9-27-2023

A Psycholinguistic Analysis of Neuropsychological Tests

Brette Lansue
University of Windsor

A Psycholinguistic Analysis of Neuropsychological Tests

By

Brette E. Lansue

A Dissertation
Submitted to the Faculty of Graduate Studies
through the Department of Psychology
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy
at the University of Windsor

Windsor, Ontario, Canada

2023

A Psycholinguistic Analysis of Neuropsychological Tests

by

Brette E. Lansue

APPROVED BY:

V. Taler, External Examiner
University of Ottawa

P. Weir
Department of Kinesiology

D. Jackson
Department of Psychology

C. Miller
Department of Psychology

L. Buchanan, Advisor
Department of Psychology

August 9, 2023

DECLARATION OF ORIGINALITY

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

Neuropsychological assessment often depends on language-based measures of cognitive functioning and proper diagnosis of certain disorders relies on patterns of impairment in language and memory on these measures. The current project was motivated by the relative lack of literature integrating psycholinguistic experimental findings and clinical neuropsychological research. It has been well documented that word-level characteristics impact language processing and memory. Therefore, it is critical that neuropsychologists begin to understand how the measures currently in use can be confounded by the underlying lexical and semantic characteristics of the stimuli and how, if used properly, those characteristics could aid in diagnostic specificity. Results from Studies 1 and 2 demonstrated that 1) age of acquisition, emotional valence, semantic neighborhood density, and imageability predicted better recall of items from neuropsychological tests in healthy participants, and 2) only one of the ten test lists examined adequately controlled for these influential variables. Results from Study 3 demonstrated several ways in which common semantic fluency categories differ from each other, including in overall category size, number of correct responses and set-loss errors produced, and across several lexical and semantic variables. Study 4 presented and evaluated four potential remedial options for a category switching task that move closer toward structure equality. Overall, the results of this dissertation show that the neuropsychological tests examined herein are not adequately constructed from a psycholinguistic perspective and that clinicians could be missing clinically relevant data by ignoring psycholinguistic contributions to performance.

DEDICATION

To my nephew, Jack.

You were born at a time when I needed a light in life; thank you for being that light. Hopefully one day, when you're older than 2, you will think this is kind of cool!

ACKNOWLEDGEMENTS

To Lori, there are no words that could adequately express my gratitude to you. Your guidance and support have carried me through some of my most challenging years yet. I am confident I would not have made it here without you.

To Drs. Weir, Miller, and Jackson, thank you for your valued input and thoughtful questions and for investing your time and knowledge in my research. You have all been so supportive throughout the development of this project and more broadly, throughout my academic career.

To Dr. Taler, thank you for sharing your time (even while on vacation with a 16-hour time difference), enthusiasm, and unique expertise to support the final stage of this journey.

To Owen Bortolin (not to be confused with Bortolon), thank you for spending countless hours collecting and coding data with me. You helped me meet insane deadlines and I am so grateful for your company and dedication to this project!

To my lab mates, peers, and colleagues, thank you for your constant support in the form of much needed distraction, potlucks, statistical consulting, and shared misery. I could not imagine a better group of people to share so many years with.

To my family and partner, thank you for your unwavering encouragement, patience, and support throughout this endeavor. You never question my middle of the night alarms to meet deadlines, anxious phone calls, and the never-ending phrase "my dissertation is almost done." Thank you for reminding me that I'm almost there (and that I need to eat). I could not have done this without you! PS. We are one step closer to the family cottage.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1:

INTRODUCTION TO IMPORTANT CONCEPTS & THE PROJECT

Neuropsychological Assessment and Test Development

Neuropsychological assessment is a specialized form of psychological testing focusing on the relationship between brain functions and behavior (Strauss et al., 2006). It can be a valuable tool for understanding the cognitive and emotional functioning of individuals with neurological, neurodevelopmental, and psychiatric disorders and can help guide treatment planning and rehabilitation. Neuropsychological assessments typically involve a battery of tests that assess several domains of cognitive functioning, such as intellect, language, attention, memory, and executive functioning. These neuropsychological tests evaluate how well an individual can perform specific tasks that rely on these cognitive functions (Lezak et al., 2004). For example, a test of attention might require the patient to sustain their focus on a specific task for a prolonged period of time, while a test of memory might ask the patient to recall a list of words or reproduce a design after a delay.

Neuropsychological assessments are used to diagnose a wide range of conditions and can help determine brain function and lateralization of damage or abilities. Prior to the development of neuroimaging, neuropsychologists were often called upon to use their tools to locate lesions and damage within the brain for surgical intervention purposes. Over time, the goal of neuropsychology has adapted to integrate the direct measurement of cortical functioning with findings from imaging to better understand how neuropathology affects daily functioning and behavior (Lezak et al., 2004). Obtaining a comprehensive neuropsychological profile aids clinicians in the identification of relevant strengths and weaknesses and the development of a treatment plan that addresses the patient's unique needs (Vakil, 2012).

To adequately measure cognitive functioning, several tests have been designed and developed following psychometric theory (Nunnally & Bernstein, 1994); specifically, Classical Test Theory (CTT; Crocker & Algina, 1986) and Modern Test Theory (MTT, also known as Item Response Theory [IRT]; Embretson & Reise, 2000). The framework addresses goals related to reliability and validity through various analyses, including item analysis, factor analysis, reliability analysis, and validity analysis (Crocker & Algina, 1986).

Classical Test Theory (CTT) is based on the idea that an individual's score on a psychological test has two underlying components: their *true score* and *error* (Crocker & Algina, 1986). The *true score* represents the individual's ability or trait, whereas the *error* represents extraneous factors that affect test performance in unpredictable ways, including measurement error, ambiguous test stimuli, or interfering behaviors (e.g., test anxiety, distraction, fatigue; Crocker & Algina, 1986). Modern Test Theory (MTT) is a similar evaluative theoretical framework for psychological tests and measures that has become more popular over time. While CTT focuses on how reliable test scores are, MTT focuses on the relationship between an individual's test score and the characteristics of the specific test items (Embretson & Reise, 2000). The distinguishing assumption of MTT is that the probability of getting any given item correct relies on the individual's true ability and the item's difficulty. MTT can be used to evaluate individual test items, but because items should gradually increase in difficulty, it can also allow us to estimate an individual's underlying ability level based on their pattern of responses (Embretson & Reise, 2000). Although both theories have different assumptions and approaches to measuring psychological traits or attributes, CTT and MTT agree that an individual's score on any given test or item encompasses both the individual's true ability and an additional error component that influences performance on the test.

The Boston Process Approach

The Boston Process Approach is a method for testing and interpreting cognitive functions in neuropsychological assessments. Edith Kaplan developed the approach in the 1970s in attempt to better capture the entire process of how a patient received a particular score on a cognitive test (Libon et al., 2013). Specifically, rather than interpreting a single final score, the Boston Process Approach promotes the interpretation of a patient's strategies, errors, and response times. For example, suppose that three patients of the same age (e.g., 50) all receive a raw score of 18 on a task that requires them to reproduce a picture with three-dimensional blocks within a specific time limit. One patient was generally able to accurately reproduce the designs long before the time limit surpassed but had a single error on the last two attempted designs resulting in discontinuation. A second patient was able to accurately reproduce the designs with no errors but was beyond the time limit on the last two attempted designs resulting in discontinuation. A third patient was unable to place more than one or two blocks in their correct position or orientation on the last two attempted designs resulting in discontinuation. Once appropriately normed to the standardization sample, all three patients would obtain a scaled score of five; however, it would be negligent to interpret all three patients' scores in the same manner, given what we know about how they each earned that score. In a crude analysis for the sake of the example, patient one may have deficits in attention to detail, patient two may have a processing speed deficit or motor deficits leading to slower movement on the task, and patient three may have a true visual-motor coordination deficit. Through the Boston Process Approach, process scores were developed for certain tests to aid clinicians in breaking down these final scores to better understand how a patient got there (Libon et al., 2013).

Using the Boston Process Approach to interpret neuropsychological tests has been shown to effectively detect cognitive changes and functioning beyond that of the typical score analysis. Using process scores developed through this approach, DeLuca et al. (2004) were able to detect subtle deficits in individuals with multiple sclerosis that were not originally apparent. The fine-grained analysis of their patients' strategies and error types was able to provide additional information that was beneficial to their individualized rehabilitation plans. Another study used the Boston Process Approach to evaluate and inform a cognitive training program, as the process scores were more sensitive to subtle changes in participants' cognitive functioning following the training (McAvinue et al., 2012).

Overall, the Boston Process Approach has been deemed invaluable in the understanding of patients' unique cognitive assets and deficits. It provides a detailed analysis that can inform a clinician's conceptualization and recommendations beyond what is typically allowed for in traditional neuropsychological test scores.

The Current Project

In neuropsychological assessment, words are often used as stimuli to test underlying abilities. For example, words are used to test verbal memory (e.g., list learning tasks; California Verbal Learning Test – Third Edition [CVLT-3; Delis et al., 2017]), confrontation naming (e.g., Boston Naming Test [BNT; Kaplan et al., 1983]), verbal fluency (e.g., Delis-Kaplan Executive Functioning System [DKEFS; Delis et al., 2001]), estimating premorbid functioning (e.g., word reading tests; Test of Premorbid Functioning [TOPF; Wechsler, 2011]), and verbal comprehension (e.g., similarities, vocabulary; Wechsler Adult Intelligence Scale – Fourth edition [WAIS-IV; Wechsler, 2008]). Some of these tests have items that intentionally increase in difficulty to better understand the patient's ability level (e.g., Boston Naming Test, Test of

Premorbid Functioning). Others assume that the difficulty of the items is equal. For example, in list learning tests, the patients are presented with words over several trials and are asked to repeat the items after a delay; thus, each word is taken to be as easy or difficult to remember as the next word on the list. Similarly, in verbal fluency tests, patients are asked to produce as many words as they can that begin with a specific letter (phonemic fluency), or that belong to a specific category (semantic fluency) within a time limit, and raw scores for multiple letters or categories are pooled to obtain an overall score; thus, for these tests, rather than the individual words, the individual categories are considered to lend themselves equally to the generation of items.

The current project focuses on neuropsychological tests that assume equal difficulty across items or tasks. Many of these tests did not take psycholinguistic variables (i.e., lexical and semantic characteristics of words or categories) into account during development. As Chapters 2 and 3 will outline, there are documented effects of such variables on experimental memory and fluency tests; thus, the underlying structure of commonly used list learning and fluency tests could be causing a violation of the assumption of equality between items and categories, making interpretation convoluted. At the very least, by ignoring these important item-level differences during interpretation, we risk missing important diagnostic information. The purpose of the current project is to explore whether the lexical and semantic effects that are found in experimental research cross over into our clinical data using well-established and commonly used neuropsychological tests.

CHAPTER 2:
EFFECTS OF LEXICAL AND SEMANTIC VARIABLES ON MEMORY TASKS

Words are believed to exist in the mind as memory representations stored in what are referred to as mental lexicons, which are analogous to mental dictionaries (Schriefers, 1992; Sommers, 1996; Elman, 2004; Libben & Jarema, 2002; Coltheart et al., 1977). There are three lexicons that are critical for processing written or spoken language: an orthographic lexicon (i.e., the visual features of a specific word), a phonological lexicon (i.e., the auditory features of a word), and a semantic system (i.e., the meaning-related features of a word; Oldfield, 1996; Treisman, 1960; Halderman & Chiarello, 2005; Buchanan et al., 2001).

Many early models of word recognition were localist in nature, such that they suggested that each word in our lexicon was represented by a specific node, and recognition occurred when this node became activated (e.g., Forster, 1976; Morton, 1969). It is now generally accepted in the psycholinguistic literature that distributed mechanisms play a more significant role. Distributed models (e.g., Seidenberger & McClelland, 1989) posit that each word is represented by networks of activation rather than by a single node. Therefore, word recognition occurs when this network reaches an activation threshold across all orthographic, phonological, and semantic characteristics of the word (Harm & Seidenberg, 2004). There is consistent agreement that various word properties (e.g., word frequency, orthographic variables, semantic richness) influence language processing (see Yap & Balota, 2015 for review) and, to a lesser extent, memory for words. Given this influence, Chapter 2 aims to complete a psycholinguistic analysis of the items in commonly used neuropsychological memory tests.

Influence of Lexical and Semantic Variables on Memory

Several variables have been implicated in influencing both memory and language processing. The following section outlines the effects of these psycholinguistic (lexical and semantic) variables on memory for words.

Word Frequency

The frequency with which a word occurs within a language (i.e., word frequency) has long been shown to affect language processing (Cattell, 1886). Words that appear more commonly in print are recognized faster and more accurately than their less frequent counterparts (Brysbaert et al., 2016; Murray & Forster, 2004; Preston, 1935), known as a word frequency effect. Word frequency has remained an influential variable in language processing, even when several other variables are controlled, and the effect is powerful and robust (MacLeod & Kampe, 1996). Given the consistency in language processing literature, it was thought that word frequency must be important for a word's representation in memory (MacLeod & Kampe, 1996).

The effect of word frequency on memory, however, is much less clear than that on language processing, and findings typically depend on task demands. First, there is a strong and consistent mirror effect of word frequency produced by recognition memory tasks, with low frequency words being recognized better than high frequency words. The pattern of results is considered a mirror effect because low frequency words are often associated with higher hits and lower false alarm rates compared to high frequency words (Glanzer & Adams, 1985, 1990; Malmberg et al., 2004). The low frequency advantage in recognition memory is quite robust for both pure lists (i.e., lists made up solely of low frequency or high frequency words) and mixed lists (i.e., lists consisting of both low frequency and high frequency words; Criss & Malmberg, 2008; Dorfman & Glanzer, 1988; Estes & Maddox, 2002; Glanzer & Adams, 1985; Gorman, 1961; Heathcote et al., 2006; Malmberg et al., 2002). Some research has found that although there is a low frequency advantage at retrieval, there may be a low frequency disadvantage at encoding. Through a series of experiments, Diana and Reder (2006) found that although low frequency words required more attention to properly encode, those that were successfully encoded were more easily recognized than their high frequency counterparts. When Diana and Reder (2006) forced resource-limited encoding conditions, low frequency words showed a greater reduction in recognition compared to high frequency words. Further, Popov and Reder

(2020) conducted a comprehensive study that found that memory for any given item is affected by the strength (or word frequency) of preceding items on the list. Specifically, memory for an item was better when the preceding item was high compared to low frequency, and this effect is continuous, meaning more high frequency words on a list lead to better memory. The authors suggest this is due to fewer resources being needed to process the preceding item (Popov and Reder, 2020). Thus, the composition of the entire list may have trickle-down effects on each individual word.

Interestingly and in contrast to the findings for recognition memory, high frequency words have been found to have an advantage over low frequency words on free recall tasks (Balota & Neely, 1980; Deese, 1960; DeLosh & Mcdaniel, 1996; Gillund & Shiffrin, 1984; Gregg et al., 1980; Ward et al., 2003). Further, there is a learning advantage to high frequency words, as they are learned at a faster rate compared to low frequency words (Sumby, 1963). However, this learning and memory advantage appears to be consistent only when measured with pure lists (MacLeod & Kampe, 1996; Watkins et al., 2000). Some research has found the high frequency advantage holds true for recall of mixed lists (Balota & Neely, 1980; Hicks et al., 2005), while others have found the opposite (low frequency advantage) on these mixed lists (DeLosh & McDaniel, 1996; Merritt et al., 2006; Ozubko & Joordens, 2007 [with random lists]), and finally, some studies have found no effect of word frequency at all on recall (May et al., 1979; Ozubko & Joordens, 2007 [with alternating lists]; Ward et al., 2003; Watkins et al., 2000).

In response to the ambiguities in the literature, Lohnas and Kahana (2013) used parametric tests and a wide range of word frequencies to examine the relationship between frequency in mixed lists and memory performance (both recognition and recall). Importantly, the authors found a robust mirror effect (increased hit rate and decreased false alarm rate for low frequency words) for recognition memory. However, Lohnas and Kahana (2013) provided a novel finding relating to recall; they found a significant advantage for both low and high frequency words as compared to those with midfrequency (i.e., a U-shaped relationship between frequency and recall performance). Considering the broad range of frequencies used in this study, these results may help to explain the inconsistencies found in the previous literature, and the authors stress the danger of using frequency as a categorical variable rather than continuous (Lohnas & Kahana, 2013).

Theories attempting to explain the mirror effect for word frequency in recognition memory typically fall into single-process or dual-process theories. Single-process theories focus on the distinct lexical characteristics (i.e., orthographic or semantic features) of low frequency words making them more likely to be correctly recognized or correctly rejected (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). In contrast, dual-process theories propose that recognition memory tests involve two cognitive processes working in tandem, recollection (i.e., contextual information related to the encoding of the item; can include how the item was presented [visually vs. auditorily], temporal information [first vs. second], etc.) and familiarity (i.e., the strength of the memory store). As one might expect, recollection is a slower process compared to familiarity (Jacoby, 1991; Joordens & Hockley, 2000; Reder et al., 2000). Two dual-process theories have been proposed to account for the mirror effect of word frequency. Joordens and Hockley (2000) postulate that recollection depends heavily on distinctiveness, which relies on the frequency of pre-experimental experiences. Thus, because low frequency words have been encountered less than the high frequency words prior to the experiment, they are more distinct, and that distinctiveness imparts a memory advantage. Whereas Reder and colleagues (Reder et al., 2000; Reder et al., 2007) argue that recollection is a function of the activation of "episode nodes," consisting of activation of both the specific word accompanied by the context in which it was studied. This theory proposes that recollection relies heavily on the context of the source of the information to be recognized. As such, low frequency words again have less competing pre-experimental information making the connection between word and source stronger, ultimately giving rise to the low frequency recollection advantage (Reder et al., 2000; Reder et al., 2007). Both dual-process models account for the higher false alarm rate for high frequency words similarly. Participants will necessarily come into experiments with more baseline familiarity to high frequency words than low frequency words, making them much more difficult to distinguish from new items at test (Coane et al., 2011).

Theories attempting to explain the effects of word frequency on recall tasks have generally landed on a redintegration process. Originally, early research theorized that the high frequency advantage in recall tasks was due to differences in speech duration, as low frequency words took longer to say (even when word length was controlled) and thus longer to rehearse than high frequency words (Wright, 1979). However, this idea was debunked using articulation suppression (i.e., repeating irrelevant information to inhibit the phonological rehearsal of visual stimuli). Specifically, Tehan and Humphreys (1988) and Gregg et al. (1989) found that although high frequency words had a higher speech rate (i.e., spoken faster) and did result in better memory span, this difference remained under articulatory suppression. Further supporting the idea that word frequency differences were likely due to influences outside of the phonological loop, Hulme et al. (1997) found that memory span was greater for words compared to nonwords, even when controlling for speech duration. The authors proposed a pattern completion or redintegration theory, such that the advantage of words compared to nonwords and of high compared to low frequency words was related to a previously stored representation of the word in long-term memory. More specifically, the pre-experimental knowledge of the word aids participants via a pattern completion or restoration process on partially decayed memory traces, which ultimately supports retrieval (Hulme et al., 1997).

Orthographic & Phonological Neighborhood Size

Lexical characteristics also provide information regarding the distinctiveness of words. As alluded to in the discussion of word frequency effects, the distinctiveness of a word can add complexity to our understanding of the memorability of words. Orthographic and phonological neighborhood size provide measures of how similar a specific word is to other words. An orthographic neighborhood refers to the words that are created by changing a single letter of the target word (e.g., *cat*, *hat*, *cot*, *cab*; Coltheart et al., 1977). Similarly, the auditory analog is the phonological neighborhood, which refers to the words that are created by changing a single phoneme of the target word (e.g., *cat, caught, chat;* Cortese et al., 2010; Cortese et al., 2014; Goh & Pisoni, 2003). The structure of a word's neighborhood can be characterized in terms of neighborhood density, which describes the number of orthographically or phonologically similar words in a target word's neighborhood (i.e., dense neighborhoods have more similar words than sparse neighborhoods; Luce & Pisoni, 1998). The structure of these neighborhoods can also be characterized by neighborhood frequency, which describes the average frequency of the words in a target word's neighborhood (i.e., high frequency neighborhoods have mostly high frequency words, low frequency neighborhoods have mostly low frequency words; Goh & Pisoni, 2003). Goh and Pisoni (2003) suggest that these two dimensions can be used to classify words as either "easy" or "hard" to recognize based on competition. The authors pose that words that have sparser neighborhoods and are higher frequency relative to their neighbors stand out more, making them easier to recognize (i.e., a target word with a sparse and low frequency neighborhood). In contrast, target words with dense and high frequency neighborhoods tend to be more difficult to recognize because they get confused with or "swamped by" their many similar neighbors (Goh & Pisoni, 2003).

Lexical competition has been shown to affect word recognition across several experiments, including naming and lexical decision tasks (Cluff & Luce, 1990; Luce et al., 1990), priming tasks (Goldinger et al., 1989), and perceptual identification tasks (Luce & Pisoni, 1998). These results suggest that words do compete with other words in the lexicon for recognition (Luce & Pisoni, 1998; Vitevitch & Luce, 1998). Using the concept of "easy" and "hard" words described above, Goh and Pisoni (2003) demonstrated that lexical competition, based on the structure of lexical neighborhoods (orthographic and frequency neighborhoods), impacted the immediate memory span for words (i.e., participants could remember more "easy" words in a row than "hard" words). Cortese et al. (2004) found similar effects on both recognition and recall using a measure of both phonological and orthographic neighborhood size. Across four experiments, words with fewer phonological and orthographic neighbors were recalled better than those with more. Glanc and Greene (2007) also replicated a low orthographic neighborhood size advantage in recognition memory across several tasks with differing demands (i.e., a standard *yes/no* recognition task, a forced-choice recognition task, and *remember/know* judgment task). Results showed a mirror effect, such that words with smaller orthographic neighborhood size had a higher hit rate and a lower false alarm rate compared to words with larger orthographic neighborhood size (Glanc & Greene, 2007). The authors also demonstrated a novel finding, where the stable mirror effect disappeared with the introduction of a semantic processing task. These results have been explained using an interference framework, whereby target words that have several phonological and orthographic neighbors are more likely to experience interference due to the coactivation of their similar neighbors (Cortese et al., 2004; Glanc and Green, 2007). Glanc and Green (2007) argue that their novel finding related to the semantic processing task provides support for the idea that the standard mirror effect is due to orthographic distinctiveness, as the task pulled attention away from the orthographic features of the target words by requiring a deeper level of processing. A recent study by Ballot et al. (2021) examined how semantic and orthographic information differentially affect memory performance, both in recognition and recall tasks. The authors used imageability, or the ease with which a word creates a visual image in the mind, as a proxy for semantic information; this variable will be discussed in more depth below. Results showed that while imageability consistently facilitated memory regardless of the task demands, orthographic neighborhood only had an effect on the recognition task, not the recall task. Further, Ballot and colleagues (2021) found that the effect of orthographic neighborhood was dependent on imageability, suggesting that when semantic information is readily available (e.g., high-imageability), semantic information is used, and when it is not (e.g., low-imageability), orthographic distinctiveness is used.

Interestingly, some research has found conflicting effects. Roodenrys et al. (2002) found that target word frequency, phonological neighborhood density, and neighborhood frequency supported recall; specifically, words that had higher word frequency, denser neighborhoods, and higher neighborhood frequency were more likely to be recalled. However, neighborhood density and neighborhood frequency also lead to a higher likelihood that the neighbor of a target word would be incorrectly recalled (i.e., an intrusion). Thus, Roodenrys and colleagues' (2002) results suggest both an effect of lexical competition and a recall advantage for dense neighborhoods. Roodenrys (2009) proposes that the differences in findings are likely due to different constraints put on stimulus sets and on the operationalization of variables.

