shares with others based on context. These norms include orthographic neighborhood, phonographic neighborhood, cue and target set sizes, and feature set size. These values were selected on the basis of previous research suggesting their impact on retrieval accurarcy; their importance is elaborated upon below.

First, we are interested in assessing the impact of base word norms. Chief amongst these is word frequency. Several sets of norms currently exist for measuring the frequency with which words occur in everyday language, and it is important to determine which of these offers the best representation of everyday language. One of the most commonly used collections of these norms is the Kučera and Francis (1967) frequency norms. This set consists of frequency values for 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 analyzed, and the overall age of these norms. First, these norms were created from an 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 best 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 exist. One popular method is to use frequency norms obtained from the HAL corpus, which consists of 131 million words (Burgess & Lund, 1997; Lund & Burgess, 1996). Other collections of frequency norms include CELEX (Baayen, Piepenbrock, & Gulikers, 1995) 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 taken from books published in the United States (see Brysbaert, Keuleers, and New (2011) for an overview and comparison of these norms). For the present study, we plan to use data taken from the both the SUBTLEX project (Brysbaert & New, 2009), which is a collection of frequency norms derived from a corpus of approximately 51 million words, which were generated from movie and television subtitles and the HAL corpus. 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 corpus size of both SUBTLEX and HAL contributes to the validity of these norms compared to Kučera and Francis (1967) frequency norms.

Next, we are interested in testing the effects of several measures of lexical information related 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 (measured as the number of individual characters a word consists of) and the number of syllables a word contains will be investigated, as previous research has suggested that the number of syllables 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). Finally, we are interested in the part of speech of each word, as nouns are often easier to remember (Paivio, 1971). Formall defined, part of speech refers to a word's categorization in language based on its syntactic functions.

Third, we will examine the effects of norms measuring word properties that are rated by participants. The first of these is age of acquisition, which is a measure of the age at which a word is learned. This norm is measured by presenting participants with a word and having them enter the age (in years) in which they believe that they would have learned the word (Kuperman et al., 2012). Age of acquisition ratings have been found to be predictive of recall; for example, Dewhurst, Hitch, and Barry (1998) found recall to be higher for late 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. These include research on emotion, the impact of emotion of 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 (see Warriner, Kuperman, and Brysbaert (2013) for a review). The next of these rated measures is concreteness, which refers to the degree that a word relates to a perceptible object (Brysbaert, Warriner, & Kuperman, 2014). Similar to concreteness, imageability is 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 recalled (D. L. Nelson & Schreiber, 1992). Finally, familiarity norms can be described as an application of word frequency. These norms measure the frequency of exposure to a particular word (Stadthagen-Gonzalez & Davis, 2006).

The final group of norms that will be investigated are 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 created by changing a single letter in word (i.e., *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 measures have been found to effect processing speed for items (Adelman & Brown, 2007; Buchanan et al., 2013;

Coltheart, Davelaar, Jonasson, & Besner, 1977). Next, we are interested in examining two single word norms that are directly related to item associations. These norms measure the number of associates a word shares connections with. Cue set size refers to the number of cue words that a target word is connected to, while target set size 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 (D. L. Nelson, Schreiber, & Xu, 1999). As such, feature list sizes will be calculated for each word overlap norm from the Buchanan et al. (2013) semantic feature norm set.

In summary, this study seeks to expand upon previous work by examining how single word norms belonging to these three neighborhoods of item information impact the accuracy of item judgments and recall. These findings will be assessed within the context of associative, semantic, and thematic memory systems. Specifically, we utilize a three-tiered view of the interconnections between these systems as it relates to processing concept information. First, semantic information is processed, which provides a means for categorizing concepts based on feature similarity. Next, processing moves into the associative memory network, where contextual information pertaining to the items is added. Finally, the thematic network incorporates information from both the associative and semantic networks to generate a mental representation of the concept containing both the items meaning and its place in the world.