Word Length

Word length (i.e., the number of letters in a word) can also influence memory for words. An early study by Baddeley and colleagues (1975) demonstrated a word length effect on recall memory, such that lists with shorter words were better remembered than lists with longer words. This effect has been found in both serial recall tasks and free recall tasks (Baddeley et al., 1975; Russo & Grammatopoulou, 2003; Tehan & Tolan, 2007; Bhatarah et al., 2009). However, when lists have mixed items (i.e., consisting of both short and long words), the findings are less consistent. Hulme and colleagues (2004) conducted two experiments to examine the word length effect on memory by comparing recall for words in pure lists (i.e., all short or all long words) and mixed lists. The authors found that although the word length effect held true for pure lists, it was eliminated for the mixed lists, such that both long and short words were remembered as well as the short words in a pure list (Hulme et al., 2004). Further, Jalbert and colleagues (2011) found 1) words with larger orthographic neighborhoods were recalled better than words with smaller orthographic neighborhoods and 2) when controlling for orthographic neighborhood size, the word length effect was eliminated. Moreover, Katkov et al. (2014) found a typical word length effect (i.e., short words were better remembered than long) for pure lists, and a mirrored effect (i.e., long words were better remembered than short) was found for the mixed lists.

There are broadly two prevailing explanations of the word length effect in the literature. The first is a time-based account, which maps directly onto Baddeley's (1986) theory of working memory, such that verbal retention is dependent upon the phonological loop and rehearsal ability. Specifically, lists of short words are less susceptible to time-based decay because participants can rehearse more of them in a short period of time (Burgess & Hitch, 1999; Page & Norris, 1998). One assumption of a time-based account is that the word length effect will remain stable, even when words are controlled on all other complexity variables (Lovatt et al., 2000; Neath et al., 2003); however, this does not appear to be the case. Although those studies mentioned above found effects of word length, several other studies using different stimulus sets found no effect or even a reversed effect (Lovatt et al., 2000; Service, 1998; Neath et al., 2003). These conflicting findings have given rise to a complexity-based account, which suggests that lists of longer words are more difficult to remember due to their increased phonological complexity, and there is a limit on the amount of phonological information that can be kept in a retrievable manner (Caplan et al., 1992; Neath et al., 2003). Hulme et al. (2004) propose the latter explanation with an added distinctiveness component to explain the difference in effects when using pure versus mixed lists. Specifically, the authors suggest that distinctiveness aids memory, where the more distinct each word on a list is, the easier the words will be to retrieve (Hulme et al., 2004). Thus, although a word length effect in memory was once thought to be a staple of memory research, more recent literature has begun to disentangle a true word length effect from a complexity effect.

Age of Acquisition

Another lexical variable to consider when attempting to understand the processing and memorability of words is the chronological age at which a target word is typically learned (i.e., age of acquisition). Language processing literature has consistently documented that words with earlier age of acquisition are processed more efficiently than those with later age of acquisition (Gilhooly & Watson, 1981; Johnston & Barry, 2006; Juhasz et al., 2019). Participants have faster reaction times (RTs) for words acquired earlier on both picture naming (Meschyan & Hernandez, 2002; Pérez, 2007) and word naming (Brysbaert & Cortese, 2011; Cortese & Khanna, 2007) tasks. Similarly, words with early age of acquisition are processed faster than those with later age of acquisition in lexical decision tasks, even after controlling for several other lexical variables (e.g., word frequency, word length, neighborhood size; Johnson & Barry, 2006). As is the case with several of the variables discussed above, the effects of age of acquisition on memory tasks are much more ambiguous.

There are several studies in which age of acquisition has been found to affect memory, while others have found contradictory results. For example, Coltheart and Winograd (1986) found null effects of age of acquisition on recognition memory for both pure lists (i.e., lists with only early age of acquisition or only late age of acquisition words) and mixed lists (i.e., lists containing both early and late age of acquisition words) after controlling for frequency, imagery, and word length. Similarly, Gilhooly and Gilhooly (1979) used regression analysis and found no effect of age of acquisition on either recognition or recall memory using mixed lists. Likewise, Dewhurst and colleagues (1998) also showed no effect of age of acquisition in recall of pure lists.

Conversely, Dewhurst et al. (1998) demonstrated that age of acquisition had an effect on recall with mixed lists, such that words acquired later were better remembered than those acquired earlier. Morris (1981) used a regression analysis and found the same effect on recall of mixed lists. Morris (1981) attributed the differing results from Gilhooly and Gilhooly (1979) to when specific lexical variables were entered into the regression. In contrast, Almond and Morrison (2014) found an opposite effect of age of acquisition (i.e., early-acquired words > late-acquired words) for recall of pure lists. Results of studies using recognition memory tasks have more consistently demonstrated an advantage of words with late age of acquisition compared to early age of acquisition (Dewhurst et al., 1998; Cortese et al., 2010; Cortese et al., 2015).

In a series of experiments, Macmillan and colleagues (2021) explored the effects of age of acquisition on recognition and recall (both serial and free recall) memory using norms from newly developed databases. The authors found a strong late-age of acquisition advantage in both pure and mixed lists for recognition memory, but no effect of age of acquisition in either type of list for serial or free recall (Macmillan et al., 2021). Taken together, it appears that age of acquisition differentially affects memory based on retrieval context, where words acquired later have an advantage in recognition but not recall tasks.

Semantic Richness Variables

Several psycholinguistic models agree on the notion that the structure of semantic space is composed of organizational influences that can be defined (Buchanan et al., 2001). However, there is debate regarding exactly which principles govern the organizational structure of the space. Specifically, there are generally two schools of thought: object-based models and language-based models. Object-based models (also known as feature- or category-based models) suggest that semantic information is conceptually organized based on the similarity between the objects' physical properties or attributes – this similarity can come from feature overlap or category membership. For example, *dog* and *cat* are close semantic neighbors because they share many physical features (i.e., fur, a tail, walk on four legs, etc.) and because they come from the same category of objects (i.e., household pets).

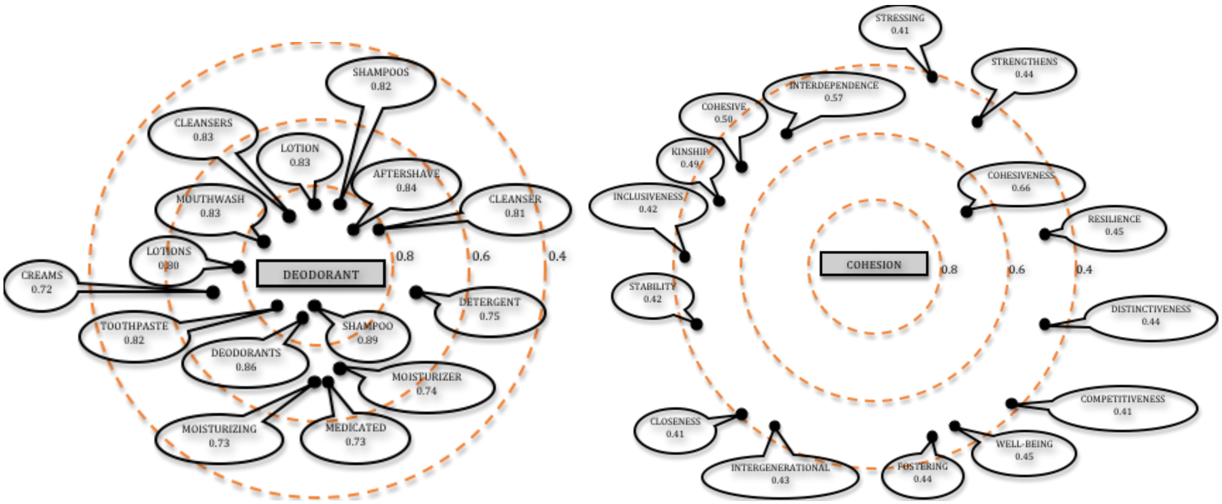
In contrast, language-based models (also known as association-based or co-occurrence models) propose that the organization of semantic information is based on how objects occur in language; thus, concepts are classified as semantically related based on the statistical co-occurrence of words. Based on these types of models, *dog* and *cat* are close semantic neighbors because they are often presented in similar contexts based on large samples of text (e.g., global co-occurrence; Burgess & Lund, 2000; Buchanan et al., 2001; Landauer & Dumais, 1997). A particular strength of the language-based view is that it allows physically unrelated concepts or concepts that do not have featural overlap or share category membership to be semantic neighbors. Buchanan and colleagues (2001) provide the example of *cat* and *scratch* to illustrate this advantage. These two concepts do not share any physical features and, thus, according to the object-based view, would not be considered semantically related, nor would they show priming effects; however, based on our experience with language, we recognize this to be untrue. Indeed, semantic-priming research has demonstrated facilitative effects for these types of word pairs that do not share overlapping features, such as *cat* and *scratch* or *hair* and *brush* (e.g., McNamara, 1994; Moss, Ostrin, Tyler & Marslen-Wilson, 1995).

Language-based views of semantics tend to allow for more precise definition and quantification. Further, object-based models also tend to be subsumed in language-based models, as Durda and colleagues (2009) showed mapping of featural knowledge in co-occurrence vectors. One way that language-based models quantify the structure of semantic space is by using lexical co-occurrence information or, more specifically, computational co-occurrence models (Buchanan et al., 2001). These co-occurrence models use computational analysis of large bodies of printed text to derive representations of the meaning of a given word based on the frequency with which words occur close to one another. As such, words are characterized as vectors in semantic space, and the relative distance between two words represents how similar they are in meaning. Words that frequently occur together are considered to be related in meaning and are deemed semantic neighbors (Durda & Buchanan, 2008). Thus, in addition to generating a target word's semantic neighbors, co-occurrence models also determine the distance between a word and its neighbors. This produces information about a word's semantic neighborhood, which refers to a hypothetical space within semantic memory that corresponds to the target word and a set of words that are close to it (Lund & Burgess, 1996). For example, the word *whale* co-occurs with semantically related neighbors like *swimming* and *sea*, but *sea* may be more closely related, and so it is situated closer in semantic space relative to *swimming*.

One such co-occurrence variable that is central to the current study is semantic neighborhood density (SND), which is a measure of the average distance of a target word's semantic neighbors as defined by a global co-occurrence model (e.g., WINDSORS; Durda & Buchanan, 2008). SND captures the variability in the overall distribution of related words within a word's semantic neighborhood. Thus, words can be characterized as having dense semantic neighborhoods with more closely distributed neighbors (i.e., high SND), whereas others are characterized as having sparsely distributed neighbors (i.e., low SND). Danguecan and Buchanan (2016) provided a visual depiction of semantic neighborhood densities, as seen in Figure 1. WINDSORS is a global co-occurrence model used in the current study that controls for frequency effects, a common confound found in other models (Durda & Buchanan, 2008).

Figure 1.

Visual representation of dense and sparse semantic neighborhood densities.



High SND **Low SND**

Note. A simplified illustration of a high SND target word with more closely distributed neighbors (i.e., dense semantic neighborhood) and a low SND target word with fewer neighbors more sparsely located (i.e., sparse semantic neighborhood; Danguecan & Buchanan, 2016).

Influence of Semantic Richness on Memory. Most research on the effects of semantics on memory has focused on explicit encoding and retrieval manipulations (e.g., using task instructions that require more or less semantic elaboration; Hargreaves et al., 2012; Schacter & Tulving, 1994). However, rather than attempting to activate semantic knowledge through encoding or retrieval strategies, the current project focuses on the connection between the underlying semantic information of words (i.e., semantic richness) and memory. Semantic richness variables demonstrate the amount of semantic information associated with a particular word (Pexman et al., 2008). Thus, more semantically rich words have greater meaning-related information associated with them and will activate more of that information than words that are less semantically rich. Semantic richness has consistently been shown to affect language processing across tasks (e.g., lexical decision, semantic categorization), such that words with more semantic-related information are recognized faster and more accurately than words with less semantic-related information (Danguecan & Buchanan, 2016; Pexman et al., 2008; Rabovsky et al., 2012; Yap et al., 2011). How this underlying semantic information influences memory for words is less clear. Understanding the influence of word-level semantic characteristics on memory is particularly important for the field of neuropsychology, given that there is such semantic emphasis on tasks like list learning to better understand a patient's memory deficits.

Relatively few studies have examined the impact of word-level semantic richness on memory (Hargreaves et al., 2012; Lau et al., 2018; Nelson et al., 1998; Wong Gonzalez, 2018). There are various ways researchers have operationalized semantic richness to manipulate the semantic information associated with their study items, including the number of semantic associates (Nelson & Schreiber, 1992), number of features (Hargreaves et al., 2012; Lau et al., 2018), and semantic neighborhood density (SND; Wong Gonzalez, 2018).

In several experiments, Nelson and colleagues explored the effects of semantic richness on memory, using the number of semantic associates. The number of semantic associates was found to influence cued-recall memory, such that words with smaller sets of associates were better remembered compared to words with larger sets (Nelson et al., 1998; Nelson et al., 2013). However, this effect was not found for free-recall and was reversed for recognition memory, whereby words with larger sets were better recognized than words with smaller sets (Nelson et al., 1998; Nelson et al., 2013). As such, in a series of experiments, Nelson and colleagues demonstrated that the effect of semantic richness on memory appears to be task dependent.

Relatedly, Hargreaves et al. (2012) and Lau et al. (2018) have studied the effect of semantic richness on memory using number of features. First, Hargreaves et al. (2012) found an effect of semantic richness on free recall, such that words with more features were better recalled than words with fewer features. Similarly, in Lau and colleagues (2018) megastudy, number of features was found to influence recall memory, but not recognition memory. That is, words with more features were recalled better than those with fewer. As such, research by Hargreaves et al. (2012) and Lau et al. (2018) suggests that semantic richness (as operationalized by number of features) facilitates recall but may also be task dependent.

Wong Gonzalez (2018) further explored the effect of word-level semantic richness on memory using a global co-occurrence approach. Semantic richness was operationalized using semantic neighborhood density (SND), discussed above. Wong Gonzalez (2018) found a facilitatory effect for memory, such that high SND nouns were recognized better and showed a greater priming effect than low SND nouns across both explicit (recognition) and implicit (lexical decision) memory tasks.

Another piece of semantic information that has been implicated in memory for words is imageability (i.e., how easily a word can produce an image in the mind; Richardson, 1975). Highly imageable words have consistently been found to be remembered better than words with low imageability (Cortese et al., 2010; Klaver et al., 2005; Lau et al., 2018; Paivio et al., 1994). The widely accepted explanation for this robust finding is a dual-coding theory (Paivio, 1991). Words are processed and remembered using information obtained from the senses both verbally and nonverbally. Thus, imageability is an important and influential factor in the nonverbal code, allowing the word to be rehearsed verbally and represented visually in the mind, ultimately leading to stronger associations and memory traces (Paivio, 1991).

Lastly, the level of emotion and arousal associated with particular words adds critical semantic information that can influence the memorability of such words. Research has confirmed that participants remember emotional stimuli (e.g., pictures, sentences, narrated stories) better than neutral stimuli, known as an emotional memory enhancement effect (Buchanan & Adolphs, 2002; Hamann, 2001). Two aspects of emotionality are important: emotional valence (i.e., the positivity or negativity associated with a word) and emotional arousal (i.e., the relative emotional intensity attributed to a word; Lang et al., 1993). Specific to single words, studies have found better recognition (Kensinger & Corkin, 2003; Ochsner, 2000) and recall (Doerksen & Shimamura, 2001) for emotional words and their source compared to neutral words. Phelps and colleagues (1997) demonstrated this emotion enhancement effect for words even for patients with unilateral temporal lobectomy who previously showed impaired fear acquisition. In contrast, event-related potential (ERP) data has found no effect on emotionality on yes-no recognition tasks (Leiphart et al., 1993; Windmann & Kutas, 2001). The researchers who have found emotion enhancement effects use the dual-process theory (described above; recollection and familiarity) to interpret their results. A common memory task used in attempt to measure differences between recollection versus familiarity is a *remember-know* procedure (Tulving, 1985). A *remember* response indicates that the participant specifically remembers encoding the item, whereas a *know* response denotes that the participant believes the item to have been studied based on a “sense.” Using this procedure, Dougal and Rotello (2007) measured memory differences based on emotionality, including differences in response bias toward emotional words. Results initially showed that emotionally arousing words had lower recognition accuracy compared to neutral words. However, after controlling for semantic similarity between emotional and neutral words, there was only a difference in response bias (i.e., participants consistently respond more liberally to negative words), probability of *remember* responses as opposed to *know* responses (i.e., participants reported more *remember* responses for negative words), but the difference in accuracy disappeared. Interestingly, Windmann and Kutas (2001) also controlled for semantic relatedness, suggesting a potential explanation for the conflicting findings.

Thus, this growing body of literature consistently suggests that many variables reflecting semantic information facilitate memory, at least for free-recall tasks (Hargreaves et al., 2012; Lau et al., 2018; Nelson et al., 1998; Nelson et al., 2013). Some authors (Hargreaves et al., 2012; Wong Gonzalez, 2018) have used a levels-of-processing framework to explain this effect. Specifically, this framework proposes that semantic elaboration during encoding produces more enriched processing and therefore aids memory at retrieval (Craik & Lockhart, 1972; 1990; Craik & Tulving, 1975). Hargreaves and colleagues (2012) and Wong Gonzalez (2018) suggest that this is likely the case for word-level semantic richness as well. It is hypothesized that the facilitatory effect of semantic richness is due to a greater level of activation in the semantic neighborhood of target words (i.e., more semantic information is activated) leading to a stronger memory trace (Hargreaves et al., 2012; Wong Gonzalez, 2018) analogous to the rich processing produced by semantic elaboration.

Summary and Conclusions

Although the impact of lexical and semantic variables on memory are less consistent and well known compared to those in the language processing literature, the above review demonstrates that using memory for words as a diagnostic component for underlying neuropsychological conditions without taking these variables into account may be problematic. Specifically, performance may be influenced by the underlying structure of the word lists rather than inherent deficits/abilities. A summary of the literature is presented in Table 1.

Table 1.

Summary of effects of lexical and semantic variables on memory.

Lexical Variable	Recognition Memory	Recall Memory
Word Frequency	✓	✓= pure lists ? = mixed lists
Orthographic/Phonological Neighborhood	✓	?
Word Length	-	✓= pure lists ? = mixed lists
Age of Acquisition	✓	? = pure lists ? = mixed lists
Semantic Variable	**Recognition Memory**	**Recall Memory**
Number of Associates	✓	✓
Number of Features	X	✓
Semantic Neighborhood Density	✓	-
Imageability	?	✓
Emotional Valance & Arousal	?	✓

Note. ✓ = consistent findings of an advantage or disadvantage; X = consistent findings of no effect; ? = inconsistent findings; - = no findings documented.

Neuropsychological Tests of List-Learning

Clinically, neuropsychologists almost always include measures of verbal learning and memory in their test batteries, given their importance and relevance in neurological, academic, and adaptive functioning. Verbal memory is typically assessed in one of two ways (and sometimes both): 1) list learning, where examinees listen to and recall a list of words over several trials and after a delay; and 2) contextual memory, where examinees listen to and recall verbal information in paragraph (or story) form (Strauss et al., 2006). For both list learning and contextual memory tests, several variations can be made to include semantic cueing and/or interference lists for list learning and single vs. multiple paragraph or multiple trials for contextual memory, as well as other additions to tap into different aspects of verbal learning (Strauss et al., 2006). The current project will only focus on verbal list learning tasks, given their particular psycholinguistic relevance. The subsequent section highlights the concern that neuropsychological batteries and tasks often do not consider the importance of lexical characteristics of target words. The following tests were chosen to represent a sample of neuropsychological memory tasks based on both their popularity in clinical use, as well as the availability and accessibility of test manuals and test items.

Child and Adolescent Memory Profile (ChAMP)

The Child and Adolescent Memory Profile (ChAMP; Sherman & Brooks, 2015) is a battery used to assess both verbal and visual learning in children and adolescents from age five through 21 (Sherman & Brooks, 2015). There are three verbal list-learning subtests on the ChAMP: 1) *Lists*, which consists of three learning trials, each with the same 16 items; 2) *Lists Delayed*, administered approximately 20 minutes following the completion of *Lists*, which requires examinees to recall the original 16 items presented during *Lists*; and 3) *Lists Recognition*, administered immediately following the administration of *Lists Delayed*, which requires the examinee to choose which items were presented at study out of three choices (i.e., the target word, a semantically related word, and an unrelated word). There is no rationale or explanation provided in the test manual (see Sherman & Brooks, 2015) for the choice of test items.

California Verbal Learning Test – Children's Version (CVLT-C)

The California Verbal Learning Test – Children's Version (CVLT-C; Delis et al., 1994) is a list learning task that measures immediate and delayed verbal recall (free- and cued-recall) and recognition of two lists (*Monday list* and *Tuesday list*) in children aged five to 16. Target words on the CVLT-C fall into one of three semantic categories (i.e., Things to wear, Things to play with, Fruits). Examinees are presented with five learning trials of a *Monday Shopping List* (15 items). Examinees are then immediately given a single learning trial of a *Tuesday Shopping List* (15 new items). The interference trial (*Tuesday List*) is then immediately followed by *Short*

Delay Free- and *Cued-Recall* trials, where the examinee is asked to first recall the *Monday List*, then asked to recall the *Monday List* with semantic cues. After a 20-minute delay, examinees are given *Long Delay Free*- and *Cued-Recall* trials. Examinees are then immediately administered a *Yes/No Recognition* trial, consisting of 45 items (i.e., items from both the *Monday List* and the *Tuesday List,* as well as new items; Delis et al., 1994).

Delis and colleagues (1994) used two criteria to develop the lists of target words (i.e., *Monday* and *Tuesday* lists). First, mean word frequency was controlled between the two lists. Second, prototypicality ratings were obtained by asking children to produce as many items from each semantic category as possible within a specific amount of time. These ratings were then used to eliminate the three most prototypical responses from each category to limit highly prototypical intrusions and to minimize the effect of potential confabulation (Delis et al., 1994).

California Verbal Learning Test – Third Edition (CVLT-3)

The California Verbal Learning Test – Third Edition (CVLT-3; Delis et al., 2017) is a list-learning task that measures immediate and delayed verbal recall (free- and cued-recall) and recognition of two lists (*List A* and *List B*) in individuals aged 16 to 90 years, 11 months. Target words on the CVLT-3 fall into one of four semantic categories (i.e., Furniture, Vegetables, Transportation, Animals). The procedure is generally the same as that described for the CVLT-C, with the addition of a forced-choice recognition trial. Examinees are presented with five learning trials of *List A* (16 items). Examinees are then immediately given a single learning trial of *List B* (16 new items). The interference trial (*List B*) is then immediately followed by *Short Delay Free*- and *Cued-Recall* trials, where the examinee is asked to first recall *List A*, then asked to recall *List A* with semantic cues. After a 20-minute delay, examinees are given *Long Delay Free*- and *Cued-Recall* trials. Examinees are then immediately administered a *Yes/No Recognition* trial, consisting of 48 items (i.e., items from both *List A* and *List B,* as well as new items). Finally, following a ten-minute delay, participants are given a *Forced Choice Recognition* trial (Delis et al., 2017).

The word lists used on the CVLT-3 are the same as those on the CVLT-II; the CVLT-3 test manual discusses a single criterion used to develop the target words in List A and List B. The items were required to not be highly prototypical of their respective semantic categories (Delis et al., 2017). Thus, the four most prototypical items from each semantic category were excluded from the target lists, as rated by 154 participants as part of an unpublished study in 1995 (Delis et al., 2017). No other criteria were discussed for development.

Hopkin's Verbal Learning Test – Revised (HVLT-R)

The Hopkin's Verbal Learning Test – Revised (HVLT-R; Brandt & Benedict, 2001) is another list-learning task that measures immediate and delayed verbal learning. Examinees are presented with three learning trials, each with the same 12 items, followed by a 20- to 25-minute delay, at which point delayed recall and recognition are assessed.

For both the HVLT (Brandt, 1991) and the HVLT-R (Brandt & Benedict, 2001), 18 semantic categories were selected from the Battig and Montague (1969) category exemplar collection to create six forms of the test, each consisting of three semantic categories. For each of these six lists, four high frequency responses to each category were chosen to be included. However, the two most common responses to the semantic categories were excluded from the target list and instead were included in the distractor list for the recognition trial (Brandt & Benedict, 2001). Further, the authors report having controlled for average frequency of response to the category (i.e., how commonly they were produced as a member of the category), as well as for word frequency (Brandt & Benedict, 2001).

Repeatable Battery for the Assessment of Neuropsychological Status (RBANS)

The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS; Randolph et al., 1998) is a battery of subtests used to assess various neuropsychological domains (e.g., memory, language, visuospatial, attention). The battery was originally used to identify and assess decline related to dementia in older adults but was eventually adapted for use with younger adults (Strauss et al., 2006). The *List Learning* subtest consists of four learning trials of ten items. The *List Recall* subtest is then administered after a 20-minute delay, requiring examinees to freely recall the original ten items. The *List Recognition* subtest is then administered, where examinees hear an additional 20 items and are asked to identify the ten items that were from the list. The test manual indicates that items were selected based on their semantic and phonemic distinctiveness; however, there is no information regarding how this information was delineated.

Wide Range Assessment of Memory and Learning – Second Edition (WRAML-2)

The Wide Range Assessment of Memory and Learning – Second Edition (WRAML-2; Sheslow & Adams, 2003) is a battery of subtests that are used to assess different aspects of memory and learning in examinees aged five to 90, including working memory and attention, verbal learning and memory, and visual memory. The *Verbal Learning* subtest consists of four learning trials of 13 items for individuals aged eight and younger and 16 items for individuals aged nine and older. The *Verbal Learning Delay Recall* subtest is then administered after a 20-minute delay, requiring examinees to freely recall the original 13 or 16 items. The *Verbal Learning Recognition* subtest is then administered. In this phase, examinees determine which of 34 items (for ages eight and younger) or 40 items (for ages nine and older) were on the original studied list. The test manual indicates that the target words included in the list were adapted from Rey (1985); there is no available information on how the original words were chosen.

A Developmental Neuropsychological Assessment – Second Edition (NEPSY-II)

A Developmental NEuroPSYchological Assessment – Second Edition (NEPSY-II; Korkman et al., 2007) is a battery of tests used to measure various neuropsychological domains (i.e., language, verbal and visual memory, working memory, executive functioning and attention, sensorimotor functioning, visuospatial skills, phonological processing, and social processing) in children aged three years to 16 years, 11 months depending on the subtest. There are two relevant list-learning subtests on the NEPSY-II: 1) *List Memory,* where examinees are presented with four learning trials of 15 items, followed by a single trial of an interference list of 15 items; and 2) *List Memory Delayed*, administered after a 20-minute delay, where examinees are asked to recall the original 15 items presented. Like most of the tests discussed above, there is no available information regarding the development of the test lists.

Summary and Conclusions

As reviewed above, it has not been common practice for neuropsychological tests to consider lexical and or semantic variables in the development of target words and lists. See Table 2 for a summary of variables considered in the development of the neuropsychological tests discussed.

Table 2.

Summary of lexical variables considered or controlled for in relevant neuropsychological tests.

	Lexical & Semantic Variables						
Test	Word Freq.	Word Length	ON & PN	AoA	Imageability	Emotion Variables	Semantics
ChAMP	X	X	X	X	X	X	X
CVLT-C	✓	X	X	X	X	X	(✓)
CVLT-3	X	X	X	X	X	X	(✓)
HVLT-R	✓	X	X	X	X	X	(✓)
RBANS	X	X	(✓)	X	X	X	(✓)
WRAML-2	X	X	X	X	X	X	X
NEPSY-II	X	X	X	X	X	X	X

Note. ✓ = indicates that test developers considered the variable in creation of word lists or test items; (✓) = aspects of the variable were considered, but not word-level characteristics as operationalized in the current study. X = indicates that they did not report any considerations. For CVLT-C, CLVT-3, HVLT-R, and RBANS the semantic richness variable (and phonological neighborhood for RBANS) was generously evaluated, as prototypical responses to relevant semantic categories were considered, but a quantitative measure of word-level semantic richness was not controlled for.

Lexical and semantic properties of test items may not be obviously important for interpreting performance on these memory tests due to the use of standardized normative data; however, given that the diagnosis of several disorders rely on a pattern of language and memory deficits, taken with the fact that memory assessment typically uses words (language) as stimuli, it is crucial that neuropsychological measures are not confounded by psycholinguistic properties. Further, if, using the Boston Process Approach discussed previously, test developers could develop specific test lists that are carefully and thoughtfully manipulated and introduce process scores that are particularly sensitive to certain psycholinguistic variables, it may allow for a more fined-grain analysis that is sensitive to more delicate changes in memory leading to earlier identification or more targeted intervention plans. To better understand this concept, take the

Beck Depression Inventory – Second Edition (BDI-II; Beck et al., 1996), for example. The BDI-II is a self-report inventory that includes questions related to symptoms of depression (e.g., sadness, crying, anhedonia, suicidal ideation, sleep dysfunction, sexual dysfunction, etc.) and is used to assess level of depression. The inventory is scored by tallying up the level of severity indicated across all items; therefore, all items are weighted the same in the scoring system, including the question related to suicidal ideation. Thus, a patient could technically indicate minimal symptoms across items accompanied by severe suicidal ideation and still obtain a final score within the minimal or no depression range. It would be extremely important that a clinician look more carefully at the individual items rather than using that final score to indicate level of functioning and to understand their patient's clinical presentation. Although that example is related to a different domain of functioning, the idea proposed here is similar. Currently, each item within each neuropsychological test list is weighted the same in the scoring system; it could be fruitful to look more specifically at which words patients are recalling and which ones they are not.

More specifically, several specific deficits have been implicated in different patient populations that may necessarily lead to differential processing of certain words within lists. For example, patients with Mild Cognitive Impairment (MCI), Alzheimer's disease, semantic dementia, primary progressive aphasia, and Parkinson's disease exhibit diminished semantic processing and memory (Angwin et al., 2006; Croisile et al., 1996; Huff et al., 1986; McNamara et al., 1992). Despite semantic processing being remarkably well-preserved in healthy aging, some research has also shown that it begins to decline very late in life (Robert & Rico Duarte, 2016). Further, bilingual speakers are more sensitive to semantic information (Johns et al., 2016). Using a semantic priming task, Vandenberghe et al. (2005) found a semantic interference effect for individuals with primary progressive aphasia, such that they were slower to name a picture after a related prime compared to an unrelated prime. This effect was not present for individuals with Alzheimer's disease, nor for people with MCI, which suggests a distinguishing marker between these disorders and primary progressive aphasia (Vandenberghe et al., 2005). Although all three types of disorders have weakened semantic processing, only one condition leads to a semantic interference effect. Similarly, Price and Grossman (2005) found subtle differences in language processing between patients with Alzheimer's disease and those with frontotemporal dementia. They found that patients with frontotemporal dementia have a heightened sensitivity to violations in verb agreement compared to those with Alzheimer's disease (Price & Grossman, 2005). Specific to the current study, these populations may show differential learning and memory for certain test items based on underlying semantic richness information. Thus, moving beyond typical scoring systems and investigating more precise differences in lexical-semantic processing could be beneficial to furthering our understanding of brain-behavior relationships.

Moreover, patients with conditions that affect access and awareness to phonological information, such as those with dyslexia or variants of aphasia (Mendez et al., 2003; Buchanan et al., 1999), may show learning and memory differences based on orthographic and phonological properties of the items. Further, word frequency effects have been shown to be different based on several patient populations. Both individuals with Alzheimer's disease and those who have had a stroke with left-hemisphere damage have been shown to have a decreased ability to recognize and recall low frequency words as compared to high frequency words (Bayley et al., 2003; Burke et al., 1991; Breining & Rapp, 2019; Heidlmayr et al., 2015). In contrast, research has found that individuals with schizophrenia show impaired recall for high frequency words, while recall for low frequency words is generally intact (Brébion et al., 2005).

Neuropsychologists are potentially capturing confounding information and perhaps may be missing informative and clinically relevant details by ignoring the effects of word-level lexical and semantic characteristics. Recall the Boston Process Approach and the three-patient example presented earlier. A similar interpretation approach may be useful in list learning tasks, where word-level characteristics and potentially process scores could provide important diagnostic clarification or guide treatment recommendations for unique patient populations. Thus, the purpose of the current project is to explore these effects using well-established and commonly used neuropsychological tests of memory and list learning.

STUDY 1

Objective

The objective of Study 1 was to gather, document, and examine all relevant lexical and semantic characteristics (i.e., word frequency, word length, orthographic and phonological neighborhood size, age of acquisition, emotional valence and arousal, imageability, semantic diversity, SND, and sensorimotor information) for each item on each neuropsychological test mentioned above (*NOTE*: test items are not provided herein to maintain test security). This information was necessary for subsequent studies.

Method

The stimulus set consisted of the items from seven neuropsychological memory (i.e., list learning) tasks, including the CVLT-C, CVLT-3, ChAMP, WRAML-2, RBANS, HVLT-R, and NEPSY-II. Word frequency was obtained from Lund and Burgess (1996). Word length and orthographic and phonological neighborhood size variables were obtained from Wordmine2 (Durda & Buchanan, 2006). Age of acquisition ratings were obtained from Kuperman et al. (2012). Emotional valence and emotional arousal values were obtained from Warriner et al. (2013). Imageability ratings were obtained from Scott et al. (2019). Semantic diversity values were obtained from Hoffman et al. (2012). Semantic neighborhood density values were obtained from the Word Pair App (Lutfallah et al., 2018). Sensorimotor information values were obtained from Lynott et al. (2020). See Table 3 for a summary of predictor variables and operational definitions. Of note, concreteness was not examined, as all items across all lists were concrete nouns.

Table 3.

Summary of predictor variables.

Variable	***Operational Definition***
Log-frequency (word frequency; WF)	Ratings regarding the frequency with which a word occurs in a language. Values were obtained from the Hyperspace Analogue Language (HAL) norms (Lund and Burgess, 1996).
Orthographic neighborhood (ON) size	The total number of words that are exactly one letter different from the target word. Values obtained from Durda and Buchanan (2006).
Phonological neighborhood (PN) size	The total number of words that are exactly one phoneme different from the target word. Values obtained from Durda and Buchanan (2006).
Word length (WL)	The number of letters in a word.
Age of acquisition (AoA)	The average chronological age at which a target word is learned. Ratings were obtained from Kuperman et al. (2012).
Emotional valence (EV)	The positivity or negativity associated with a target word. Ratings were obtained from Warriner et al. (2013).
Emotional arousal (EA)	The relative emotional intensity associated with a target word. Ratings obtained from Warriner et al. (2013).
Imageability (Img)	The ease with which a word can produce an image in the mind. Ratings obtained from the Glasgow Norms (Scott et al., 2019).
Semantic diversity (SD)	The number and contexts of meaning that can be attributed to one word of the same spelling. Ratings obtained from Hoffman et al. (2012).
Semantic neighborhood density (SND)	The average distance of a target word's semantic neighbors as defined by a global co-

	occurrence model. Values were obtained from the Word Pair App site (Lutfallah et al., 2018).
Sensorimotor information (SI)	An overall measure of the connection between a word and perceptual (e.g., gustatory, auditory, olfactory, etc.) and motoric or action parts of the body (e.g., hand, foot, torso, etc.). Values were obtained from the Lancaster Sensorimotor Strength Norms (Lynott et al., 2020).

Results

To better understand the underlying structure of the target words within the neuropsychological test lists, each value mentioned above was standardized into a Z-score as compared to variable's entire corpus using descriptive statistics obtained from each respective source provided above. For example, the word "Bananas" has a word frequency Z-score of 1.20; meaning that compared to all words rated in the HAL norms (Lund & Burgess, 1996), "Bananas" has a word frequency value 1.20 standard deviations above the mean. Tables 4 through 10 present Z-scores for word frequency, age of acquisition, emotional valence, emotional arousal, imageability, semantic diversity, semantic neighborhood density, and sensorimotor information for each word in each neuropsychological test. Word length, orthographic neighborhood, and phonological neighborhood variables were not standardized as the raw values can logically be interpreted. These tables allow for visualization of within list differences between test items.

Table 4.

CVLT-C item breakdown.

Item	List	WL	WF	ON	PN	AoA	EV	EA	Img	SemD	SND	SI
Item 1	A	7	1.20	0	0	-2.00	1.28	-1.11	1.48	0.24	2.07	2.47
Item 2	A	7	0.93	4	5	-1.52	0.87	-1.57	1.39	-0.95	2.31	1.77
Item 3	A	6	1.41	3	5	-1.55	1.14	0.24	0.64	0.86	0.76	1.91
Item 4	A	6	1.64	2	2	-1.95	0.61	-0.95	1.51	-0.38	2.03	2.26
Item 5	A	6	1.01	10	9	-1.95	1.27	-0.79	1.43	-1.14	1.85	2.10
Item 6	A	6	1.76	3	10	-1.66	-0.47	0.05	0.25	0.86	-0.32	0.60
Item 7	A	10	0.51	0	0	-1.85	1.31	0.49	-3.55	-	2.27	2.93
Item 8	A	6	1.43	7	19	-1.95	0.68	0.03	-0.01	-0.27	0.93	1.17
Item 9	A	7	0.43	0	1	-2.20	0.53	-1.45	1.32	-1.38	0.88	1.09
Item 10	A	7	1.04	7	18	-1.86	1.37	0.55	1.49	-0.53	2.15	2.38
Item 11	A	8	1.37	0	5	-1.80	1.38	-0.34	1.33	0.40	-0.01	0.98
Item 12	A	3	1.69	23	36	-2.16	0.48	-1.48	1.48	0.27	0.40	1.79
Item 13	A	12	1.00	0	0	-1.86	1.70	-0.17	1.55	-0.60	2.37	2.88
Item 14	A	4	1.66	11	17	-1.72	-0.50	-0.84	1.16	0.64	-0.05	1.27
Item 15	A	7	1.32	3	4	-0.90	0.57	-0.71	-	-0.32	0.18	1.28
Item 1	B	9	0.59	0	1	-1.45	1.90	0.33	-	0.01	2.42	2.82
Item 2	B	8	1.35	2	7	-1.40	1.54	0.79	1.36	-1.28	2.34	1.74
Item 3	B	5	2.23	3	6	-1.80	0.32	-1.34	1.49	0.74	-0.49	1.44
Item 4	B	5	0.77	5	7	-1.42	0.84	-1.61	1.14	-0.66	2.47	0.99
Item 5	B	7	1.21	3	5	-2.14	1.76	0.55	-	-0.54	2.21	2.61
Item 6	B	4	1.49	12	13	-1.93	0.52	-1.67	1.47	-0.14	0.92	1.33
Item 7	B	8	-	-	-	-2.15	-	-	-	-	1.02	-
Item 8	B	5	0.83	15	42	-1.93	1.27	-0.50	1.43	-0.98	2.32	1.83
Item 9	B	3	2.03	16	40	-2.30	1.63	-1.34	1.39	-	0.05	2.78
Item 10	B	5	1.39	7	7	-1.93	1.72	0.92	1.38	-0.50	1.79	2.54
Item 11	B	3	0.96	16	24	-1.72	-0.06	-1.08	-	-0.74	1.00	1.37
Item 12	B	6	1.32	1	0	-1.68	1.01	0.35	1.39	-0.93	2.14	2.20
Item 13	B	8	1.03	0	2	-1.66	-	-	-	-1.61	1.83	1.53
Item 14	B	9	0.68	0	0	-1.17	1.43	0.23	1.55	-0.43	1.91	3.07

Item 15	B	4	1.74	2	2	-1.40	0.38	-1.96	1.27	-0.28	0.12	1.54

Table 5.

CVLT-3 item breakdown.

Item	List	WF	WL	ON	PN	AoA	EV	EA	Img	SemD	SND	SI
Item 1	A	1.64	5	0	7	-2.00	0.06	-0.50	1.32	0.08	0.66	1.29
Item 2	A	0.80	7	0	2	-1.61	0.57	-0.86	1.29	-1.68	2.31	2.69
Item 3	A	0.50	7	0	0	-1.59	1.13	-1.45	1.58	-0.34	2.02	0.57
Item 4	A	0.48	8	0	0	-1.38	0.80	-0.92	1.38	-0.85	0.69	0.84
Item 5	A	1.17	5	1	0	-1.24	0.23	0.83	1.37	-2.21	2.05	2.77
Item 6	A	1.29	10	0	0	-1.57	0.57	1.78	1.48	0.48	1.33	2.64
Item 7	A	1.48	7	0	0	-1.23	0.02	-0.51	1.24	-0.25	0.59	0.91
Item 8	A	0.81	5	0	0	-1.66	1.09	-0.34	1.50	-0.70	1.38	0.48
Item 9	A	1.07	6	0	1	-0.69	0.28	0.23	-	-0.24	0.69	0.26
Item 10	A	1.49	4	12	13	-1.93	0.52	-1.67	1.47	-0.14	0.92	1.33
Item 11	A	0.73	6	0	1	-1.33	0.50	-1.56	1.38	-2.69	2.53	2.07
Item 12	A	1.52	3	24	18	-1.95	0.27	-1.40	1.55	-0.30	0.74	0.82
Item 13	A	1.74	4	2	2	-1.40	0.38	-1.96	1.27	-0.28	0.11	1.54
Item 14	A	1.79	4	10	34	-1.98	1.00	-0.17	1.51	-0.57	1.24	0.21
Item 15	A	1.07	8	0	0	-1.78	0.50	0.31	1.40	-0.49	1.72	0.74
Item 16	A	0.88	7	0	3	-1.33	-0.37	-1.45	-	-0.57	2.19	2.15
Item 1	B	1.17	6	0	0	-0.76	1.16	-0.89	1.54	-1.84	1.67	2.21
Item 2	B	0.71	8	0	1	-1.35	1.09	-1.16	-	-1.22	2.10	2.74
Item 3	B	1.51	8	0	3	-1.66	0.86	0.03	1.54	0.02	1.06	1.12
Item 4	B	1.41	6	3	0	-1.59	0.02	-0.73	1.65	0.33	0.69	0.15
Item 5	B	0.40	6	0	1	-0.92	-0.35	-0.99	-	-0.31	2.16	0.78
Item 6	B	1.95	6	0	2	-1.48	1.58	0.22	1.44	-2.44	0.76	2.13
Item 7	B	1.37	8	1	2	-1.00	-0.21	-0.98	1.01	0.27	0.45	0.48

Item 8	B	1.53	5	7	23	-1.84	0.19	-1.40	1.54	-0.43	1.42	1.22
Item 9	B	1.01	8	0	0	-0.17	-0.26	-1.28	1.38	-2.46	1.97	2.24
Item 10	B	1.45	6	0	0	-1.50	0.33	-0.12	-	0.10	-0.27	0.28
Item 11	B	1.39	4	14	54	-1.72	0.68	-0.86	1.11	-0.32	1.21	2.65
Item 12	B	1.44	6	1	5	-1.95	1.67	-0.25	1.49	-0.24	0.98	0.31
Item 13	B	0.71	5	1	3	-0.58	1.20	-1.85	-	-1.23	0.68	0.59
Item 14	B	0.63	9	0	0	-0.21	0.97	0.52	1.29	-1.41	1.52	1.91
Item 15	B	1.42	5	3	5	-1.93	0.72	1.50	1.44	0.28	0.45	1.02
Item 16	B	0.15	8	1	1	-1.51	-0.29	-0.55	-	-	2.48	1.59

Table 6.

CHAMP item breakdown

Item	List	WF	WL	ON	PN	AoA	EV	EA	Img	SemD	SND	SI
Item 1	A	2.35	5	3	26	-1.22	0.51	-0.41	1.36	0.92	-0.60	1.04
Item 2	A	2.16	3	11	15	-1.52	-0.51	0.16	1.30	0.80	1.14	1.75
Item 3	A	1.86	6	2	7	-1.40	0.28	-0.81	1.41	0.48	0.13	0.55
Item 4	A	1.90	4	18	36	-1.72	1.61	-1.75	1.43	0.02	0.15	1.22
Item 5	A	1.22	7	0	0	-1.00	-0.38	2.78	1.46	-1.17	1.28	-0.16
Item 6	A	1.82	6	1	1	-1.16	1.25	0.26	1.37	-0.77	0.42	1.46
Item 7	A	1.53	6	2	9	-1.29	-0.47	-0.13		-0.10	0.24	0.44
Item 8	A	2.11	4	1	36	-1.82	0.15	-1.13		1.43	-0.50	1.39
Item 9	A	2.21	5	13	16	-1.67	0.68	-0.86	0.23	0.63	1.36	0.41
Item 10	A	1.63	6	0	0	-0.39	1.25	-0.52	1.23	-0.81	0.35	0.64
Item 11	A	1.16	4	14	31	-1.76	0.85	-0.71	1.36	-0.51	0.24	0.39
Item 12	A	1.34	8	1	1	-0.86	0.93	-1.23	1.45	0.38	-0.21	1.20
Item 13	A	1.93	5	6	11	-2.70	1.00	-0.17	1.42	0.03	0.80	1.57
Item 14	A	1.53	5	8	8	-1.95	1.09	-0.91	1.47	0.02	0.30	2.05
Item 15	A	1.64	4	11	18	-1.72	0.85	-0.86	1.22	-0.24	0.30	1.54

Item 16	A	1.41	5	2	11	-1.16	-0.02	-1.68	1.36	0.22	-0.01	0.89

Table 7.

WRAML-2 item breakdown.

Item	List	WF	WL	ON	PN	AoA	EV	EA	Img	SemD	SND	SI
Item 1	A	1.64	4	11	18	-1.72	0.85	-0.86	1.22	-0.24	0.30	1.54
Item 2	A	2.67	4	15	21	-1.84	1.34	0.94	0.45	0.12	0.59	1.69
Item 3	A	1.69	3	23	36	-2.16	0.48	-1.48	1.48	0.27	0.40	1.79
Item 4	A	1.98	4	5	10	-2.07	1.97	-1.71	1.41	-0.02	0.54	1.56
Item 5	A	1.68	3	11	11	-2.05	0.61	-0.79	1.53	0.49	0.03	1.30
Item 6	A	1.00	4	6	24	-1.42	0.45	-2.09	1.39	-0.44	-0.12	2.38
Item 7	A	1.78	4	9	7	-1.48	0.80	-0.52		0.90	0.10	0.57
Item 8	A	1.94	4	11	22	-1.73	0.58	-0.79	1.35	0.37	0.31	1.35
Item 9	A	1.96	3	23	28	-1.39	0.57	-0.28		0.26	0.19	1.07
Item 10	A	2.22	4	6	40	-2.25	0.28	-1.13	1.41	0.11	0.11	1.91
Item 11	A	2.10	3	4	12	-1.98	0.77	-1.01	1.21	0.12	-0.03	2.43
Item 12	A	1.36	4	11	37	-1.45	-0.37	-1.29	1.04	0.59	0.44	2.54
Item 13	A	1.79	4	10	34	-1.98	1.00	-0.17	1.51	-0.57	1.24	0.21
Item 14	A	2.59	4	11	17	-1.54	0.79	-0.84	1.23	0.73	0.10	0.86
Item 15	A	1.19	3	7	5	-1.82	-0.92	-1.04	1.45	0.08	2.30	0.50
Item 16	A	1.90	4	18	36	-1.72	1.61	-1.75	1.43	0.02	0.15	1.22

Table 8.