Therefore, the present study has two aims. First, we seek to replicate the interaction results from the previous study using a new set of stimuli. Second, we wish to expand upon these findings by extending the analysis to include neighborhood information for the item pairs. The extended analysis will be analyzed by introducing the different types single word norms through a series of steps based on the type of neighborhood they belong to. First, base word norms will be analyzed. Next, measures of word ratings will be analyzed. Third, single word norms measuring connections between concepts will be analyzed. Finally, network norms and their interactions will be reanalyzed. The end goal is to determine both which neighborhood of norms have the greatest overall impact on recall and judgment ability, and to further assess the impact of network connections after controlling for the various neighborhoods of single word information.

Methods

Participants

A power analysis was conducted using the *simr* package in *R* (Green & MacLeod, 2016), which uses simulations to calculate power for mixed linear models created from the *lme4* and *nlme* packages (Bates, Mächler, Bolker, & Walker, 2015; Pinheiro, Bates, Debroy, Sarkar, & Team, 2017). The results of this analyses suggested a minimum of 35 participants was required to detect the interaction at 80% power (α = .05). However, because power is often underestimated (Bakker, Hartgerink, Wicherts, & Maas, 2016; Brysbaert & Stevens, 2018), we plan to extend the analysis to include approximately 200 participants, a number determined by the amount of available funding. Consistent with the design of the pilot study, participants will be recruited from Amazon's Mechanical Turk, which is a website where individuals can host projects and be connected with a large respondent pool who complete tasks for small amounts of money (Buhrmester, Kwang, & Gosling, 2011). Participants will be paid $2.00 for their participation. Participant responses will be screened for a basic understanding of study instructions and for automated survey responses. Data will be excluded for participants who respond with words when asked to make numerical judgements, respond with numerical ratings during the recall task, or fail to complete either the judgment or recall tasks.

Materials

First, mimicking the design of the original pilot study, sixty-three word pairs of varying associative, semantic, and thematic overlap were created to use as stimuli. As with the pilot study, these word pairs were created using the Buchanan et al. (2013) word norm database. Next, neighborhood information for all cue and target items was collected. 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 was derived from the Buchanan et al. (2013) word norms (originally contained in The English Lexicon Project, Balota et al., 2007). For items with multiple parts of speech (for example, *drink* can refer to both a beverage and the act of drinking a beverage), part of speech was coded as the most commonly used form. Following the design of Buchanan et al. (2013), this part of speech was determined using Google's define feature. Concreteness, cue set size, and target set size were taken from the South Florida Free Association Norms (D. L. Nelson et al., 2004). Feature set size (i.e., the number of features listed as part of the definition of a concept) and cosine set size (i.e., 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 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. Stimuli information for cue and target words can be found in Tables 1 and 2.

After gathering neighborhood information, network norms measuring associative, semantic, and thematic overlap were generated for each pair. Forward strength (FSG) was used as a measure of associative overlap. FSG is a value ranging from zero to one which measures of the probability that a cue word will elicit a particular target word in response to it (D. L. Nelson et al., 2004). Cosine (COS) strength was used to measure semantic overlap between concepts (Buchanan et al., 2013; McRae et al., 2005; Vinson & Vigliocco, 2008). As with FSG, this value ranges from zero to one, with higher values indicating more shared features between concepts. Finally, thematic overlap was measured with Latent Semantic Analysis (LSA), which is a measure generated based upon the co-occurrences of words within a document (Landauer & Dumais, 1997; Landauer et al., 1998). Like the measures of associative and semantic overlap, LSA values range from zero to one, with higher values indicating higher co-occurrence between items. The selected stimuli contained a range of values across both the network and neighborhood norms. As with the previous study, stimuli will be arranged into three blocks, with each block consisting of 21 word pairs. The blocks will be structured to have seven words of low COS (0 - .33), medium COS (.34 - .66), and high COS (.67 - 1). COS was chosen due to both limitations with the size of the available data set across all norm sets, and the desire to recreate the selection process used for the previous study. The result of this selection process is that values for the remaining network norms (FSG and LSA) and information neighborhood norms will be contingent upon the COS strengths of the selected stimuli. To counter this, we selected stimuli at random based on the different COS groupings so as to cover a broad range of FSG, LSA, and information neighborhood values. Stimuli information for word pair norms can be found in Table 3. All stimuli and their raw values can be found at https://osf.io/j7qtc/.

Procedure

This study will be divided into three sections. First, participants will be presented with word pairs and will be asked to judge how related the items are to one another. This section will comprise three blocks, with each block containing 21 word pairs. Each item block will be preceded by a set of instructions explaining one of the three types of relationships. Participants will also be provided with examples illustrating the type of relationship to be judged. The associative instructions explain associative relationships between concepts, how these relationships can be strong or weak, and the role of free association tasks in determining the magnitude of these relationships. The semantic instructions will provide participants with a brief overview of how words can be related by meaning and will give participants examples of item pairs with low and high levels of semantic overlap. Finally, the thematic instructions will explain how concepts can be connected by overarching themes. These instruction sets are modeled after Buchanan (2010) and Valentine and Buchanan (2013).

To clarify, the association instruction set includes the following instructional explanation focusing on the co-occurrence in language: "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."

While the semantic instructions focus on the definition and feature overlap of a set of concepts: "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."

Last, the thematic instructions contain a blend of the two instruction sets to focus on both semantic and associative relation: "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."

Participants will then rate the relatedness of the word pairs based on the set of instructions they receive at the start of each judgment block. These judgments will be made using a scale of zero (no relatedness between pairs) to one hundred (a perfect relationships). The instructions for association were: "Assume 100 college students from around the nation gave responses to each CUE (first) word. How many of these 100 students do you think would have given the RESPONSE (second) word?" The semantic instructions were: "Assume both CUE and RESPONSE words have various features like you filled in before. What percent of those features that are the same? Use a scale of 0 to 100, with 0 indicating no relationship, and 100 indicating a perfect relationship." Finally, the thematic instructions were: "Using the two words provided, think about how often those two words would be written together in the same story. Please rate the thematic strength of the following word pairs using a scale of 0 to 100, with 0 indicating no relationship, and 100 indicating a perfect relationship." The complete instructions and examples provided can be found on our OSF page for replication.

Judgments were recorded by the participant typing it into the survey. Participants will complete each of the three judgment blocks in this manner, with judgment instructions changing with each block. Three versions of the study will be created to counterbalance the order in which judgment blocks appear. Stimuli are counterbalanced across blocks, such that each word pair is seen once per subject but evenly spread across all three judgment types. Word pairs are randomized within each block. Participants will be randomly assigned to survey conditions. After completing the judgment blocks, participants will be presented with a short distractor task to account for recency effects. This section will be timed to last two minutes and will task participants with alphabetizing a scrambled list of the fifty U.S. states. Once two minutes elapses, participants will automatically progress to a cued recall task, in which they will be presented with each of the 63 cues that had previously been judged as cue-target pairs. Participants will be asked to complete each word pair with the appropriate target word, based on the available cue word. Presentation of these pairs will be randomized, and participants will be informed that there is no penalty for guessing. The Qualtrics surveys are uploaded at https://osf.io/j7qtc/.

Results

First, the results from the recall section will be coded as zero for incorrect responses and one for correct responses. NA will be used to denote missing responses from participants who did not complete the recall section. Responses that are words instead of numbers in the judgment phase will be deleted and treated as missing data. Data will then be screened for out of range judgment responses (i.e., responses greater than 100). Recall and judgment scores will be screened for outliers using Mahalanobis distance at $p < .001$ (Tabachnick & Fidell, 2012); all outliers will be removed. Next, multicollinearity between predictor variables will be measured with Pearson correlations. It is expected that the meausres word length will correlate highly, as words with a higher number of characters are natuarlly more likely to contain more syllables, morphemes, or phonemes. Predictor variables will be excluded from the analysis if correlations exceed $r > .60$. Finally, data will be screened for assumptions of normality, linearity, homogeneity, and homoscedasticity. Descriptive statistics of mean judgment and recall scores will be reported for each judgment condition.