RBANS item breakdown.

Item	List	WF	WL	ON	PN	AoA	EV	EA	Img	SemD	SND	SI
Item 1	A	2.43	6	0	1	-0.86	0.89	-0.73	1.03	1.07	0.08	0.89
Item 2	A	2.27	7	0	3	-1.05	0.07	0.59	0.78	0.39	-0.29	1.79
Item 3	A	1.26	5	0	0	-1.67	0.24	-1.12	1.16	0.15	1.74	1.54
Item 4	A	2.29	5	2	4	-1.88	1.21	-0.76	1.57	-0.10	0.12	2.76
Item 5	A	2.42	5	4	3	-1.97	1.72	-0.76	-0.73	0.64	0.01	1.45
Item 6	A	1.61	6	1	2	-1.24	0.68	-1.60	1.15	-0.21	0.57	1.22
Item 7	A	1.45	6	5	9	-2.00	1.06	-0.02	1.33	0.85	-0.37	0.95
Item 8	A	1.68	7	0	1	-1.08	0.09	0.08	1.18	0.10	0.97	1.14
Item 9	A	1.20	6	2	3	-1.11	-0.10	-1.23		-1.24	0.66	0.98
Item 10	A	1.56	6	2	4	-1.57	0.14	-1.60		0.22	0.02	1.25

Table 9.

NEPSY-II item breakdown.

Item	List	WF	WL	ON	PN	AoA	EV	EA	Img	SemD	SND	SI
Item 1	A	2.21	5	13	16	-1.67	0.68	-0.86	0.23	0.63	1.36	0.41
Item 2	A	1.58	5	4	9	-2.17	2.17	-1.77	1.47	-1.84	2.08	2.88
Item 3	A	2.09	6	6	3	-2.12	0.57	-0.06	1.47	0.41	0.41	2.11
Item 4	A	2.22	6	1	2	-1.68	1.09	-1.04	1.41	0.20	0.04	1.46
Item 5	A	1.53	5	8	8	-1.95	1.09	-0.91	1.47	0.02	0.30	2.05
Item 6	A	2.24	6	6	20	-1.68	0.47	-1.13	1.21	0.24	0.21	1.22
Item 7	A	2.02	4	4	12	-1.91	1.05	-0.98	1.44	-0.48	1.46	2.83
Item 8	A	0.85	5	1	3	-0.56	0.35	-0.50	0.83	-0.83	-0.33	1.18
Item 9	A	1.93	6	3	4	-1.80	0.33	-1.60	1.05	0.39	0.83	1.18

Item 10	A	2.10	3	23	36	-2.04	1.47	0.33	1.47	0.14	0.35	2.35
Item 11	A	1.22	6	0	0	-1.91	0.45	-1.22	1.58	0.02	0.72	1.30
Item 12	A	1.41	5	2	11	-1.16	-0.02	-1.68	1.36	0.22	-0.01	0.89
Item 13	A	1.82	7	1	6	-1.74	1.79	-1.46	0.36	-0.56	0.01	0.93
Item 14	A	2.45	5	8	15	-2.48	1.50	-0.55	1.56	0.75	-0.21	2.95
Item 15	A	1.79	4	10	34	-1.98	1.00	-0.17	1.51	-0.57	1.24	0.21
Item 1	B	1.44	6	1	5	-1.95	1.67	-0.25	1.49	-0.24	0.98	0.31
Item 2	B	2.24	3	15	18	-2.33	1.50	1.37	1.56	-0.87	2.31	2.91
Item 3	B	2.29	5	2	4	-1.88	1.21	-0.76	1.57	-0.10	0.12	2.76
Item 4	B	1.47	6	2	14	-1.10	0.80	-0.37	1.22	-0.66	-0.05	0.94
Item 5	B	1.19	4	2	3	-0.72	-0.22	-0.46	1.34	-0.01	0.39	1.34
Item 6	B	2.22	4	6	40	-2.25	0.28	-1.13	1.41	0.11	0.11	1.91
Item 7	B	1.26	4	7	29	-1.88	0.96	-0.99	1.53	-0.04	0.91	1.47
Item 8	B	0.96	8	4	7	-1.02	0.20	-0.14		-0.35	-0.26	0.43
Item 9	B	1.74	6	1	2	-2.18	1.36	-0.18	1.36	0.28	0.34	1.65
Item 10	B	1.85	5	2	27	-2.12	0.64	-1.50	1.34	0.23	-0.01	2.02
Item 11	B	1.05	5	7	9	-1.82	0.68	-0.86	1.36	-0.76	1.00	0.57
Item 12	B	2.02	6	0	4	-2.08	1.69	0.21	0.45	-1.86	0.49	2.52
Item 13	B	2.23	5	4	6	-1.80	0.32	-1.34	1.49	0.74	-0.49	1.44
Item 14	B	1.41	4	6	32	-2.40	0.55	-2.01	1.49	0.22	0.23	1.89
Item 15	B	1.39	5	7	7	-1.93	1.72	0.92	1.38	-0.50	1.79	2.54

Table 10.

HVLT-R item breakdown.

Item	List	WF	WL	ON	PN	AoA	EV	EA	Img	SemD	SND	SI
Item 1	A	1.54	4	4	7	-1.79	0.60	1.21	1.56	0.70	0.72	1.78
Item 2	A	1.47	7	0	0	-0.49	1.74	1.26	1.02	-0.13	0.36	0.38
Item 3	A	1.93	5	8	24	-1.88	0.76	-0.05	1.51	-0.80	1.76	2.31
Item 4	A	1.24	4	16	21	-1.54	0.90	-0.92	1.34	-0.32	0.04	1.42
Item 5	A	1.20	8	0	2	-0.17	1.44	0.75	1.15	-0.46	0.70	0.64
Item 6	A	1.88	5	2	1	-1.24	1.19	0.39	1.37	-0.15	0.57	0.77
Item 7	A	1.53	4	17	22	-1.00	0.06	0.21	1.18	-0.06	0.04	0.31
Item 8	A	0.69	4	2	3	-0.15	0.72	-1.19		-0.51	0.34	-0.66
Item 9	A	1.42	5	3	5	-1.93	0.72	1.50	1.44	0.28	0.45	1.02
Item 10	A	1.66	5	1	31	-1.16	0.76	-0.79	1.33	0.33	1.27	0.83
Item 11	A	1.52	3	24	18	-1.95	0.27	-1.40	1.55	-0.30	0.74	0.82
Item 12	A	1.05	3	17	31	-0.54	0.02	-1.45		-0.20	-0.02	-0.09

STUDY 2

Objective

The objective of Study 2 was to evaluate the effect of lexical and semantic variables on memory performance for words from common neuropsychological memory tests. Based on the current literature, it was hypothesized that both lexical and semantic variables would account for unique variance in overall free recall of test items. Specifically, it was hypothesized that semantic richness variables (i.e., SND, imageability, semantic diversity, and sensorimotor information), word frequency, and orthographic and phonological neighborhood size would account for unique variance in recall performance, given their more robust and consistent effects documented in the literature.

Participants and Inclusion Criteria

All participants in the study were undergraduate students at the University of Windsor who volunteered to participate in return for bonus credits toward their eligible psychology courses. Inclusion criteria required participants to report English as their first language. One hundred twenty-one participants were recruited for the study. One participant's data was removed due to failure to meet inclusion criteria (i.e., English was their second language). Thus, the final sample included 120 participants (M_{age} = 23.5, 89% female, 22% bilingual).

Method

Stimuli

The stimulus set described in Study 1 was used to create study lists. Study lists were created by combining three test lists (i.e., the number of target words per participant varied from 42 to 47, depending on which test lists were included). Study lists were created to ensure that the number of items at study aligned with experimental memory tasks. Interference lists (e.g., List B in the CVLT-3) were used as independent lists, making ten test lists to be combined into study lists. The combination of test lists and the order of the target words included in the study lists were randomized per participant.

Procedure

Before beginning data collection, Study 2 was approved by the University of Windsor's Research Ethics Board (REB). Participants volunteered to participate through the University of Windsor Participant Pool and provided oral and signed consent. Participants were randomly assigned to a study list. There were three phases to Study 2: Learning, Short-Delay Recall, and Long-Delay Recall. The entire procedure for Study 2 was designed to simulate clinical procedures of list learning tasks as much as possible to maximize the generalizability of findings, though some modifications were made to ensure feasibility for research (e.g., number of test items, shorter delay periods).

Learning Trials. Participants entered an individual testing room with the examiner. They were given typical list learning instructions: "I am going to read a list of words to you. When I am through, I will ask you to say as many of them as you can. You can say them in any order. Ready?" The examiner then presented the participants' assigned study list, with each word at a one-to-two-second interval. After being presented with the list, participants were told: "Now tell me all the words you can remember." This process was repeated for a total of three learning trials. For trials two and three, participants were given the following modified instructions: "Now I am going to read the same list again. When I'm through, say as many words as you can remember, including words from the list you said before. You can say them in any order. Ready?" Following the third learning trial, participants were given instructions to answer demographic questions on a computer and to stay in the testing room until the examiner returned.

Participants were left for three minutes. Participants were not made aware of the upcoming short- or long-delay recall trials.

Short-Delay Recall. After the three-minute delay, the examiner returned to the testing room and gave the following instructions: “Remember that list of words I read to you three times? Tell me as many of those words as you can remember now.” Following the short-delay recall trial, participants were told they were moving to the second part of the study and were asked to complete as many mazes as they could before the examiner returned. Mazes were obtained from an online source called Maze Generator. Participants completed mazes for approximately seven minutes (i.e., a ten-minute total delay from the end of learning trials).

Long-Delay Recall. After the seven-minute delay, the examiner returned to the testing room and gave the following instructions: “I want you to say as many of those words as you can remember now.” Following the long-delay recall trial, participants were given an opportunity to ask questions and were assigned their credit on the Participant Pool for participating.

Data Analysis and Results

Data Cleaning and Preparation

For each participant, items were coded as either one or zero for each trial depending on performance (i.e., Learning Trial 1 [LT1], Learning Trial 2 [LT2], Learning Trial 3 [LT3], Short-Delay Recall [SDR], Long-Delay Recall [LDR]). Specifically, for each trial, if the participant appropriately recalled the item, it was coded as one, and if they did not recall the item, it was coded as zero. Average performances were then calculated per word per trial per participant, and an overall average per word was calculated for overall performance. These averages were used as the dependent variables (overall recall [i.e., LT1 + LT2 + LT3 + SDR + LDR], Learning Trials 1, 2, and 3, and overall delayed recall [i.e., SDR + LDR]) in the following analyses.

The values for each variable populated from Study 1 were used as predictor variables in the analysis. Predictor variables included: word frequency, word length, orthographic and phonological neighborhood size, age of acquisition, emotional valence and arousal, imageability, semantic diversity, SND, and sensorimotor information. Each variable was described in Study 1 with reference to the available corpora. Due to the relatively smaller corpora available for some variables, values for some items were not populated (i.e., 22 items for imageability [15.2%], one item for emotional valence and emotional arousal [0.7% each], and three items for semantic diversity [2.1%]. Although Little's Missing Completely At Random (MCAR) test revealed the data was not MCAR ($p < .05$), there were no common patterns observed in the missing values. For emotional valence, emotional arousal, and semantic diversity variables, the average rating for words was substituted to replace missing data. Due to the relatively large amount of missing data (i.e., 15%) for imageability, multiple imputation (i.e., predictive mean matching using the *mice* package in R [R Core Team, 2021; van Buuren & Groothuis-Oudshoorn, 2011]) was used to minimize data loss. The original regression results with listwise deletion were similar to the imputed regression results, and thus, both model fit results are reported below, but the imputed results were interpreted. One item (ice cream) was removed from the analyses, as values were unavailable across most variables due to having two constituents ("ice" and "cream").

Statistical Analyses

First, a correlation analysis was conducted (See Table 12) on the imputed data to determine the extent to which predictor variables correlated with one another and to assess for multicollinearity among the variables. Based on an examination of the correlations, orthographic neighborhood size was removed from the analysis because of its high correlation with phonological neighborhood size ($r = .74$). All other variables remained in the analysis.

Overall Accuracy. The *Enter* method was used to regress the ten predictor variables onto the criterion variable (overall accuracy) to determine the contribution of these variables on overall recall. On examination of the initial output, there were no studentized deleted residuals greater than +/- two standard deviations and no values with Cook's distance above one. One observation had a leverage value of .37, which is considered to be in the "risky" range (Huber, 1981); however, this observation was maintained due to studentized deleted residuals and Cook's distance values within appropriate ranges, and removal did not change the results. A visual inspection of the regression standardized residuals histogram and a P-P plot suggested the data approximated a normal distribution. The Durbin-Watson statistic of 1.88 for the original data and 1.75 to 1.76 for imputed data iterations indicated independence of observations. There was homoscedasticity as assessed by visual inspection of a plot of studentized residuals versus unstandardized predicted values. There was no evidence of multicollinearity observed, as all tolerance values (i.e., for original and imputed data iterations) were greater than .1. The original multiple regression model (i.e., with listwise deletion) significantly predicted memory performance, $F(10, 95) = 4.32$, $p <.001$, $R^2 = .31$, adjusted $R^2 = .24$. Similarly, the multiple regression model was significant for the imputed data, $F(10, 127) = 4.64$, $p < .001$, $R^2 = .29$, adjusted $R^2 = .23$. Three of the ten variables (i.e., emotional valence, imageability, and SND) added to the prediction of overall recall performance, $p \leq .05$. Results suggest that greater emotional valence, imageability, and SND are associated with better recall performance across all trials. Pooled regression coefficients and standard errors can be found in Table 11.

Table 11.

Pooled multiple regression results for overall memory performance.

	B	SE	*t*	Sig.	Correlations		
					Zero-Order	Partial	Part
Constant	-.222	.224	-.993	.323			
WF	-.002	.009	-.265	.792	-.269	-.030	-.026
PN	-.001	.001	-.991	.323	-.236	-.093	-.079
WL	-.001	.007	-.153	.879	.270	-.019	-.016
AoA	.012	.007	1.689	.094	.110	.148	.127
EV	**.026**	**.012**	**2.193**	**.030***	.242	.194	.168
EA	.012	.012	.990	.324	.177	.092	.078
Img	**.041**	**.020**	**2.009**	**.047***	.217	.154	.132
SemD	-.021	.045	-.481	.632	-.345	-.044	-.037
SND	**.184**	**.095**	**1.942**	**.054***	.405	.182	.157
SI	.015	.015	1.053	.295	.229	.103	.087

Note. WF = word frequency, PN = phonological neighborhood, WL = word length, AoA = age of acquisition, EV = emotional valence, EA = emotional arousal, Img = imageability, SemD = semantic diversity, SND = semantic neighborhood density, SI = sensorimotor information. Two-tailed significant predictors are presented in bold (α = .05). There were no significant one-tailed predictors (α = .10; the relationship between AoA and recall was not predicted, given the inconsistencies in the literature).

Table 12.

Pooled correlations of 11 lexical and semantic predictor variables

	ACC	WF	WL	ON	PN	AoA	EV	EA	Img	SemD	SND	SI
ACC	--	-.27*	.27*	-.24*	-.24*	.11	.25*	.18*	.22*	-.34*	.41*	.23*
WF	-.27*	--	-.45*	.29*	.31*	-.33*	.11	.05	-.21*	.48*	-.53*	.06
WL	.27	-.45	--	.67	-.61	.27	.17	.29	-.05	-.25	.36	.12
ON	-.24*	.29*	-.67*	--	.74*	-.32*	-.01	-.12	.04	.20*	-.17*	.01
PN	-.24*	.31*	-.61*	.74*	--	-.33*	-.03	-.21*	.06	.24*	-.24*	.08
AoA	.11	-.33*	.27*	-.32*	-.33*	--	-.25*	.08	-.06	-.22*	-.06	-.39*
EV	.25*	.11	.17*	-.01	-.03	-.25*	--	.16*	-.03	-.21*	.14	.35*
EA	.18*	.05	.29*	-.12	-.21*	.08	.16*	--	-.09	-.03	.15*	.05
Img	.22*	-.21*	-.05	.04	.06	-.06	-.03	-.09	--	-.16*	.19*	.12
SemD	-.34*	.48*	-.25*	.20*	.24*	-.22*	-.21*	-.03	-.16*	--	-.54*	-.23*
SND	.41*	-.53*	.36*	-.17*	-.24*	-.06	.14	.15*	.19*	-.55*	--	.38*
SI	.23*	.06	.12	.01	.08	-.39*	.35*	.05	.12	-.23*	.38*	--

Note. WF = word frequency, WL = word length, ON = orthographic neighborhood size, PN = phonological neighborhood size, AoA = age of acquisition, EV = emotional valence, EA = emotional arousal, Img = imageability, SemD = semantic diversity, SND = semantic neighborhood density, SI = sensorimotor information.

* p < .05.

Learning Trial 1. Similar to the analysis presented above, a multiple regression was conducted to determine the contribution of the ten lexical and semantic variables on memory performance on the first learning trial. As with overall recall, all assumptions were checked and met. Again, the original regression with listwise deletion of missing values was similar to the imputed regression. Both models were significant. The imputed multiple regression model significantly predicted memory performance $F(10, 127) = 6.43$, $p < .001$, $R^2 = .36$, adjusted $R^2 = .31$ (original model: $F(10, 95) = 7.11$, $p < .001$, $R^2 = .42$, adjusted $R^2 = .37$). Four of the ten variables (i.e., age of acquisition, emotional valence, SND, and imageability) added to the prediction of recall at learning trial 1, $p < .05$. Results suggest that higher ratings of age of acquisition, emotional valence, and imageability and higher SND led to better recall performance on learning trial 1. Pooled regression coefficients and standard errors can be found in Table 13.

Table 13.

Pooled multiple regression results for Learning Trial 1.

	B	SE	*t*	Sig.	Correlations		
					Zero-Order	Partial	Part
Constant	-.482	.202	-2.382	.019			
WF	-.006	.008	-.766	.445	-.334	-.074	-.060
PN	-.001	.001	-.844	.400	-.261	-.079	-.064
WL	-.002	.007	-.277	.782	.315	-.029	-.023
AoA	**.017**	**.007**	**2.554**	**.011***	.159	.222	.183
EV	**.033**	**.011**	**3.010**	**.003***	.251	.264	.220
EA	.013	.012	1.197	.234	.213	.109	.088
Img	***.032***	***.0182***	***1.752***	***.083****	.191	.120	.097
SemD	.037	.041	.901	.370	-.316	.082	.066
SND	**.301**	**.087**	**3.439**	**< .001***	.478	.308	.260
SI	.007	.014	.501	.617	.185	.054	.044

Note. WF = word frequency, PN = phonological neighborhood, WL = word length, AoA = age of acquisition, EV = emotional valence, EA = emotional arousal, Img = imageability, SemD = semantic diversity, SND = semantic neighborhood density, SI = sensorimotor information.

Two-tailed significant predictors are presented in bold (α = .05). One-tailed significant predictors are presented in bold and italics (α = .10).

Learning Trial 2. Another multiple regression analysis was conducted to determine the contribution of the ten lexical and semantic variables on memory performance for the second learning trial. All assumptions were checked and met. The imputed multiple regression model significantly predicted memory performance $F(10, 127) = 3.13$, $p < .01$, $R^2 = .22$, adjusted $R^2 = .16$ (original model: $F(10, 95) = 2.74$, $p < .05$, $R^2 = .21$, adjusted $R^2 = .12$). Two of the ten variables (i.e., emotional valence and SND) added to the prediction of recall at learning trial 3, $p < .05$ (one-tailed). Pooled regression coefficients and standard errors can be found in Table 14.

Table 14.

Pooled multiple regression results for Learning Trial 2.

	B	SE	*t*	Sig.	Correlations		
					Zero-Order	Partial	Part
Constant	-.174	.276	-.362	.529			
WF	-.003	.010	-.306	.760	-.230	-.034	-.030
PN	-.001	.001	-.785	.434	-.192	-.074	-.065
WL	-.004	.009	-.453	.651	.217	-.044	-.039
AoA	.015	.009	1.692	.093	.102	.148	.133
EV	***0.28***	***.015***	***1.925***	***.057****	.220	.172	.155
EA	.009	.014	.607	.545	.136	.058	.051
Img	.024	.026	.933	.353	.134	.057	.050
SemD	-.024	.051	-.438	.662	-.318	-.044	-.039
SND	***.202***	***.115***	***1.766***	***.079****	.362	.164	.147
SI	.023	.018	1.268	.207	.225	.122	.109

Note. WF = word frequency, PN = phonological neighborhood, WL = word length, AoA = age of acquisition, EV = emotional valence, EA = emotional arousal, Img = imageability, SemD = semantic diversity, SND = semantic neighborhood density, SI = sensorimotor information. One-tailed significant predictors are presented in bold and italics (α = .10).

Learning Trial 3. Again, a multiple regression analysis was conducted to determine the contribution of the ten lexical and semantic variables on memory performance on the third learning trial. All assumptions were checked and met. The imputed multiple regression model significantly predicted memory performance $F(10, 127) = 2.69$, $p < .01$, $R^2 = .20$, adjusted $R^2 = .13$ (original model: $F(10, 95) = 2.88$, $p < .01$, $R^2 = .23$, adjusted $R^2 = .15$). Two of the ten variables (i.e., imageability and SND) added to the prediction of recall on learning trial 3, $p < .05$ (one-tailed). Pooled regression coefficients and standard errors are presented in Table 15.

Table 15.

Pooled multiple regression results for Learning Trial 3.

	B	SE	*t*	Sig.	Correlations		
					Zero-Order	Partial	Part
Constant	-.010	.260	-.039	.969			
WF	.003	.010	.257	.797	-.209	.016	.015
PN	-.001	.001	-.667	.506	-.191	-.062	-.056
WL	.003	.008	.335	.738	.228	.025	.023
AoA	.004	.008	.494	.622	.039	.040	.036
EV	.019	.014	1.364	.175	.194	.121	.109
EA	.008	.014	.606	.545	.136	.056	.051
Img	***.041***	***.023***	***1.748***	***.083****	.176	.117	.106
SemD	-.050	.052	-.960	.339	-.308	-.089	-.081
SND	***.188***	***.110***	***1.700***	***.092****	.355	.159	.144
SI	.004	.017	.217	.828	.182	.026	.023

Note. WF = word frequency, PN = phonological neighborhood, WL = word length, AoA = age of acquisition, EV = emotional valence, EA = emotional arousal, Img = imageability, SemD = semantic diversity, SND = semantic neighborhood density, SI = sensorimotor information. Two-tailed significant predictors are presented in bold (α = .05; none). One-tailed significant predictors are presented in bold and italics (α = .10).

Overall Delay. Lastly, a multiple regression was conducted to determine the contribution of the ten lexical and semantic variables on overall delayed recall. All assumptions were checked and met. The imputed multiple regression model significantly predicted memory performance $F(10, 127) = 2.94$, $p < .01$, $R^2 = .22$, adjusted $R^2 = .16$ (original model: $F(10, 95) = 2.70$, $p < .01$,

$R^2 = .22$, adjusted $R^2 = .14$). One of the ten variables (i.e., imageability) added to the prediction, $p < .05$. Pooled regression coefficients and standard errors can be found in Table 16.

Table 16.

Pooled multiple regression results for overall delayed memory performance.