Multilevel modeling will then be used to analyze the data (Gelman, 2006) to control for the nested structure of the data using the *nlme* library. Each participant's judgment and recall ratings will be treated as a data point, using participants as a nested random intercept factor. As part of our replication, we will reanalyze these new stimuli using COS, FSG, LSA, and their interaction to predict judgments and recall separately as the dependent variables. Just as in Maxwell and Buchanan (2018), judgment condition will be used as a control variable. Variables will be mean centered prior to analysis to control for multicollinearity. If a significant three-way interaction occurs, simple slopes analyses will be used to explore that interaction. We will examine low (-1SD), average (mean), and high (+1SD) COS values for two-way interactions of FSG and LSA. If these values are significant, LSA will be further broken into low, average, and high simple slopes to examine FSG. α is set to .05 for all analyses. We predict that the interactions found previously will replicate with the new set of stimuli.

A second set of analyses will be performed using the Maxwell and Buchanan (2018) stimuli set and this new stimuli set combined, examining the hypothesis of interactive networks after controlling for base word norm information. Stimuli sets from both studies will be combined to create a larger range of stimuli and values across normed information. These neighborhood norms will be introduced into each model in steps, after controlling for judgment condition. Initially, base word norms will be added, followed by lexical information, rated properties, and norms measuring neighborhood connections, as described in the introduction and methods. Each set of variables will be used to predict the dependent variables of judgment and recall, again as a multilevel model. Each variable will be discussed in the step of the analysis it was entered. We expect that many of these variables will significantly predict judgments and recall, but do not predict which ones in particular. Last, the interaction of network norms will be added to the model with the prediction that the interaction of COS, FSG, and LSA may be significant, even after controlling for single concept information.

This analysis plan was pre-registered as part of the Pre-Registration Challenge through the Open Science Foundation and may be found at: https://osf.io/24sp9/. This manuscript was written in *R* markdown using the *papaja* package by Aust and Barth (2017).

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. doi:10.3758/BF03194088

Aust, F., & Barth, M. (2017). papaja: Create APA manuscripts with R Markdown. Retrieved from https://github.com/crsh/papaja

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

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & Maas, H. L. J. van der. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069–1077. doi:10.1177/0956797616647519

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

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., & 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., 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., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. doi:10.3758/s13428-013-0403-5

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

Coltheart, M., Davelaar, E., Jonasson, T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance vi* (pp. 535–555). Hillsdale, NJ: Earlbaum.

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

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language*, *33*(4), 545–565. doi:10.1006/jmla.1994.1026

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. doi:10.1111/2041-210X.12504

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

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

Maxwell, N. P., & Buchanan, E. M. (2018). *Modeling memory: Exploring the relationship between word overlap and single word norms when predicting relatedness judgments and retrieval*. Retrieved from http://osf.io/j7qtc

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/BF03192726

Nelson, D. L., & Schreiber, T. A. (1992). Word concreteness and word structure as independent determinants of recall. *Journal of Memory and Language*, *31*(2), 237–260. doi:10.1016/0749-596X(92)90013-N

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*, *36*(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

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely acurate at predicting subsequent recall: The delayed-JOL effect. *Psychological Science*, *2*(4), 267–270. doi:10.1111/j.1467-9280.1991.tb00147.x

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. doi:10.1006/jmla.1997.2516

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

Schreiber, T. A., & Nelson, D. L. (1998). The relation between feelings of knowing and the number of neighboring concepts linked to the test cue. *Memory & Cognition*, *26*(5), 869–83. doi:10.3758/BF03201170

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. (2012). *Using Multivariate Statistics* (6th ed.). Boston, MA: Pearson.