	B	SE	*t*	Sig.	Correlations		
					Zero-Order	Partial	Part
Constant	-.184	.267	-.688	.493			
WF	-.002	.011	-.177	.860	-.226	-.022	-.020
PN	-.001	.001	-.498	.620	-.181	-.051	-.046
WL	.002	.009	.234	.815	.216	-.002	-.002
AoA	.011	.009	1.186	.238	.106	.105	.094
EV	.019	.015	1.273	.205	.181	.124	.112
EA	.019	.015	1.312	.192	.171	.117	.105
Img	**.061**	**.023**	**2.652**	**.009**	.231	.182	.166
SemD	-.059	.056	-1.053	.295	-.311	-.092	-.082
SND	.093	.119	.787	.433	.308	.087	.078
SI	.014	.018	.758	.450	.175	.071	.064

Note. WF = word frequency, PN = phonological neighborhood, WL = word length, AoA = age of acquisition, EV = emotional valence, EA = emotional arousal, Img = imageability, SemD = semantic diversity, SND = semantic neighborhood density, SI = sensorimotor information. Two-tailed significant predictors are presented in bold ($\alpha = .05$). One-tailed significant predictors are presented in bold and italics ($\alpha = .10$; none).

STUDY 1 & 2 DISCUSSION

The goal of Studies 1 and 2 was to examine the psycholinguistic characteristics of common list learning tests and to explore the effects of these characteristics on memory performance in healthy participants. Results from Study 1 clearly demonstrate the wide variety of differences in values across most variables in most of the test lists examined; specifically, most test lists have several variables with values ranging more than a standard deviation away from each other. The results from Study 2 indicate that out of the ten variables analyzed (four lexical and six semantic), emotional valence, imageability, and SND predicted overall recall across learning and delayed memory trials. All three variables had a positive relationship with recall – meaning that more associated semantic information (in the form of higher emotional valence and imageability and denser semantic neighborhoods) supported better recall. Further, the methodology of Study 2 allowed for a breakdown of when specific variables become influential (i.e., which learning and/or delayed recall trial). Results found that emotional valence, SND, imageability, and age of acquisition predicted recall at learning trial 1. Again, these variables had a positive relationship with performance; there is an advantage for words with higher emotional valence ratings (i.e., more positive words), denser semantic neighborhoods, higher imageability ratings, and words with older age of acquisition. Similarly, emotional valence and SND continued to influence recall at learning trial 2 and imageability and SND at learning trial 3. Given the temporal sequence, the results suggest that the effect of age of acquisition is most prevalent at initial encoding (i.e., first time hearing the list), whereas emotional valence, imageability, and SND support memory even as information is repeated. Lastly, imageability was found to independently predict overall delayed recall, suggesting that words with higher imageability either aid the retrieval process at delay or are more likely to "stick" in memory.

These findings are generally in line with experimental memory research. As described in depth in Chapter 2, for mixed-list recall specifically, there are consistent effects for semantic richness (e.g., emotional information, imageability, number of features). Although there is no published research analyzing the effects of SND on recall, it has been found to play a role in recognition memory (Wong Gonzalez, 2018). More specifically, previous research has consistently found an advantage or facilitatory effect for high imageability words (Lau et al., 2018), emotional words (Doerksen & Shimamura, 2001; Lau et al., 2018), words with dense semantic neighborhoods (recognition only; Wong Gonzalez et al., 2018), and words with more associated features (Hargreaves et al., 2012; Lau et al., 2018) on memory performance. It is not surprising that these strong semantic effects also aid recall in the current study.

There is less consistency in the research regarding the impact of lexical variables (e.g., word frequency, orthographic/phonological neighborhood size, word length, and age of acquisition) on memory, and the results of Study 2 may offer some further explanation. First, previous research has found conflicting effects of age of acquisition on mixed-list recall; some studies have suggested an advantage for words learned later in life (Dewhurst et al., 1998; Morris, 1981), while others have found no effect (Gilhooly & Gilhooly, 1979; Macmillan et al., 2021). In the current study, there was a late age of acquisition advantage, but only for learning trial 1, suggesting that this effect may be weakened with increased opportunity for rehearsal. In the two studies that found no effect of age of acquisition on recall of mixed lists, the methodology theoretically allowed more time for additional rehearsal compared to the current study. For example, Macmillan and colleagues (2021) presented lists with only 12 items, which would logically be easier to rehearse compared to the 42 to 47 items in Study 2. Further, although participants were presented a similar number of items in Gilhooly and Gilhooly's (1979) study, participants were given the recall test ten seconds following list presentation and were allowed ten minutes to write down as many items as they could remember. Theoretically, participants could spend much of those ten minutes rehearsing items from the list. Thus, the relationship between age of acquisition and memory could be mediated by rehearsal.

Like age of acquisition, the effects of word frequency, orthographic/phonological neighborhood size, and word length on mixed-list recall have been inconsistent. Regarding word frequency, some studies have found a recall advantage for high frequency words (Balota &

Neely, 1980; Hicks et al., 2005), others have found a recall advantage for low frequency words (Delosh & McDaniel, 1996; Merritt et al., 2006; Ozubko & Joordens, 2007), while still, others have found no effect of word frequency on recall at all (May et al., 1979; Ward et al., 2003; Watkins et al., 2000). Lohnas and Kahana (2013) suggested that the relationship may actually be U-shaped, as they found that both high and low frequency words were better recalled compared to midfrequency words. Upon visually inspecting the makeup of the test lists (Study 1), word frequency is one of the most adequately managed variables in the test lists examined herein; many of the items appear to be midfrequency with some high frequency words in each list. Therefore, the lack of a word frequency effect in Study 2 could be due to the large number of midfrequency words in the stimulus set. Further, given that the current project analyzed word frequency as a continuous rather than categorical (e.g., high vs. low, or high vs. low vs. mid) variable, it is possible that the data follows this U-shaped trend, and any high or low frequency effects were balanced out. In terms of orthographic/phonological neighborhood size, the current results are consistent with recent research by Ballot et al. (2021), whereby semantic information, specifically imageability, was more influential for recall than orthographic distinctiveness. Similar to those presented above, findings are inconsistent regarding the effect of word length on mixed-list recall. A study conducted by Jalbert and colleagues (2011) found that any effect of word length was eliminated after controlling for orthographic neighborhood size, suggesting that word length may be a proxy for orthographic distinctiveness, which, as mentioned above, may be a lower level processing benefit overpowered by rich semantic information.

Across all analyses, four out of the ten variables used in the current study were influential for recall at some time point (e.g., either for learning trials 1, 2, 3, delayed recall, or overall). Therefore, these four variables (emotional valence, SND, imageability, and age of acquisition) are those that appear to be most concerning for test construction. Differences in Z-scores (i.e., the lowest Z-score for that variable within that list subtracted from the highest) for each of these four variables on each test are presented in Table 17. Visual examination of these differences, even with a conservative approach (difference of 2 SDs), shows that nine of the ten lists were constructed with large differences in at least one influential variable. Age of acquisition is particularly important for List "B" construction given that it is only ever administered one time (i.e., single trial), and age of acquisition was shown to be especially influential on initial encoding (i.e., learning trial 1).

Table 17.

Difference between the highest and lowest Z-score for each variable in each test list.

Test	List	AoA	EV	Img	SND
CVLT-C	A	1.3	2.2	5.1	2.69
CVLT-C	B	1.13	1.96	.35	2.91
CVLT-3	A	1.31	1.5	.34	2.2
CVLT-3	B	1.67	2.02	.64	2.75
ChAMP	A	2.31	2.12	1.24	1.96
WRAML-2	A	.86	2.89	1.08	2.42
RBANS	A	1.14	1.82	2.3	2.11
NEPSY-II	A	1.92	2.19	1.35	2.41
NEPSY-II	B	1.68	1.94	1.2	2.8
HVLT-R	A	1.8	1.72	.54	1.75

In line with the Boston Process Approach, recent research has begun to demonstrate that a finer-grain analysis of patient responses provides more sensitivity to the discrimination of some disorders compared to traditional scoring systems. For instance, Taler et al. (2021) used a discourse-theoretic approach to analyze story recall responses of both younger and older adults and older adults and those with MCI. Story recall tasks are typically scored by summing the total number of units recalled (a unit was either recalled or not; binary system). Taler and colleagues (2021) compared the unit-based scoring approach to a more in-depth scoring system, in which they distinguished between veridical (i.e., verbatim), gist (i.e., not verbatim but general meaning), and distorted responses (i.e., incorrect). The authors found that healthy older adults had greater recall for both veridical and gist information compared to patients with MCI. Further, older adults' responses had more gist information and distortions at delayed recall compared to immediate, whereas patients with MCI showed a decline in their gist and distorted responses at delay. Thus, the reduction in gist and distorted responses over time could be a sensitive marker for MCI (Taler et al., 2021) that would likely be missed when using a more traditional scoring approach. Moreover, as discussed previously, research using experimental tasks has revealed specific deficits related to certain patient populations. Patients with disorders including MCI, Alzheimer's disease, semantic dementia, primary progressive aphasia, and Parkinson's disease show deficits in semantic processing (Angwin et al., 2006; Croisile et al., 1996; Huff et al., 1986; McNamara et al., 1992); however, these deficits can be distinguished using a more detailed analysis (Price & Grossman, 2005; Vandenberghe et al., 2005). Further, semantic processing has also been shown to differentially affect individuals as they age, as well as individuals who speak more than one language (Johns et al., 2016). Additionally, psycholinguistic research has shown differential processing for other lexical and semantic information across patient populations (e.g., word frequency, phonological processing). As demonstrated by the results of Taler et al. (2021) discussed above, analyzing patient responses at a more specific level can lead to better sensitivity of neuropsychological measures.

Given these differences and the relative lack of knowledge of and consideration for how psycholinguistic characteristics can affect clinical data in neuropsychological research and test development, these results are important to consider. The results of Study 1 demonstrate that items in many of the test lists examined were not adequately controlled across psycholinguistic variables. Even for those tests that attempted to control for semantic distinctiveness (RBANS) or prototypicality (CVLT-C, CVLT-3, HVLT-R) in their test items, most of them (HVLT-R excluded) still have items with considerable variability across semantic variables, suggesting their methods to do so were unsuccessful. The results of Study 2 demonstrate that specific psycholinguistic properties affect performance on neuropsychological test lists for healthy populations, which could have important clinical relevance. First, certain patient populations could be more sensitive to the underlying psycholinguistic structure of the test lists, ultimately confounding true "memory" results. For example, given the known deficient semantic processing in Alzheimer's disease, it is likely that the performance of patients with Alzheimer's disorder would look different when using a list composed of items that differ in their semantic richness (i.e., current neuropsychological tests) compared to using a list composed of items that are well controlled for semantic information. Put another way, specific patient populations may be pre-emptively at a disadvantage based on variables related to how tests are constructed. Second, clinicians could be missing out on additional data that would allow for earlier differential diagnosis. Specifically, if neuropsychological test lists were developed that thoughtfully manipulated a single psycholinguistic property at a time, process scores could be developed that help clinicians understand what type of information is contributing to their patient's scores. For example, if a test list was well controlled across semantic and orthographic characteristics and manipulated the phonological property of the items, a score could be developed that indicates the impact of such manipulation. Therefore, when assessing an individual with dyslexia, for example, a clinician could use their results to conclude that although general rote memory is intact, memory for words with high phonological complexity is not. Future research would be needed to further examine how such information could be used to aid early detection of certain disorders.

CHAPTER 3:

EFFECTS OF LEXICAL AND SEMANTIC VARIABLES ON VERBAL FLUENCY

Verbal Fluency tests are commonly used in neuropsychological batteries as measures of language productivity, semantic memory, and executive functioning in the assessment of several disorders (Strauss et al., 2006). These tests typically consist of phonemic fluency tasks, where participants are asked to say words that begin with a specific letter within a specified timeframe, and semantic fluency tasks, where participants are asked to say words that belong to a designated semantic category within a specified timeframe (Strauss et al., 2006). They are typically scored by counting the number of correct items produced, and, in some cases, the number of correct items produced in several letters or categories is pooled to obtain an overall phonemic or semantic fluency score. Thus, although number of set-loss errors (i.e., producing incorrect items) and repetition errors are typically examined, the specific items produced, and their lexical properties are not necessarily considered. Important to the current study, clinicians have also begun including a category switching verbal fluency task in assessments to specifically target executive functioning (i.e., cognitive flexibility and shifting). These tasks require examinees to fluently switch between one semantic category and another. For example, in a popular neuropsychological test battery aimed at measuring different aspects of executive functioning, the D-KEFS (Delis et al., 2001), examinees are required to switch between producing as many *fruits* and as many *pieces of furniture* as they can in 60-seconds. Clinicians can then obtain a score for both total correct responses and category switching accuracy.

Anecdotally, neuropsychologists often recognize a pattern of more efficient production of items from certain categories (e.g., *fruit*) over others (i.e., *furniture*). This pattern is often seen for categories in both standard semantic fluency tasks and category switching tasks, although the impact of this difference appears greater on the latter given that a disruption in fluent production of items in only one of the two categories necessarily slows the examinee down and impacts their overall raw score. Chapter 2 of this project will explore and examine potential lexical and semantic variables, as well as the underlying psycholinguistic structure of the categories themselves, that may be contributing to differences in fluency ability.

Theoretical Mechanisms Behind Verbal Fluency

Given the popularity of the verbal fluency task in clinical practice and the great diagnostic utility of the task in some disorders, researchers have begun to explore the potential neuropsychological mechanisms behind the retrieval of correct items. Troyer and colleagues (1997) observed that people tended to produce items in bursts or “clusters” of semantically related words and that pauses between “cluster switches” were longer than pauses between items within the same cluster. For example, in the semantic fluency category *animals*, participants produced a set of animals that could be considered house pets (e.g., cat, dog, fish) before switching to African animals (e.g., elephant, zebra, cheetah) before switching to farm animals (e.g., pig, cow, horse) and so on. Based on this pattern, Troyer et al. (1997) suggested that the fluent retrieval of items required two processes: one process that produces words within a cluster and one that jumps between related clusters. Therefore, although the total number of correct responses is still an important piece of the task, understanding differences in cluster sizes and number of cluster switches would provide a more comprehensive and fine-grained analysis of responses. Troyer and colleagues (1998) went on to use this approach to distinguish qualitative differences in verbal fluency between patients with Alzheimer’s disease and Parkinson’s disease. Other research has also used this approach to discover fluency abnormalities in patients with Huntington’s disease (Ho et al., 2002) and traumatic brain injury (Zakzanis et al., 2011).

Based on this pattern of clustering, Hills et al. (2012) used a computational model to compare the process of searching through semantic memory for an appropriate item to the way animals forage for food (i.e., optimal foraging). They argued that participants switch categories when the value of finding another item within their current cluster becomes less than the expected value of searching in a different cluster (marginal value theorem; Charnov, 1976). Specifically, participants make a strategic decision to switch to a new cluster when the current one becomes depleted. Hills and colleagues (2012) had participants produce as many animals as they could within three minutes. Using a computational approach, the authors were able to analyze the retrieval path through memory as indicated by the time between items. Participants whose response times did not indicate the use of the marginal value theorem (i.e., switched clusters either too early or too late) produced significantly fewer words than those who did. Based on their results, Hills et al. (2012) proposed that semantic fluency output, and therefore semantic memory, is based on spatial representations and requires two distinct processes, "clustering" and "switching," with the strategy for switching following that described above.

Alternatively, Abbott and colleagues (2015) proposed that the behavioral phenomenon described by Troyer et al. (1997) could be explained by a "random walk" on a semantic network. Random walks suggest an undirected search through memory (Anderson, 1972), as opposed to strategic (like optimal foraging). Abbott et al. (2015) showed that results similar to those found by Hills et al. (2012) are produced by a random-walk model on a network where words that are semantically related are close together in that network (e.g., free association data). Moreover, their results <u>did not</u> mimic optimal foraging when the random-walk was completed on a spatial network, suggesting the importance of the semantic network. Therefore, Abbott and colleagues

(2015) posit that semantic memory is represented by a network, and fluency output (or search) is solely dependent on a random walk on that network.

As discussed above, analyzing patients' fluency output in terms of clusters and switches has been shown to provide insight into patient differences above and beyond what can be garnered from the total number of correct responses. However, there are several drawbacks to using this approach: it is vulnerable to problems in inter-rater reliability given the subjective nature of determining clusters, the resulting analysis is relatively crude due to the binary scoring system (i.e., the item is within a cluster or not; Taler et al., 2020), and it does not allow for a distinction between a processing speed deficit and a more specific cluster-switching deficit (Mayr, 2002). Therefore, the development of computation models and approaches, like those discussed above, address those drawbacks and allow for a fine-grained analysis of fluency output that can help researchers better understand underlying processes (Taler et al., 2020).

Influence of Lexical and Semantic Variables on Semantic Fluency

As discussed at length in Chapter 2, several lexical and semantic variables influence language processing and memory. The following section explores the effects of those same variables in relation to verbal fluency output. Based on a review of the available literature, it appears that clinical researchers have only recently begun to consider psycholinguistic variables in their methodology. There is extensive research on verbal fluency output and performance; however, it has largely focused on comparing and understanding deficits in patient populations and disentangling the underlying latent variables of the task. Despite the research in a vast array of patient and demographic populations, very few areas of verbal fluency research have attempted to understand the contributing effects of various word-level characteristics on verbal fluency performance. Thus, there is a significant gap in the literature regarding the influence of psycholinguistic properties on verbal fluency and specifically in neurotypical participants. Given the lack of targeted investigation of these psycholinguistic characteristics, the following section outlines the available and relevant literature organized according to pertinent patient populations rather than specific variables, as in Chapter 2.

As discussed briefly in Chapter 1, patients with Alzheimer's disease exhibit diminished semantic processing (Croisile et al., 1996; McNamara et al., 1992). Although verbal fluency generally declines with healthy aging, the decline for individuals with Alzheimer's is qualitatively different. Specifically, healthy older adults typically perform better on semantic fluency compared to phonemic fluency tasks (Rosen, 1980), whereas in patients with Alzheimer's disease, semantic fluency performance declines much more rapidly than phonemic fluency performance (Butters et al., 1988; Hart et al., 1988; Henry et al., 2004; Monsch et al., 1992). Patients with Alzheimer's disease also make more perseverative and set-loss errors compared to healthy controls (Forbes-McKay et al., 2005). Further, performance on semantic fluency tasks has been found to be impaired before the onset of the disease (Vogel et al., 2005) and to decline as a patient's disease advances (Perry et al., 2000). Henry and colleagues (2004) suggest that the specific deficit in semantic fluency for patients with Alzheimer's disease is due to the progressive degeneration of semantic knowledge, as the task necessarily requires intact semantic processing for associations (Rohrer et al., 1999). Semantic fluency has been found to be a profoundly sensitive measure of semantic memory, with sensitivity at 100% and specificity at 92.5% for detecting the presence of Alzheimer's disease (Monsch et al., 1992). As such, research has begun to investigate specific aspects of semantic fluency performance beyond the traditional number of correct responses that may help to discriminate Alzheimer's patients from other populations. Forbes-McKay and colleagues (2005) examined differences in lexical (i.e., word length, word frequency, age of acquisition) and semantic (i.e., typicality, which reflects the degree to which the item represents a good exemplar of the semantic category) characteristics of words produced by healthy older adults and patients with varying degrees of Alzheimer's disease in a semantic fluency task (categories: *animals, fruit*). Words produced by patients were shorter in length, had earlier age of acquisition, higher frequency, and higher typicality ratings of the semantic category compared to their healthy counterparts (Forbes-McKay et al., 2005). Moreover, although each of those lexical and semantic variables was found to predict group membership to some degree, the strongest predictor was age of acquisition, even above and beyond the number of words generated (Forbes-McKay et al., 2005). Venneri and colleagues (2008) found the same pattern of results, as they demonstrated effects of typicality, word frequency, and age of acquisition on the words produced by patients with early Alzheimer's disease. The authors also demonstrated a significant correlation in the relationship between these lexical variables impacting early Alzheimer's patients' verbal fluency and regions of the medial temporal lobes, particularly in areas of the perirhinal and parahippocampal cortex (Venneri et al., 2008). Further, Sailor et al. (2011) compared words produced on a semantic fluency task (category: *vegetables*) and phonemic fluency task (letter: *F*) between patients with Alzheimer's disease and healthy controls. Overall, they found that words produced in the semantic fluency task had earlier age of acquisition, while words produced in the phonemic fluency task had higher word frequency, and differences in age of acquisition between patients and controls were larger for semantic fluency tasks than phonemic fluency tasks. Research by Gomez and White (2006) found that measures including clustering, cluster switching, and size of clusters were able to discriminate healthy aging from even mild Alzheimer's disease. Taken together, findings in the Alzheimer's research highlight the potential importance of examining the specific items produced by examinees, in addition to total scores.

Relatedly, Mild Cognitive Impairment (MCI) is a disorder that is characterized by impairments in memory without dementia (Petersen et al., 1999) and is known to be a strong predictor for the development of Alzheimer's disease (Peterson et al., 2001) making it an important early marker and target for early intervention. Although the pattern of results is not as clear as those found in Alzheimer's research, there appear to be language-based, and more specifically, semantically related deficits associated with MCI (see Taler & Phillips, 2008 for a review). Using the cluster-based interpretation approach (Troyer et al., 1998) discussed previously, Murphy and colleagues (2009) found that semantic fluency output in MCI patients has more disorganization within clusters and fewer cluster switches compared to healthy controls. Further, Johns et al. (2018) used a novel computational model to comprehensively examine changes in verbal fluency performance over time in a population of older adults in which some went on to develop MCI (pre-MCI patients), and some did not (healthy controls). The participants completed a semantic fluency test (category: *animals*) annually as part of a larger neuropsychological battery. Neither the number of correct responses nor the slope of the number of items across assessments significantly predicted the development of MCI. Rather, cognitive modeling demonstrated that changes in parameter values across time differentiated those who went on to develop MCI and those who did not; specifically, pre-MCI patients produced a pattern of items that were guided by high frequency cues that are from closely connected regions of semantic memory (Johns et al., 2018). These findings suggest, again, that there is important and potentially wasted diagnostic information lying beneath the traditional scoring system.

Semantic dementia is another form of dementia that is specifically characterized by deficient semantic memory (Hodges et al., 1992). These patients show a specific lexical and semantic language impairment with relatively intact syntactic and phonological language (Hodges et al., 1994). The deterioration of lexical and semantic processing is progressive and appears to show an effect of frequency, where low frequency words are lost first, followed by high frequency words as the disease progresses (Bird et al., 2000). Not surprisingly, patients with semantic dementia demonstrated decreased performance on semantic fluency tasks relative to controls and other patient populations (Laisney et al., 2009). Moreover, patients with semantic dementia have shown preserved episodic memory (Graham & Hodges, 1997), which would theoretically predict fewer perseveration errors compared to other dementia populations (Marczinski & Kertesz, 2006).

Primary progressive aphasia is a neurological condition that is characterized by the gradual loss of language function, often beginning with anomia (i.e., word-finding difficulties) and gradually progressing to impaired fluency and even mutism in some cases (Kertesz et al., 2003). Rather than semantic impairment, patients with primary progressive aphasia have deficits in access to phonological information (Mendez et al., 2003). As would be predicted, patients with this specific disease showed greater impairment in phonemic fluency compared to semantic fluency (Scheffel et al., 2021).

Marczinski and Kertesz (2006) investigated differences in verbal fluency performance and the effect of word frequency across the three patient populations described above (i.e., those with Alzheimer's disease, semantic dementia, and primary progressive aphasia). Foremost, all three patient populations showed reduced verbal fluency output in both phonemic and semantic fluency tasks (i.e., categories: *animals, groceries;* letter: *S*) compared to healthy controls. In all three tasks, patients with Alzheimer's disease generated more items than both semantic dementia and primary progressive aphasia patients. There were no differences found in semantic fluency between patients with semantic dementia and primary progressive aphasia; however, those with semantic dementia showed better performance on the phonemic fluency task compared to patients with primary progressive aphasia. These results are consistent with the documented impairments in each of these patient populations, including phonological access in primary progressive aphasia and semantic representations in semantic dementia (Marczinski & Kertesz, 2006). Relevant to the current study, healthy controls produced lower frequency items, followed by those with Alzheimer's disease, those with primary progressive aphasia, and those with semantic dementia producing higher frequency items.

Another patient population that has been found to demonstrate impairment in verbal fluency is those with schizophrenia. Language disorganization is a hallmark feature of schizophrenia (Bleuer, 1950), making verbal fluency tasks especially useful when assessing neuropsychological deficits and impairment in the population. Several studies have found that patients with schizophrenia produce fewer words overall on both phonemic and semantic fluency tasks compared to control groups (Heinrichs & Zakzanis, 1998; Bokat & Goldberg, 2003; Henry & Crawford, 2005; Vogel et al., 2009); however, the mechanisms responsible for the difference in performance are less clear. A meta-analysis conducted by Doughty and Done (2009) revealed that studies typically demonstrate similar impairment across phonemic fluency, standard semantic fluency, and category switching tasks. As a result, the authors suggest that patients with schizophrenia may experience deficits in both executive functioning and semantic knowledge (Doughty & Done, 2009). Findings from other studies suggest that patients with schizophrenia have distinct semantic network patterns, despite intact semantic representations (i.e., deficient connections between representations; Holmes & Ellis, 2006; Kiang & Kutas, 2006; Kiang et al., 2012). Researchers have also considered the potential of a size reduction in the overall mental lexicon. Some research has found no difference in the size of the lexicon (i.e., number of unique words produced over a series of trials) between patients and controls, despite overall lower output (Allen et al., 1993; Elvevag et al., 2001). In contrast, other studies using different semantic tasks (e.g., picture naming) have found smaller overall lexicon sizes in patients with schizophrenia (Chen et al., 2000; Leeson et al., 2005). Thus, although impairments in verbal fluency have been well documented within this population, the cause of such deficits remains ambiguous. This ambiguity has led some researchers to begin exploring the lexical characteristics of generated items.

Paulson et al. (1996) found that both healthy controls and patients with schizophrenia generated largely high frequency items in a semantic fluency task (category: *animals*). In contrast, in a study conducted by Baskak and colleagues (2008), patients with schizophrenia produced more "peculiar" words compared to controls on a phonemic fluency task. Although the operationalization of "peculiar" is inherently different than word frequency, the findings may suggest that patients with schizophrenia produce lower frequency items (Baskak et al., 2008). Juhasz et al. (2012) examined lexical (i.e., word frequency, age of acquisition, word length) and semantic (i.e., typicality) properties of words generated in a semantic fluency task (category: *animals*) by patients with schizophrenia compared to demographically matched controls. Results showed the expected overall reduction in both phonemic and semantic fluency output for patients with schizophrenia, as well as effects of age of acquisition (more early acquired words produced by patients) and typicality (more typical exemplars produced by patients) in the semantic fluency task only. Interestingly, no differences were found based on word frequency or word length in either fluency task (Juhasz et al., 2012).

Research on autism spectrum disorder has also led to ambiguous results on verbal fluency tasks – some studies suggest impaired verbal fluency in the population, while others suggest the ability is intact (e.g., Beacher et al., 2012; Borkowska, 2015; Spek et al., 2009). Due to these inconsistencies, Tóth and colleagues (2022) examined the data in a more comprehensive way, investigating the idea that there may be differences in imageability and/or concreteness of the words produced by the patients compared to healthy controls. Although the authors did not find underlying differences in these variables, they suggest that future research should continue to use a comprehensive approach (i.e., analyzing qualitative differences in the words produced) to better understand the ambiguities in the literature.

A recent study by Henderson et al. (2023) examined performance on semantic fluency and phonemic fluency tasks and psycholinguistic properties (i.e., word frequency, imageability, age of acquisition, familiarity, concreteness, word length, orthographic distance, and phonological distance) across several different patient populations (i.e., Alzheimer's disease, behavioral variant frontotemporal dementia, primary progressive aphasia, progressive supranuclear palsy, corticobasal syndrome, and healthy controls). The authors found that overall, the total words produced differentiated controls and patient populations and was associated with severity of the disease. Additionally, word frequency was the only psycholinguistic variable to add diagnostic clarification. Frequency was a moderately strong discriminator in both phonemic and semantic fluency tasks for patients with primary progressive aphasia, as these patients were more likely to produce higher frequency words (Henderson et al., 2023).

Taler et al. (2020) conducted a large-scale study examining the role of psycholinguistic variables on semantic fluency output (category: *animals*) across the aging spectrum. The authors used data from the Canadian Longitudinal Study on Aging (CLSA) which included fluency data for a sample size over 10,000 and an age range of 45- to 85-years-old. Using computational modeling, Taler and colleagues (2020) found that at the participant-level, the best predictor of aging was pairwise similarity, followed by total correct responses, number of Troyer clusters (discussed above), and average word frequency. Further, at the item-level, they found that the best predictor of how often a word was produced by their participants was the word's semantic neighborhood, followed by word frequency (Taler et al., 2020). Thus, the authors identified several additional measures that are fruitful for further understanding the experience of aging and further address the idea that a closer examination of data can provide a more nuanced clinical interpretation.

Research has also documented differences in semantic fluency output in bilingual speakers. Specifically, bilingual speakers, regardless of age, typically show reduced output on semantic fluency tasks compared to individuals who speak only one language (monolinguals), even when the task was administered in their native language (Bialystok et al., 2008; Gollan et al., 2002; Portocarrero et al., 2007; Rosselli et al., 2002). Performance on phonemic fluency is less consistent, with some research showing the same pattern (i.e., bilinguals have fewer responses; Bialystok et al., 2008; Gollan et al., 2002), while others document similar performance between bilinguals and monolinguals (Portocarrero et al., 2007; Rosselli et al., 2000). Sandoval et al. (2010) reported that words produced by bilingual speakers had lower frequency and were often cognates (i.e., words with similar form in their two spoken languages). Further, Taler et al. (2013) used a computational approach to examine verbal fluency output for

English monolinguals and English-French bilinguals. The authors used several fluency tasks that varied depending on executive demands. For the standard semantic fluency task, both groups performed similarly; however, in a condition with high executive demands (i.e., switching between English words and French words), bilinguals produced responses that had higher word frequency and lower semantic similarity compared to monolinguals. Taler and Johns (2022) used data from the CLSA to analyze more subtle differences in semantic fluency (category: *animals*) between English monolinguals, bilinguals with English as their first language (L1 bilinguals), and bilinguals with English as their second language (L2 bilinguals). Results showed that L1 bilinguals had the best performance, followed by monolinguals, and then L2 bilinguals. Further, L1 bilinguals produced items that had lower word frequency and items with lower semantic similarity between them compared to monolinguals, and L2 bilinguals produced items with higher word frequency and pairwise semantic similarity (Taler & Johns, 2022). Taken together across studies, Taler and colleagues suggest that as memory search becomes more difficult (high executive demands) or weaker, people produce items that are higher in frequency and closer together in semantic space due to diminished access to low frequency words and semantic connections (Taler et al., 2013; Taler & Johns, 2022).

Interestingly, researchers have begun to investigate emotion as playing a role in verbal fluency output. In a study conducted by Fishman and colleagues (2013), apathy was found to predict semantic but not phonemic fluency performance after controlling for depressive symptoms in patients with cerebrovascular disease. Sass et al. (2013) proposed a new "Emotional Verbal Fluency" task to examine the relationship between emotion and executive functioning. The authors had participants complete a standard semantic fluency task (categories: *plants, toys, vehicles*) and an emotional fluency task, where participants produce words that trigger certain emotions (emotions: *joy, anger, sadness, fear, disgust*). There were no differences in total correct responses between tasks, suggesting that the two tasks were comparable in terms of difficulty. For the emotional categories, *joy* generated the most correct items suggesting a positivity bias for healthy control participants. Thus, the authors conclude that more research is needed, but a new "Emotional Verbal Fluency" task could include fruitful information to delineate populations who suffer from emotional deficits/challenges (Sass et al., 2013).

An important concept and variable that is not often accounted for in verbal fluency research is category size (i.e., the total number of items in a category), which has been shown to affect fluency in other recall tasks (Wixted & Roher, 1994). Although many studies have found informative differences in performance on verbal fluency tasks between patients with Alzheimer's disease and controls or patients with schizophrenia and controls, Diaz et al. (2003) suggest the utility of such findings may be marginal due to methodological flaws. First, the authors point out the use of overlapping categories (e.g., in a review of 22 studies, 17 used *F, A*, or *S* for letters in the phonemic fluency task, and 21 included *animals*, *fruits*, or *vegetables* as semantic categories in the semantic fluency task). These data are then almost always aggregated across categories (i.e., responses are not analyzed separately for each category); this practice makes results difficult to interpret meaningfully across studies as they become vulnerable to differences in category size based on the specific category used in each study (Diaz et al., 2003). For example, Bayles et al. (1989) examined performance on categories individually and found that participants generated the most responses for *animals*, followed by the letters *S, F,* and *A*, followed by *vegetables,* and lastly, *fruit*. Importantly, when collapsed across categories, there was not a significant difference between semantic and phonemic fluency categories. But, if performance for only the letter *S* was used for phonemic fluency and only performance for the category *fruit* was used for semantic fluency, results would suggest an advantage of phonemic over semantic fluency. Moreover, the opposite would be true of the comparison between the letter *A* and the category *animals* (Bayles et al., 1989). Thus, Diaz and colleagues (2003) directly investigated the effects of category size on semantic and phonemic fluency performance in patients with Alzheimer's disease. Rather than aggregating across letters and categories, Diaz and colleagues (2003) used a hierarchical linear model to analyze differences in fluency across categories. Overall, the patient group still demonstrated a more severe deficit in semantic fluency compared to phonemic fluency, as they recalled more items for the letter categories than from the semantic categories. Healthy controls showed the opposite effect, as they recalled more items from semantic compared to letter categories. Further, Diaz et al. (2003) found that although this difference held stable after controlling for category size, the category size slope was more than double for controls compared to patients, meaning that patients with Alzheimer's disease will show more severe deficits for larger categories than for smaller ones (Diaz et al., 2003).

Although there have not been many thorough examinations covering the effects of psycholinguistic variables on verbal fluency tasks, there is significant evidence to suggest that differences in response to lexical (i.e., word frequency, word length, age of acquisition) and semantic (i.e., typicality, emotional information) properties could be useful for understanding impairments and distinguishing similar patient populations beyond what can be garnered from standardized normative data.

STUDY 3

Objective

The objective of Study 3 was 1) to determine whether differences exist in total responses and errors between seven commonly used categories in semantic fluency tasks (i.e., *animals, clothing, fruit, vegetables, furniture, instruments,* and *things you can eat or drink*) and 2) to explore potential underlying differences in the semantic and lexical structure of those categories. It was hypothesized that there would be differences in the total responses and errors between categories, specifically set loss errors. It was also hypothesized that the semantic and lexical structure of the categories differ from each other, especially in semantic richness variables.

Participants

Participants were undergraduate students at the University of Windsor who volunteered to participate in exchange for bonus credits toward their eligible courses. All participants were required to report English as their first language. As described below, the data for Study 3 was collected in two phases (~one year apart). In phase one, 102 participants were recruited. Two participants' data were removed due to technical issues during administration (i.e., audio loss, poor internet connection affecting timing). In phase two, 51 participants were recruited. One participant's data was removed due to failure to meet inclusion criteria (i.e., English was not reported to be their first language). In the final sample (N = 150), 42 participants (28%) reported being bilingual (with English as their first language).

Procedure

Before collecting data, Study 3 was approved by the University of Windsor's REB. Participants volunteered to participate through the University of Windsor Participant Pool and provided consent through a Qualtrics survey. They completed the study via Microsoft Teams with a live research assistant. After joining the meeting, participants answered eligibility and demographic questions. Participants were given 60 seconds to verbally produce as many items from a specific semantic category as possible. Of note, data was collected in two phases, approximately one year apart, due to additional categories being added (i.e., *instruments, vegetables*, and *things you can eat/drink* categories were added) for the sake of completeness of this dissertation and to allow for more specific recommendations regarding test development. In phase one, 100 participants completed the task for four semantic categories (i.e., *furniture, fruits, animals, clothing*). In phase two, a total of 50 participants completed the task; 25 completed the task for four categories, including *instruments, vegetables, things you can eat or drink,* and *furniture,* while the other 25 completed the task for *instruments, vegetables, things you can eat or drink*, and *clothing.* This procedure with two overlapping categories (i.e., *clothing* and *furniture*) was developed to ensure there were no relevant differences between the demographics of participants and total response counts and errors in the data collected in the two phases (see *Preliminary Analyses* section). The seven categories were chosen based on the popularity of use in common neuropsychological test batteries (e.g., D-KEFS [standard and alternate forms], NEPSY-II, Woodcock-Johnson – Fourth Edition). Before beginning each category, participants were instructed: "When I say begin, I want you to tell me as many (INSERT CATEGORY) as you can. You will have 60 seconds before I tell you to stop. Do you have any questions? Ready? Begin." The order of the categories was randomly presented for each participant.

Participants' responses for each category were then scored for accuracy using the scoring rules from the D-KEFS scoring manual (Delis et al., 2001). For *furniture* and *clothing*, only items that could be purchased at typical furniture and clothing stores were counted as correct (e.g., appliances, foundational items [e.g., countertops, doors], jewelry, or highly specific sports gear would be counted as incorrect according to this rule). For both *fruits* and *vegetables*, items were counted as correct if they technically met the botanical definition of fruit (i.e., develops from the flower of a plant and contains seeds) or vegetable (i.e., the edible product of a herbaceous plant), respectively. Three scores were generated for each category per participant, including total correct items, set-loss errors (i.e., an item that does not fit within the semantic category), and repetition errors.

Next, all unique correct responses from all 150 participants were compiled by category, and the values for the eleven lexical and semantic variables outlined previously were populated for each word. Category size (number of unique correct responses) for each semantic category is presented in Table 18.

Table 18.

Category size for each semantic category.

Instruments	Animals	Clothing	Furniture	Fruit	Vegetables	Eat/Drink
60	220	190	92	67	72	274

Preliminary Analyses

Several t-tests were conducted to ensure there were no relevant differences in the data collected between phases. The final sample included 150 participants (see Table 19 for demographic characteristics).

Table 19.

Demographic characteristics per phase.

	n	Gender	Age
Phase 1	100	92% F	22.05 (*6.02*)
Phase 2	50	97% F	22.99 (*7.53*)

Note. Mean age is presented with standard deviation in parentheses.

In phase one, a total of 100 participants completed the task for four semantic categories (i.e., *furniture, fruits, animals,* and *clothing*). In phase two, a total of 50 participants completed the task; 25 completed the task for four categories, including *instruments, vegetables, things you can eat or drink,* and *furniture*, while the other 25 completed the task for *instruments, vegetables, things you can eat or drink*, and *clothing.*, Twenty-five participants were randomly chosen from phase one to compare to those in phase two to analyze potential differences between phases. Total correct counts, set-loss errors, and repetition errors were compared between phases for the overlapping categories (i.e., *clothing* and *furniture*). The average total correct score for both clothing ($t(1, 48) = 1.73$, $p = .09$) and furniture ($t(1, 48) = .76$, $p = .45$) did not differ between phases. The average set-loss score for both clothing ($t(1, 48) = 0.86$, $p = 0.39$) and furniture ($t(1, 48) = 1.41$, $p = 0.17$) did not differ between phases. The average repetition error score for both clothing ($t(1, 48) = 0.31$, $p = 0.76$) and furniture ($t(1, 48) = 1.79$, $p = 0.8$) did not differ between phases. See Table 20 for descriptive statistics. The specific procedure documented here was developed, and these preliminary analyses were conducted in order to preserve data obtained during both phases while ensuring there were no relevant demographic or response differences that would contraindicate doing so. Thus, based on these analyses, the data and responses from both phases were pooled to analyze differences across all seven categories.

Table 20.

Descriptive statistics for total correct responses and errors by phase and semantic category.

	Clothing			Furniture		
	TC	SL Errors	R Errors	TC	SL Errors	R Errors
Phase 1	19.36 (3.99)	1.24 (1.67)	.16 (.47)	12.08 (3.34)	2.08 (2.33)	.12 (.33)
Phase 2	21.28 (3.85)	.88 (1.27)	.12 (.44)	11.40 (2.97)	1.32 (1.38)	.40 (.71)

Note. Mean is presented with standard deviation in parentheses. TC = total correct, SL = set-loss, R = repetition.

Data Analysis and Results

Score Analysis

Three one-way Analysis of Variance (ANOVA) tests were conducted to examine differences in total correct responses, set-loss errors, and repetition errors by semantic category (i.e., *instruments, animals, clothing, furniture, fruit, vegetables,* and *things you can eat or drink*).

Fifty participants' responses were randomly chosen from phase one to be included in these analyses.

First, a one-way Welch ANOVA was conducted to determine if the total correct responses differed by semantic category. No outliers were observed, as assessed by boxplot. The data were normally distributed for all categories, with the exception of the *vegetables* category (Shapiro-Wilk $p < .001$). Levene's test indicated the assumption of homogeneity of variances was violated ($p < .001$). Results revealed a main effect, $F(6, 151.59) = 50.84$, $p < .001$, $\omega^2 = .48$. Participants generated more correct responses in the *things you can eat or drink* category than all other semantic categories ($p < .001$) except for *animals*. More correct responses were generated in the *animals* category compared to *fruit*, *vegetables, instruments,* and *furniture* (*p*s < .001); there was no difference in correct responses between *animals* and *clothing* ($p = .66$). There were more correct responses in the *clothing* category compared to *fruit*, *vegetables*, *instruments,* and *furniture* ($p < .001$). There were no differences between the number of correct responses generated in *fruit, vegetables,* and *instruments* categories, but all three categories had more correct responses than *furniture* ($p \leq .05$). See Figure 2 for a visual representation of the data. See Table 21 for descriptive statistics.

Figure 2.

Total correct responses by semantic category.

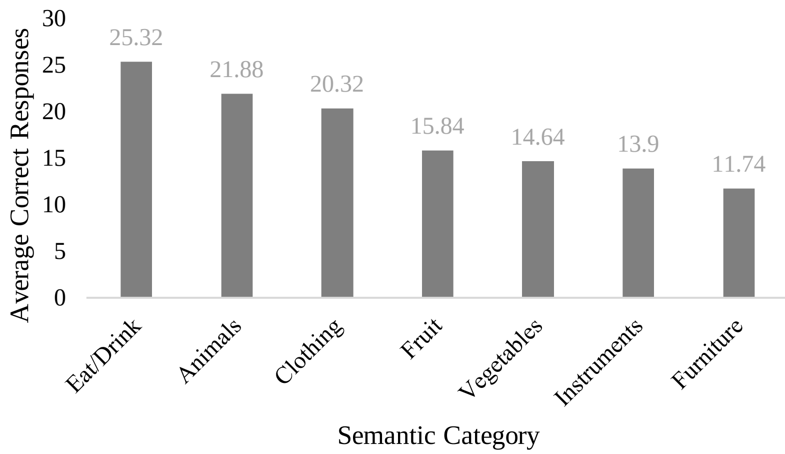


Table 21.

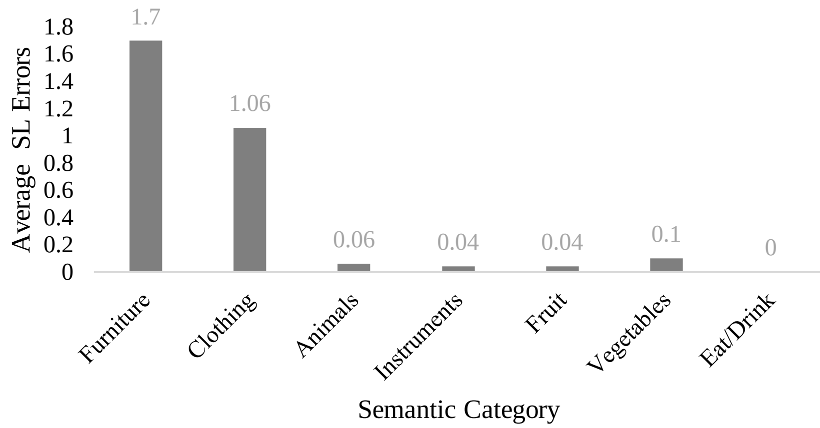
Descriptive statistics for total correct responses per semantic category.

Category	N	M	SD	Min	Max
Eat/Drink	50	25.32	7.13	7	40
Animals	50	21.88	5.45	12	35
Clothing	50	20.32	4.00	13	29
Fruit	50	15.84	3.35	9	23
Vegetables	50	14.64	4.59	7	25
Instruments	50	13.90	4.03	6	21
Furniture	50	11.74	3.15	5	20

Next, two additional one-way ANOVAs were conducted to examine differences in set-loss (i.e., incorrect) errors and repetition errors by semantic category. Results showed a main effect for set-loss errors, $F(6, 343) = 25.43$, $p < .001$, $\omega^2 = .24$. Participants made significantly more set-loss errors in the *furniture* category compared to all other categories (*p*s < .001). They also made more set-loss errors in the *clothing* category compared to the *animals, instruments, fruit, vegetables,* and *things you can eat or drink* categories (*p*s < .001). There were no other significant differences in set-loss errors. See Figure 3 for a visual representation of the data. Results demonstrated no differences in number of repetition errors across semantic categories.

Figure 3.

Set-Loss errors by semantic category.



Category Structure Analysis

Data Cleaning and Preparation. As mentioned above, all unique correct responses from all 150 participants were compiled, and the values for the eleven lexical and semantic variables outlined in Study 1 were populated for each word. Like issues arising in Study 2, many values were unavailable across variable corpora. Due to software limitations (i.e., pooling function after imputation unavailable across all accessible statistical software for the Kruskal-Wallis H test), case-wise deletion of missing values was used. Due to violations of normality and homogeneity of variance and the presence of several outliers across most semantic categories for most variables, a non-parametric Kruskal-Wallis H test was chosen to analyze differences between the lexical and semantic variables between semantic categories (Field, 2013). The Kruskal-Wallis H test was conducted for each of the 11 variables discussed throughout this dissertation. The interpretation of the individual analyses depends on the assumption of similar distributions being met, such that when this assumption is met, the group medians are compared versus mean rank comparison when it is violated (Sheskin, 2011). Descriptive statistics for all variables and semantic categories are presented following the analyses in Table 30.

Word Frequency. First, a Kruskal-Wallis H test was run to determine if there were differences in frequency between the seven semantic categories (i.e., *instruments*, *animals*, *clothing*, *furniture, fruit, vegetables*, and *things you can eat or drink*). Distributions of word frequency were relatively similar for all categories, as assessed by visual inspection of a box plot. The median frequency was different between categories, $X^2(6) = 33.31$, $p < .001$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons (i.e., the Dunn-Bonferroni post-hoc test). Post hoc analysis revealed several differences (see Table 22). See Figure 4 for a visual representation of the analysis. Median frequency was lower in the *fruit* category (6.46) compared to *animals* (7.04), *instruments* (7.28), *furniture* (7.15), and *things you can eat or drink* (7.29) categories. Median frequency was also lower in the *vegetables* category (6.47) compared to *animals*, *furniture*, and *things you can eat or drink* categories. No other differences were significant.