Toglia, M. P. (2009). Withstanding the test of time: The 1978 semantic word norms. *Behavior Research Methods*, *41*(2), 531–533. doi:10.3758/BRM.41.2.531

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

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. doi:10.3758/s13428-012-0314-x

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

Table 1

Summary Statistics of Single Word Norms for Cue Items

Variable	Citation	Mean	SD	Min	Max
QSS	Nelson et al., 2004	14.76	4.45	4.00	24.00
Concreteness	Nelson et al., 2004	5.35	1.00	1.98	7.00
HAL Frequency	Lund and Burgess, 1996	9.34	1.67	6.26	13.39
SUBTLEX Frequency	Brysbaert and New, 2009	3.15	0.74	1.76	5.20
Length	Buchanan et al., 2013	4.90	1.50	3.00	10.00
Ortho N	Buchanan et al., 2013	7.44	5.91	0.00	19.00
Phono N	Buchanan et al., 2013	19.00	15.11	0.00	51.00
Phonemes	Buchanan et al., 2013	3.94	1.39	2.00	9.00
Syllables	Buchanan et al., 2013	1.35	0.60	1.00	3.00
Morphemes	Buchanan et al., 2013	1.10	0.30	1.00	2.00
AOA	Kuperman et al., 2012	5.15	1.53	2.47	8.50
Valence	Warriner et al., 2013	5.77	1.23	1.91	7.72
Imageability	Toglia and Battig, 1978	5.52	0.68	3.22	6.61
Familiarity	Toglia and Battig, 1978	6.17	0.28	5.58	6.75
FSS	Buchanan et al., 2013	17.37	11.61	5.00	48.00
COSC	Buchanan et al., 2013	87.25	71.33	3.00	347.00

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

Table 2

Summary Statistics of Single Word Norms for Target Items

Variable	Citation	Mean	SD	Min	Max
TSS	Nelson et al., 2004	15.44	4.86	5.00	26.00
Concreteness	Nelson et al., 2004	5.40	1.01	1.28	7.00
HAL Frequency	Lund and Burgess, 1996	9.78	1.52	6.05	13.03
SUBTLEX Frequency	Brysbaert and New, 2009	3.34	0.64	1.59	4.74
Length	Buchanan et al., 2013	4.62	1.67	3.00	10.00
Ortho N	Buchanan et al., 2013	9.02	7.77	0.00	29.00
Phono N	Buchanan et al., 2013	21.51	16.71	0.00	59.00
Phonemes	Buchanan et al., 2013	3.70	1.50	1.00	10.00
Syllables	Buchanan et al., 2013	1.25	0.54	1.00	3.00
Morphemes	Buchanan et al., 2013	1.05	0.21	1.00	2.00
AOA	Kuperman et al., 2012	4.87	1.56	2.50	9.16
Valence	Warriner et al., 2013	5.84	1.27	1.95	7.89
Imageability	Toglia and Battig, 1978	5.50	0.71	2.95	6.43
Familiarity	Toglia and Battig, 1978	6.28	0.32	5.19	6.85
FSS	Buchanan et al., 2013	16.70	11.62	5.00	54.00
COSC	Buchanan et al., 2013	91.71	79.52	3.00	322.00

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

Table 3

Summary Statistics for Network Norms

Variable	Citation	Mean	SD	Min	Max
FSG	Nelson, McEvoy, and Schrieber, 2004	0.13	0.19	0.01	0.83
COS	Maki, McKinley, and Thompson, 2004	0.42	0.29	0.00	0.84
LSA	Landauer and Dumais, 1997	0.38	0.20	0.05	0.88

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

Table 4

Summary Statistics for Stimuli

Variable	COS Low			COS Average			COS High		
	N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
COS	21	.058	.070	21	.445	.081	21	.752	.047
FSG Low	21	.050	.044	19	.069	.073	16	.098	.088
FSG Average	NA	NA	NA	2	.623	.033	4	.542	.066
FSG High	NA	NA	NA	NA	NA	NA	1	.828	NA
LSA Low	17	.182	.074	9	.215	.070	4	.192	.053
LSA Average	3	.466	.140	10	.489	.087	14	.515	.079
LSA High	1	.717	NA	2	.685	.025	3	.772	.106

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