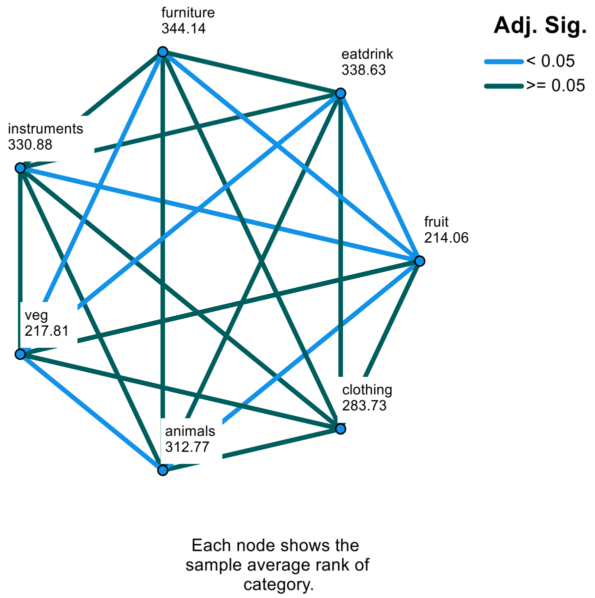
Table 22.

Significant pairwise comparisons of semantic categories for WF.

	Test Statistic	SE	Adj. Sig
Fruit-Animals	98.71	28.32	.010
Fruit-Instruments	116.82	37.30	.036
Fruit-Furniture	130.08	38.10	.013
Fruit-Eat/Drink	-124.57	28.47	< .001
Vegetables-Animals	94.97	31.05	.047
Vegetables-Furniture	126.34	40.17	.035
Vegetables-Eat/Drink	-120.83	31.19	.002

Figure 4.

All pairwise comparisons of semantic categories for WF.



Note. Blue lines indicate a significant difference between the two corresponding variables; green lines indicate a nonsignificant difference between the two corresponding variables.

Word Length. A Kruskal-Wallis H test was run to compare word length across the seven semantic categories (i.e., *instruments, animals, clothing, furniture, fruit, vegetables,* and *things you can eat or drink*). Distributions of word length did not appear similar across categories, as assessed by visual inspection of a box plot. Mean ranks of word length were determined to be different between categories, $X^2(6) = 14.85$, $p = .021$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. Post hoc analysis revealed two differences; mean ranks for word length was longer in the *fruit* category (414.92) compared to the *things you can eat or drink* (312.55, Adj. $p = .012$) and *animals* (319.58, Adj. $p = .029$) categories.

Orthographic Neighborhood Size. Another Kruskal-Wallis H test was run to compare orthographic neighborhood size across the seven semantic categories (i.e., *instruments, animals, clothing, furniture, fruit, vegetables,* and *things you can eat or drink*). Distributions of orthographic neighborhood size were not similar across categories, as assessed by visual inspection of a box plot. Mean ranks for orthographic neighborhood size were determined to be different between categories, $X^2(6) = 18.92$, $p = .004$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. Post hoc analysis revealed only one difference; orthographic neighborhood size was lower in the *fruit* category (239.49) compared to the *clothing* category (343.56), Adj. p =.009.

Phonological Neighborhood Size. A third Kruskal-Wallis H test was conducted to examine differences in phonological neighborhood size across the seven semantic categories (i.e., *instruments, animals, clothing, furniture, fruit, vegetables,* and *things you can eat or drink*). Distributions of phonological neighborhood size were not similar across categories, as assessed by visual inspection of a box plot. Mean ranks for phonological neighborhood size were determined to be different between categories, $X^2(6) = 38.41$, $p < .001$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. Post hoc analysis revealed several differences (see Table 23). See Figure 5 for a visual representation of the analysis. Mean ranks for phonological neighborhood size were lower for *fruit* (153.77) compared to *animals* (289.54), *clothing* (263.78), and *things you can eat or drink* (246.58). Phonological neighborhood size was also lower for the *instruments* category (192.28) compared to *animals.* No other differences were significant.

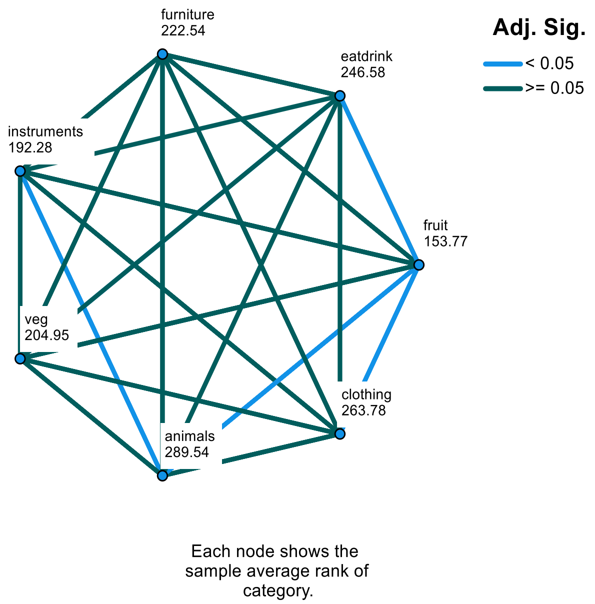
Table 23.

Significant pairwise comparisons of semantic categories for PN.

	Test Statistic	SE	Adj. Sig
Fruit-Animals	135.77	26.35	< .001
Fruit-Clothing	110.01	24.41	< .001
Fruit-Eat/Drink	-92.81	21.79	< .001
Instruments-Animals	-97.26	27.53	.009

Figure 5.

All pairwise comparisons of semantic categories for PN.



Note. Blue lines indicate a significant difference between the two corresponding variables; green lines indicate a nonsignificant difference between the two corresponding variables.

Age of Acquisition. A Kruskal-Wallis H test was run to compare age of acquisition across the seven semantic categories (i.e., *instruments, animals, clothing, furniture, fruit, vegetables*, and *things you can eat or drink*). Distributions of age of acquisition values were relatively similar across categories, as assessed by visual inspection of a box plot. The median age of acquisition was found to be different between categories, $X^2(6) = 34.23$, $p < .001$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. Post hoc analysis revealed several differences (see Table 24). See Figure 6 for a visual representation of the data. Median age of acquisition was lower for the *things you can eat or drink* category (5.61) compared to the *clothing* (7.00), *vegetables* (6.48), and *instruments* (8.06) categories. Further, the median age of acquisition was lower for the *animals* category (6.05) compared to the *instruments* category.

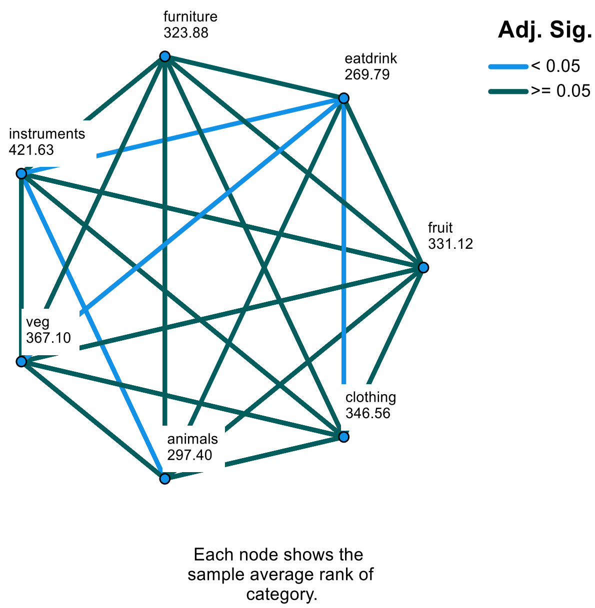
Table 24.

Significant pairwise comparisons of semantic categories for AoA.

	Test Statistic	SE	Adj. Sig
Eat/Drink-Clothing	76.77	23.66	.025
Eat/Drink-Vegetables	97.31	31.82	.047
Eat/Drink-Instruments	151.84	30.86	< .001
Animals-Instruments	124.24	30.89	.001

Figure 6.

All pairwise comparisons of semantic categories for AoA.



Note. Blue lines indicate a significant difference between the two corresponding variables; green lines indicate a nonsignificant difference between the two corresponding variables.

Emotional Valence (EV). A Kruskal-Wallis H test was run to determine if there were differences in emotional valence between the seven semantic categories (i.e., *instruments, animals, clothing, furniture, fruit, vegetables,* and *things you can eat or drink*). Distributions of emotional valence were not similar for all categories, as assessed by visual inspection of a box plot. The mean ranks for emotional valence were different between categories, $X^2(6) = 88.80$, $p <$

.001. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons was used. Post hoc analysis revealed several differences (see Table 25). See Figure 7 for a visual representation of the analysis. Mean ranks for emotional valence were higher for the *things you can eat or drink* category (347.19) compared to *clothing* (196.14), *furniture* (220.13), *animals* (236.69), *instruments* (231.74), and *vegetables* (238.05). Further, mean ranks for emotional valence were higher for *fruit* (384.72) compared to *clothing, furniture, animals, instruments,* and *vegetables.*

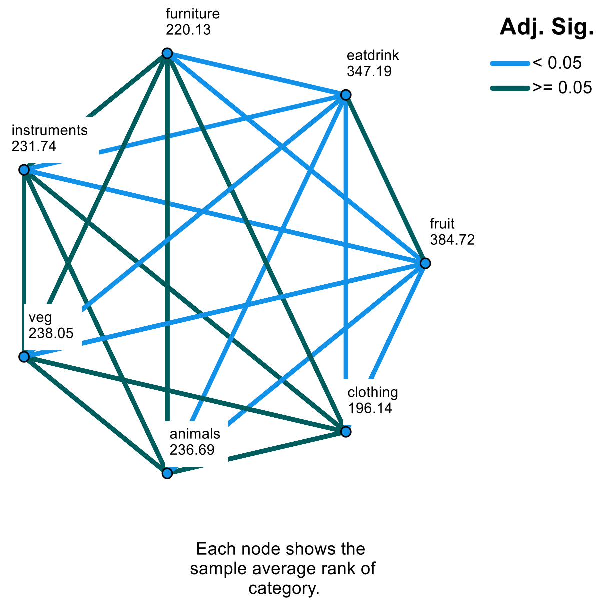
Table 25.

Significant pairwise comparisons of semantic categories for EV.

	Test Statistic	SE	Adj. Sig
Clothing-Eat/Drink	-151.05	22.59	< .001
Furniture-Eat/Drink	-127.07	30.55	.001
Instruments-Eat/Drink	-115.45	28.75	.001
Animals-Eat/Drink	-110.50	17.11	< .001
Vegetables-Eat/Drink	-109.14	31.40	.011
Clothing-Fruit	-188.57	32.38	< .001
Furniture-Fruit	-164.59	38.36	< .001
Animals-Fruit	-148.02	28.83	< .001
Instruments-Fruit	-152.97	36.95	.001
Vegetables-Fruit	146.67	39.04	.004

Figure 7.

All pairwise comparisons of semantic categories for EV.



Note. Blue lines indicate a significant difference between the two corresponding variables; green lines indicate a nonsignificant difference between the two corresponding variables.

Emotional Arousal (EA). A Kruskal-Wallis H test was run to determine if there were differences in emotional arousal between the seven semantic categories (i.e., *instruments, animals, clothing, furniture, fruit, vegetables,* and *things you can eat or drink*). The pattern of distributions of emotional arousal was relatively similar for all categories, as assessed by visual inspection of a box plot. The median emotional arousal was different between categories, $X^2(6) = 50.11$, $p < .001$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons was used. Post hoc analysis revealed several differences (see Table 26). See Figure 8 for a visual representation of the analysis. Median emotional arousal was higher for the *animals* category (4.20) compared to *vegetables* (3.31), *furniture* (3.49), *clothing* (3.50), and *things you can eat or drink* (3.91). The median emotional arousal was also higher for the *things you can eat or drink* category compared to the *vegetables* category.

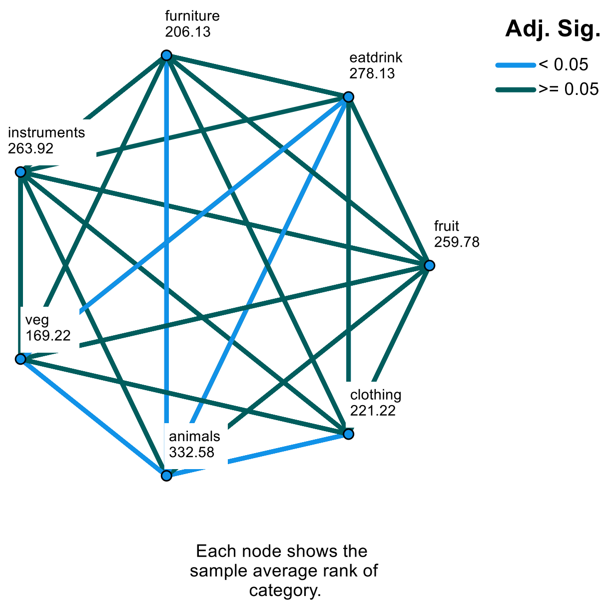
Table 26.

Significant pairwise comparisons of semantic categories for EA.

	Test Statistic	SE	Adj. Sig
Vegetables-Eat/Drink	-108.92	31.40	.011
Vegetables-Animals	163.36	34.45	< .001
Furniture-Animals	126.45	30.62	.001
Clothing-Animals	111.35	22.68	< .001
Eat/Drink-Animals	54.45	17.11	.031

Figure 8.

All pairwise comparisons of semantic categories for EA.



Note. Blue lines indicate a significant difference between the two corresponding variables; green lines indicate a nonsignificant difference between the two corresponding variables.

Imageability. A Kruskal-Wallis H test was run to determine if there were differences in imageability between the seven semantic categories (i.e., *instruments, animals, clothing, furniture, fruit, vegetables,* and *things you can eat or drink*). The pattern of distributions for imageability was relatively similar for all categories, as assessed by visual inspection of a box plot. The median imageability was different between categories, $X^2(6) = 32.67$, $p < .001$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons was used. Post hoc analysis revealed several differences (see Table 27). See Figure 9 for a visual representation of the analysis. Median imageability was higher for the *animals* category (6.69) compared to *furniture* (6.50), *instruments* (6.53), and *clothing* (6.57). *Fruit* (6.71) also had a higher median imageability rating compared to *furniture* and *instruments.*

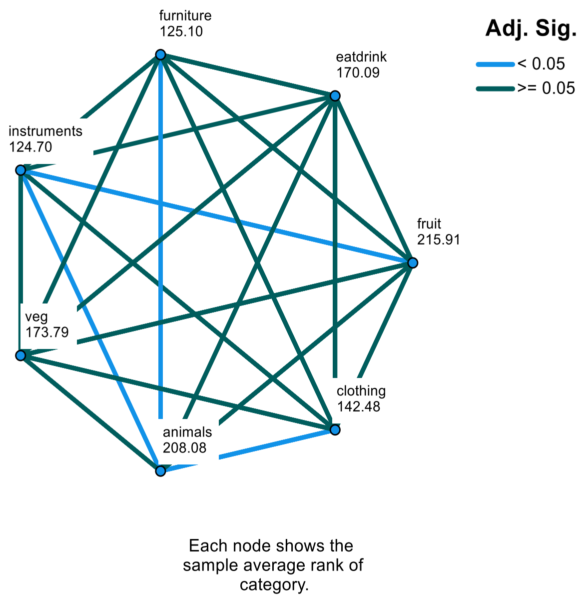
Table 27.

Significant pairwise comparisons of semantic categories for imageability.

	Test Statistic	SE	Adj. Sig
Instruments-Fruit	-91.21	29.55	.043
Instruments-Animals	-83.38	21.69	.003
Furniture-Fruit	-90.81	30.08	.053
Furniture-Animals	82.98	22.41	.370
Clothing-Animals	65.60	20.49	.029

Figure 9.

All pairwise comparisons of semantic categories for imageability.



Note. Blue lines indicate a significant difference between the two corresponding variables; green lines indicate a nonsignificant difference between the two corresponding variables.

Semantic Diversity. A Kruskal-Wallis H test was run to determine if there were differences in semantic diversity between the seven semantic categories (i.e., *instruments, animals, clothing, furniture, fruit, vegetables,* and *things you can eat or drink*). The pattern of distributions for semantic diversity was not similar for all categories, as assessed by visual inspection of a box plot. The mean ranks for semantic diversity were different between categories, $X^2(6) = 20.65$, p = 002. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. Post hoc analysis revealed two differences; the responses in the *vegetables* category (159.26) had lower semantic diversity compared to the *animals* (247.97, Adj. p = .041) and *furniture* (276.26, Adj. p =.014) categories.

Semantic Neighborhood Density (SND). A Kruskal-Wallis H test was run to determine if there were differences in SND between the seven semantic categories (i.e., *instruments,* *animals, clothing, furniture, fruit, vegetables,* and *things you can eat or drink*). The pattern of distributions for SND was not similar for all categories, as assessed by visual inspection of a box plot. The mean ranks for SND were different between categories, $X^2(6) = 92.12$, $p < .001$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons was used. Post hoc analysis revealed several differences (see Table 28). See Figure 10 for a visual representation of the analysis. Mean rank for SND was lower for *furniture* (134.73) compared to all other categories. Mean rank for SND was higher for *vegetables* (434.77) compared to *animals* (249.07), *clothing* (305.59), and *instruments* (311.05) and higher for *things you can eat or drink* (353.96) compared to *animals* (349.07).

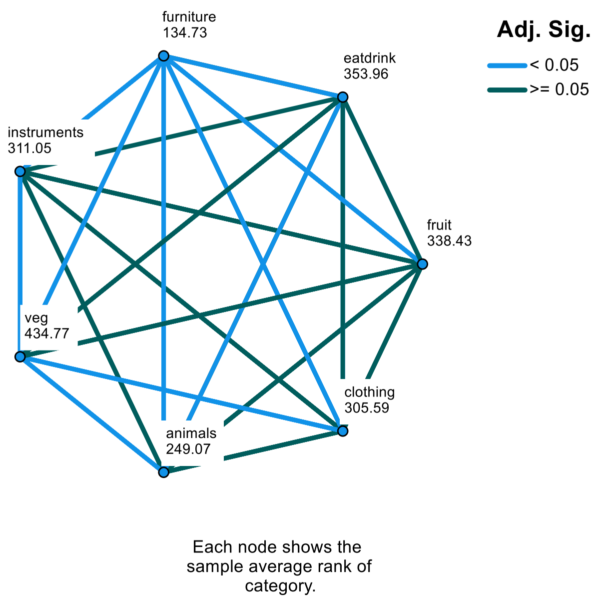
Table 28.

Significant pairwise comparisons of semantic categories for SND.

	Test Statistic	SE	Adj. Sig
Furniture-Animals	114.35	30.50	.004
Furniture-Clothing	170.86	33.66	< .001
Furniture-Instruments	176.33	39.27	< .001
Furniture-Fruit	-203.70	39.02	< .001
Furniture-Eat/Drink	-219.24	30.39	< .001
Furniture-Vegetables	-300.05	40.39	< .001
Animals-Eat/Drink	-104.89	18.20	< .001
Animals-Vegetables	-185.70	32.24	< .001
Clothing-Vegetables	-129.19	35.23	.005
Instruments-Vegetables	-123.72	40.63	.049

Figure 10.

All pairwise comparisons of semantic categories for SND.



Note. Blue lines indicate a significant difference between the two corresponding variables; green lines indicate a nonsignificant difference between the two corresponding variables.

Sensorimotor Information (SI). A Kruskal-Wallis H test was run to determine if there were differences in amount of sensorimotor information between the seven semantic categories (i.e., *instruments, animals, clothing, furniture, fruit, vegetables,* and *things you can eat or drink*). The pattern of distributions for sensorimotor information was not similar for all categories, as assessed by visual inspection of a box plot. The mean ranks for sensorimotor information were different between categories, $X^2(6) = 391.76$, $p < .001$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons was used. Post hoc analysis revealed several differences (see Table 29). See Figure 11 for a visual representation of the analysis. The mean rank for sensorimotor information was lower for *animals* (163.68) compared to all categories except *furniture*, lower for *furniture* (226.47) compared to *vegetables* (412.53), *fruit* (428.08), and *things you can eat or drink* (456.28), lower for *clothing* (259.78) compared to *vegetables*, *fruit*, and *things you can eat or drink,* and lower for *instruments* (334.78) compared to *things you can eat or drink*.

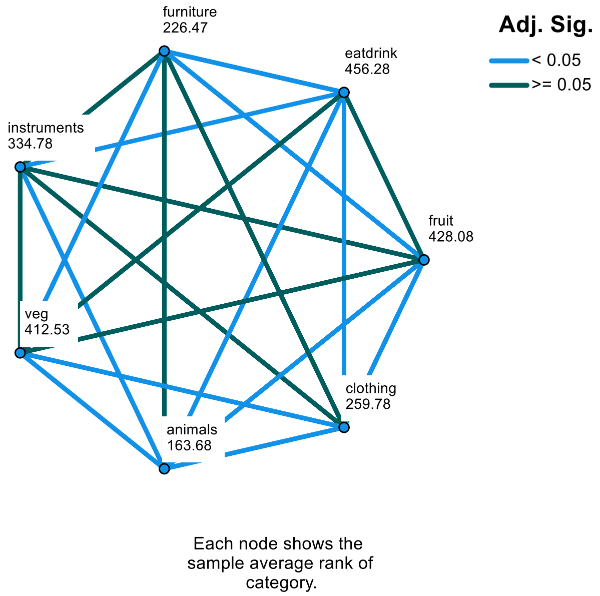
Table 29.

Significant pairwise comparisons of semantic categories for SI.

	Test Statistic	SE	Adj. Sig
Animals-Clothing	-96.10	23.49	.001
Animals-Vegetables	-248.85	32.56	< .001
Animals-Fruit	-264.40	30.05	< .001
Animals-Eat/Drink	-292.60	18.73	< .001
Animals-Instruments	171.10	31.52	< .001
Furniture-Vegetables	-186.06	40.34	< .001
Furniture-Fruit	-201.61	38.34	< .001
Furniture-Eat/Drink	-229.81	30.30	< .001
Clothing-Vegetables	-152.75	35.50	< .001
Clothing-Fruit	-168.30	33.21	< .001
Clothing-Eat/Drink	-196.49	23.47	< .001
Instruments-Eat/Drink	-121.50	31.51	.002

Figure 11.

All pairwise comparisons of semantic categories for SI.



Note. Blue lines indicate a significant difference between the two corresponding variables; green lines indicate a nonsignificant difference between the two corresponding variables.

Table 30.

Descriptive statistics for all lexical and semantic variables by category.

		WF	WL	ON	PN	AoA	EV	EA	Img	SemD	SND	SI
Instruments	N	41	44	41	41	42	37	37	27	28	39	40
	Med.	7.28	6.50	1.00	6.00	8.06	5.70	3.79	6.53	1.23	.626	5.81
Animals	N	185	192	183	55	183	170	170	135	133	180	186
	Med.	7.04	6.00	5.00	6.00	6.05	5.74	4.20	6.69	1.48	.564	4.77
Clothing	N	82	93	79	79	85	69	69	31	60	82	87
	Med.	6.76	6.00	2.00	6.00	7.00	5.58	3.5	6.57	1.42	.593	5.28
Furniture	N	38	46	37	37	45	32	32	25	35	40	44
	Med.	7.15	6.50	2.00	2.00	6.48	5.64	3.49	6.50	1.48	.472	4.98
Fruit	N	49	51	48	48	46	37	37	22	29	40	45
	Med.	6.46	7.00	6.46	.00	6.56	6.47	3.83	6.71	1.38	.633	6.20
Vegetables	N	39	41	38	38	39	30	30	17	25	35	37
	Med.	6.47	6.00	.00	1.00	6.79	5.75	3.31	6.63	1.30	.708	6.10
Eat/Drink	N	176	196	171	171	185	175	175	99	145	188	187
	Med.	7.29	6.00	1.00	3.00	5.61	6.32	3.91	6.62	1.42	.643	6.27
Total	N	610	663	597	469	625	550	550	356	455	604	626
	Med.	7.07	6.00	1.00	3.00	6.25	5.90	3.88	6.64	1.43	.604	5.53

STUDY 3 DISCUSSION

As described, these seven categories were chosen based on their popularity of use. These categories are used in common neuropsychological tests, including the D-KEFS (Delis et al., 2001), the NEPSY-II (Korkman et al., 2007), the Woodcock-Johnson Test of Achievement – Fourth Edition (WJ-ACH-IV; Schrank & Wendling, 2018) and the letter and category fluency test on the Halstead-Reitan Battery (Reitan & Wolfson, 1993). Further, two of these categories (i.e., *furniture* and *fruit*) are used together in a popular category switching task on the D-KEFS (Delis et al., 2001). In theory, participants should be able to generate an equal number of items from each category, especially on the category switching task; however, as shown in Study 3, this is not the case.

There are several notable findings in Study 3. First, participants' responses confirmed the anecdotal difference observed between these many of these categories. Across all categories, the most correct responses were in the *things you can eat or drink* category, followed by the *animals* and *clothing* categories, and the fewest correct responses were in the *furniture* category. The other three categories (i.e., *fruit, vegetables*, and *instruments*) were equal in terms of number of correct responses. Further, participants made more set loss errors in the *furniture* category compared to all others, followed by the *clothing* category. The other five categories were equal in the number of set loss errors. As expected, there were no differences in repetition errors across categories. The qualitative differences in category size (number of unique correct responses across all participants), as shown in Table 18, also demonstrate the inequity across semantic categories. To highlight the relevance of this inequity, as discussed previously, a study conducted by Diaz et al. (2003) found that patients with Alzheimer's disease showed more severe deficits for larger categories as compared to smaller ones. Therefore, the severity of the observed deficit is dependent upon which category is used and would in turn affect interpretation.

These findings demonstrated the need to further investigate differences in the underlying lexical and semantic structure of the categories themselves. Results of the category structure analysis show that there was at least one difference in all of the 11 psycholinguistic variables between categories. There are two reasons the large number of differences is important to understand. First, researchers and clinicians should be careful about how they interpret the results of semantic fluency tasks; using raw scores of either single different categories or pooled raw scores across different categories to compare across studies or patient populations may be minimally confusing or maximally may cause misleading interpretations. These quantitative and qualitative differences in responses produced by healthy participants can add to the research on patient populations. For example, three studies (Forbes-McKay et al., 2005; Sailor et al., 2011; Venneri et al., 2008) all found that patients with Alzheimer's disease produced words with earlier age of acquisition compared to healthy controls. All three studies used an aggregated raw score as the dependent variable (i.e., *animals* + *fruit* [Forbes-McKay et al., 2005; Venneri et al., 2008], *animals* + *fruit* + *vegetables* [Sailor et al., 2011]). Study 3 found that items produced by healthy controls do not differ in age of acquisition between any of these three categories, suggesting they are appropriate categories to measure the effects of that specific variable in patient populations and would lead to informative and accurate results.

In contrast, Forbes-McKay and colleagues (2005) found that items produced by patients with Alzheimer's disease had higher word frequency compared to healthy controls, whereas Henderson et al. (2023) found no difference for this same comparison. Henderson et al. (2023) used the raw score of a single category as the dependent variable (i.e., *animals*). The difference in results could be explained by the methodology. Study 3 showed an underlying difference in word frequency between items in the *fruit* and *animals* categories, where items in the *fruit* category have lower word frequency compared to those in the *animals* category. Therefore, the difference found by Forbes-McKay et al. (2005) could have been driven by the inclusion of the *fruit* category; however, because the raw scores are pooled or aggregated across categories, this difference would go undetected. Taken together with findings from recognition memory research, whereby the low frequency advantage was found to be eliminated in individuals with Alzheimer's disease, categories with lower frequency items may be more sensitive to picking up a difference between patients and healthy controls.

Similarly, there were differences in the literature regarding the impact of psycholinguistic variables on responses generated by patients with schizophrenia. Baskak and colleagues (2008) found that items produced by patients with schizophrenia were more "peculiar" compared to those produced by healthy controls on a phonemic fluency task (i.e., letter *s*). Juhasz et al. (2012) used the *animals* category and found that responses by patients had earlier age of acquisition and higher typicality ratings compared to healthy controls, whereas Kiang and Kutas (2006) used the *fruit* category and found responses were lower in typicality relative to healthy individuals. Although the current study did not investigate "typicality" or "peculiarity," the structure of the two categories used by Juhasz et al. (2012) and Kiang and Kutas (2006) differ on several variables, including word frequency, word length, phonological neighborhood size, emotional valence, and sensorimotor information. It is unclear which of these variables or which combination of these variables could be driving a difference in their results; however, word frequency may be a likely candidate (i.e., more typical items or less peculiar items may have higher word frequency). The results of Study 3 suggest that inconsistencies in the literature could be due to differences in the underlying structure of the categories used. Further, some categories may be more sensitive than others to changes in certain variables across patient populations.

Second, these results support the idea that the two categories used in the most common category switching task are not equal and, therefore, should likely not be included together. Items from the *fruit* category and items from the *furniture* category were different on five out of the 11 variables examined, in addition to correct counts and set-loss errors. Participants generated more correct responses and fewer set-loss errors in the *fruit* category compared to the *furniture* category. Further, on average, items from the *fruit* category have lower word frequency but higher emotional valence and imageability ratings, denser semantic neighborhoods, and more associated sensorimotor information compared to those in the *furniture* category. Examinees are likely better able to produce items from the *fruit* category than they are from the *furniture* category because of the constraints associated with the smaller set size for *furniture* on most of the variables. One goal of the current study was to present other potential ideas for categories to be paired together into a category switching task as guided by the results of Study 3. In looking at the categories analyzed in this study, the *things you can eat or drink* and *animals* categories were excluded as potential options for several reasons: 1) these two categories had very large category sizes (i.e., 220 for *animals* and 274 for *things you can eat or drink*) relative to the others, 2) the *things you can eat or drink* category has items that also include items in two other potential options (i.e., *fruit* and *vegetables*), and 3) the *animals* category is very commonly used as the standard category in a standalone semantic fluency task, and it is unlikely that clinicians and researchers would be willing to give this category up for inclusion in a switching task. The *furniture* category was also excluded as a potential option, as participants generated fewer correct responses and more set-loss errors in this category compared to all others. Thus, *clothing,* *instruments, fruit,* and *vegetables* were left as options. The goal was to find four categories (or two pairings) that move closer toward an assumption of equality to be used in a category switching task; thus, only pairings that had fewer differences between them compared to the original task (i.e., five psycholinguistic variables + correct counts + set-loss errors) were considered as viable options. The current project attempted to identify two sets of pairings as repeat assessments typically use a “Form B” alternative version of the same task. These four categories were paired together, creating four pairs (i.e., *fruit/clothing, fruit/instruments, vegetables*/*clothing*, *vegetables*/*instruments*) to be compared to the original task (i.e., *fruit/furniture*).

Potential Pairings

Fruit/Clothing

Based on the results from Study 3, the *fruit* and *clothing* categories were different on four out of the 11 variables analyzed; on average, items from the *clothing* category had higher word frequency, denser orthographic and phonological neighborhoods, and more associated sensorimotor information, but lower ratings of emotional valence compared to the *fruit* category. Although participants generated approximately the same number of correct responses in both categories, the *clothing* category was associated with more set-loss errors.

Fruit/Instruments

Study 3 results indicated that the *fruit* and *instruments* categories differ on three of the 11 psycholinguistic variables analyzed. On average, the *fruit* category had lower word frequency but higher ratings of emotional valence and imageability compared to the *instruments* category. These categories were deemed equivalent in terms of number of correct responses and set-loss errors.

Vegetables/Clothing

Results from Study 3 show that the *vegetables* and *clothing* categories were different on two out of the 11 variables analyzed; on average, items from the *vegetables* category had denser semantic neighborhoods and more associated sensorimotor information compared to the *clothing* category. Notably, participants were able to generate more correct responses in the *clothing* category but also had more set-loss errors compared to the *vegetables* category.

Vegetables/Instruments

Lastly, Study 3 demonstrated that the *vegetables* and *instruments* categories were different on only one variable; on average, items from the *vegetables* category had denser semantic neighborhoods compared to items in the *instruments* category. Number of correct responses and set-loss errors were deemed to be equal for these two categories.

Overall, all four pairings have <u>fewer</u> underlying differences compared to the pairing in the original switching task. The *vegetables/instruments* pair is closest to equal with only one underlying difference and no differences in correct counts or errors, while the *fruit/clothing* pair are most different with four underlying differences and differences in set-loss errors. Study 4 directly compared switching accuracy across these new four new pairings and the original *fruit/furniture* pairing to guide future test development.

STUDY 4

Objective

Based on Study 3, there are clear differences in the ease with which their prototypical responses can be generated in many of the categories that are expected to be equivalent. Thus, the objective for Study 4 was to present potential options for a modified category switching task with the intent to move closer to that assumption of equality. The objective of Study 4 was to directly compare performance (i.e., switching accuracy) on the original category switching task (*fruit*/*furniture*) to a modified version that took the results of Study 3 into account. Errors were not analyzed as they are built into the overall switching accuracy score.

Participants

All participants in Study 4 were undergraduate students at the University of Windsor who volunteered to participate in return for bonus credits toward their eligible psychology courses. All participants were required to report English as their first language. Fifty participants were recruited, and all participants' data was included in the analysis (M_{age} = 21.5, 90% female).

Procedure

Before beginning data collection, Study 4 was approved by the University of Windsor's REB. Similar to Study 3, participants completed the study via Microsoft Teams with a live research assistant. Participants signed a consent form via Qualtrics survey and, upon joining the meeting, answered inclusion and demographic questions. Participants were given 60 seconds to complete a single category switching task. Participants were randomly assigned to one of five experimental conditions (i.e., *furniture*/*fruit* condition [original version] vs. *fruit/clothing* [modified version] vs. *vegetables/clothing* [modified version] vs. *fruit/instruments* [modified version] vs. *vegetables/instruments* [modified version]). Before beginning, participants were instructed: "When I say begin, I want you to switch back and forth between saying as many (INSERT CATEGORY 1) and as many (INSERT CATEGORY 2) as you can. You will have 60 seconds before I tell you to stop. So you would say a (INSERT CATEGORY 1), then a (INSERT CATEGORY 2), then a (INSERT CATEGORY 1), then a (INSERT CATEGORY 2), and so on. You can start with either a (INSERT CATEGORY 1) or a (INSERT CATEGORY 2). Do you have any questions? Ready? Begin."

Participants' responses were scored according to the same criteria as described in Study 3. For category switching accuracy, to be considered an accurate switch, both items have to correct to be considered accurate. Meaning that a switch from an incorrect item due to set-loss or repetition to a correct item of the other category does not count as a correct switch.

Data Analysis and Results

A one-way ANOVA was conducted to examine differences in switching accuracy by condition (*furniture*/*fruit* condition [original version] vs. *fruit/clothing* [modified version] vs. *vegetables/clothing* [modified version] vs. *fruit/instruments* [modified version] vs. *vegetables/instruments* [modified version]). No outliers were observed on assessment of a boxplot. The data were normally distributed for all conditions (Shapiro-Wilk $p > .05$). Levene's test indicated the assumption of homogeneity of variances was met ($p > .05$). Results revealed a main effect, $F(4, 45) = 45.32$, $p < .001$, $\omega^2 = .38$. Tukey's post hoc analysis showed that participants had fewer correct switches in the original version (i.e., *furniture/fruit;* $M = 12.00$, $SD = 2.79$) compared to all modified versions (i.e., *fruit/clothing* [$M = 15.60$, $SD = 2.22$, $p < .01$], *fruit/instruments* [$M = 16.20$, $SD = 1.75$, $p < .01$], *vegetables/clothing* [$M = 15.80$, $SD = 1.87$, $p < .05$], *vegetables/instruments* [$M = 17.80$, $SD = 2.03$, $p < .001$]). There were no differences in the number of correct switches between any of the modified versions.

Given the lack of group differences between all modified versions of the task, the two one-sided tests (TOST) procedure was conducted for all modified versions to test for equivalence. All of the following equivalence tests were conducted using the TOSTER package in R (Caldwell, 2022; R Core Team, 2021). To determine the boundaries for the smallest effect size of interest (SESOI), the normative data for the original category switching task provided by the DKEFS (Delis et al., 2001) was consulted. Based on those norms, every correct switch after the initial six corresponded to an increase in one scaled score point. Given that switching accuracy across categories ranged from ten to 22 (all above the zero to six threshold), a mean difference of a single point was determined to be a meaningful change, and therefore, the equivalence bound (SESOI) for all of the following analyses were set to a raw count of ±1.

Fruit/Clothing vs. Fruit/Instruments

A TOST procedure was conducted to examine whether switching accuracy on the *fruit/clothing* pairing was equivalent to those in the *fruit/instrument* pairing. Results yielded a nonsignificant result for the test against Δ_{U}, $t(17.1) = .45$, $p = .33$, despite significance for Δ_{L}, $t(17.1) = -1.79$, $p < .05$. Thus, results suggest these two versions of the task are <u>not</u> statistically equivalent.

Fruit/Clothing vs. Vegetables/Instruments

Another TOST procedure was run to examine equivalence between the *fruit/clothing* pairing and the *vegetables/instruments* pairing. Results yielded a nonsignificant result for the test against Δ_{U}, $t(17.6) = -1.12$, $p = .86$ (Δ_{L}, $t(17.6) = -2.98$, $p < .01$). Results suggest these two versions of the task are <u>not</u> statistically equivalent.

Fruit/Clothing vs. Vegetables/Clothing

The TOST procedure between *fruit/clothing* and *vegetables/clothing* yielded a nonsignificant result for both Δ_{U}, $t(14) = 1.00$, $p = .17$, and Δ_{L}, $t(14) = -1.48$, $p = .08$, indicating non-equivalence.

Fruit/Instruments vs. Vegetables/Clothing

The TOST procedure between *fruit/instruments* and *vegetables/clothing* yielded a nonsignificant result for Δ_{L}, $t(16.1) = -.89$, $p = .19$ (Δ_{U}, $t(16.1) = 2.07$, $p < .05$), indicating non-equivalence.

Fruit/Instruments vs. Vegetables/Instruments

The TOST procedure between *fruit/instruments* and *vegetables/instruments* yielded a nonsignificant result for Δ_U, $t(15.9) = -.61$, $p = .73$ (Δ_L, $t(15.9) = -2.64$, $p < .01$), indicating non-equivalence.

Vegetables/Instruments vs. Vegetables/Clothing

The TOST procedure between *vegetables/instruments* and *vegetables/clothing* yielded a nonsignificant result for Δ_U, $t(12.9) = -1.11$, $p = .86$ (Δ_L, $t(12.9) = -3.33$, $p < .01$), indicating non-equivalence.

STUDY 4 DISCUSSION

The goal of Study 4 was to evaluate potential new pairings of semantic categories to be included in a modified category switching task in attempt to create a task where both categories have more similar underlying lexical and semantic structures compared to the current common version. Four new pairs were identified based on the results from Study 3. In Study 4, switching accuracy was directly compared against four new pairs (i.e., *fruit/clothing, fruit/instruments, vegetables*/*clothing*, *vegetables*/*instruments*) and the original pair (i.e., *fruit/furniture*). In healthy participants, mean switching accuracy was lower in the original pairing compared to all of the new pairings. These results suggest that all four of the presented options may be superior to the current category switching task; however, follow-up equivalence testing determined that none of the proposed pairings were statistically equivalent, suggesting subtle differences between them.

Returning to Study 3, results demonstrate that the new pairs are all more similar in their inherent lexical and semantic structure compared to the original pair. Specifically, the pairings follow a continuum from most similar to least similar: *vegetables/instruments,* followed by *vegetables/clothing,* followed by *fruit/instruments,* followed by *fruit/clothing,* followed by *fruit/furniture*. The pairs differ on different variables, so it is difficult to determine which variable is most influential; it appears to be the combination that leads to less productive semantic fluency. It is possible that the semantic characteristics of a category play a larger role than the lexical characteristics, given the constraints of the task. All four differences in the original pair (i.e., *fruit/furniture*) are on semantic variables (i.e., emotional valence, imageability, semantic neighborhood density, and sensorimotor information), whereas the next most dissimilar pair (i.e., *fruit/clothing*) have differences split between lexical and semantic variables (i.e., orthographic neighborhood, phonological neighborhood, emotional valence, and sensorimotor information). Thus, the difference in switching accuracy could arise from meeting some specific threshold of semantic difference.

Given the lack of differences in mean switching accuracy across the new pairings obtained through null-hypothesis testing, equivalence testing was conducted as follow-up to examine whether there are small, albeit meaningful, effects between the pairs. The results suggest that all modified pairings yield differences in switching accuracy that would meaningfully affect interpretation, despite the effects not appearing in hypothesis testing. However, Lakens et al. (2018) discuss that when observed effects are not statistically equivalent and not statistically different from zero, there is not enough data to draw definitive conclusions. Therefore, future research should be conducted with a much larger sample size to better quantify the SESOI and understand the differences between these pairings.

CHAPTER 4:

GENERAL DISCUSSION

The current project was motivated by the relative lack of literature integrating psycholinguistic experimental findings and clinical neuropsychological research. It has been well documented in psycholinguistic research that word-level characteristics impact language processing and memory. Despite this longstanding and productive field, neuropsychological research has only recently begun to investigate how patient populations may differ in their word-level processing. There is a significant gap in the literature regarding understanding the lexical and semantic structure of neuropsychological tests using healthy participants. Results from Studies 1 and 2 demonstrated that 1) multiple psycholinguistic variables predicted better recall of items from neuropsychological tests and 2) even when evaluated generously, only one of the ten test lists examined would be considered adequately controlled in terms of consistency in said variables. Results from Study 3 demonstrated a myriad of ways in which several common categories used in semantic fluency tasks are unequal; the categories have differences in their overall number of unique responses produced, total correct counts, set-loss errors, and across several lexical and semantic variables. Lastly, Study 4 presented and evaluated four potential remedial options for a category switching task that move closer toward structure equality. Overall, the results of this dissertation show that the neuropsychological tests examined herein are poorly constructed from a psycholinguistic perspective. Although scores are typically normed against standardized data when used for neuropsychological purposes, these results are still highly relevant. As discussed, different patient populations have shown specific deficits in processing lexical and semantic information or producing semantic fluency items that have qualitative differences in lexical and semantic characteristics. The information from the current study could be useful in continuing to investigate more comprehensive ways to discriminate conditions. Moreover, as discussed at length in Study 3, many research studies appear to use raw scores (correct counts) of a single category or aggregate category scores into a single pooled raw score as the outcome variable. These results suggest that this may be causing inconsistencies in the literature and that relevant and informative differences in responses between semantic categories could be averaged out after norming.

Further, the results of Study 3 should also influence list learning test construction when test developers wish to include semantic categories/clustering within their test lists. For example, some list learning tests (e.g., CVLT-3) purposely include test items that fit into specific semantic clusters (e.g., vegetables, animals, furniture, and transportation). The results of Study 3 clearly indicate differences in the underlying structure of three of these four categories (transportation was not studied herein). Thus, not only do the lexical and semantic characteristics of the individual words within tests differ, those of the categories within them do as well.

Limitations & Future Directions

Although the current study begins to fill a gap within the literature, there are limitations to the project and analyses. First, a large limitation across all four studies is related to the corpora used for populating psycholinguistic variables and data. Some corpora were much smaller than others. For example, the database used for imageability ratings had 5,500 total words (Scott et al., 2019), whereas the database for word frequency had over 40,000 total words (Lund & Burgess, 1996). Some of these variables were unavailable for many words, especially for those in Study 3, leading to missing data. For Study 2, multiple imputation was used as all items in the study were required to have ratings. Multiple imputation likely reduced any bias introduced by the missingness (Madley-Dowd et al., 2019); however, casewise deletion was used in Study 3 due to software limitations. Thus, the results for some variables (i.e., imageability, phonological neighborhood size) in Study 3 should be interpreted with caution due to the large number of excluded data. Further, several items in categories evaluated in the current project had two constituents (e.g., “ice” “cream”); these items could not be evaluated, as they were unavailable across all databases.

Another limitation of the current project relates to the availability of neuropsychological test lists. Only seven neuropsychological tests (ten total test lists) of list-learning and memory were evaluated. Further, some of the tests included in Studies 1 and 2 had semantic clustering (e.g., CVLT-C, CVLT-3), whereas others did not (WRAML-2). Because test lists were combined into 42 to 47 items, the effects of these clusters were likely minimal; however, due to the relatively few tests included the interaction between semantic clustering and the effects of these variables could not be investigated. Future research should attempt to quantify how lists with semantic clustering differ from those without in terms of their psycholinguistic structure.

A third limitation of the current study is related to the sample characteristics; all studies were dominated by females. Research has consistently shown that females outperform males in the immediate and delayed recall trials of verbal learning tasks and use more efficient strategies (e.g., semantic clustering; Bolla-Wilson & Bleecker, 1986; Kramer et al., 1997; Kramer et al., 1988). Further, there is evidence that women also outperform men in some aspects of verbal fluency. Early research found sex differences for phonemic fluency only (Bolla et al., 1990; Capitani et al., 1998; Crossley et al., 1997; Weiss et al., 2003; Weiss et al., 2006); however, some studies have shown superior performance on semantic and other fluency tasks, as well (Costa et al., 2004). Like the results of sex differences in verbal learning, there is evidence that women also use more effective strategies (e.g., more cluster switches) for verbal fluency (Weiss et al., 2006), although one study (Sokolowski et al., 2020) found an opposite pattern of results, with men slightly outperforming women in total correct responses, as well as cluster size and cluster switches. Thus, given the large number of females in the current studies sample and the sex differences described above, the results may have looked different had there been more heterogeneity in the sample. The notion and consideration of sex differences is particularly notable given that neuropsychological norms are only occasionally stratified by sex. Future research should examine the influence of sex on the psycholinguistic effects in verbal learning and memory and verbal fluency found here.

In attempt to include as many neuropsychological tests as possible and to ensure an appropriate number of items, all accessible list learning tests were used to create the stimulus sets and analyzed in the current study, including those developed specifically for children. The current study used an undergraduate population, and therefore, these participants are considered out of age-range for some of these tests lending to another potential limitation. Although the results of Study 2 are not necessarily impacted because raw scores were used rather than normed scores, it will be important to follow-up with a pediatric population and to investigate pediatric-based disorders. It will be especially important given that age of acquisition was one variable that was found to be influential; tests developed specifically for pediatric populations should ensure only words with appropriate age of acquisition are included and should ensure a small range on this variable to limit the influence of any effects.

Another area of focus for future research is to examine recognition memory. Many neuropsychological test lists have a recognition and/or forced choice component. Given the significant differences in effects of psycholinguistic variables between recognition and recall tasks, it would be interesting and informative to evaluate the items in those components. Similarly, the current study did not analyze the characteristics of intrusion, set-loss, or repetition errors across studies. Future studies should investigate whether lexical or semantic characteristics play a role in errors produced in memory or semantic fluency tasks.

Conclusions

In conclusion, this dissertation described the effects of lexical and semantic characteristics on memory and semantic fluency for common neuropsychological tests. Taken together and with consideration of these results in mind, there are two ways to move forward. If future research shows that these results carry over into pertinent populations, these psycholinguistic effects could potentially be useful to inform clinical diagnostics. Process scores could be developed to analyze patient responses on semantic fluency tasks and patterns of correct items on list learning tasks to better differentiate conditions and to better understand patients' experiences. Additionally, test developers should be more stringent in the requirements for choosing test items and categories; specifically, tests should be reconstructed to minimize these effects by controlling for influential psycholinguistic information to make each item equal to the next on list learning tasks and to ensure the categories included in semantic fluency tasks and importantly, included in category switching tasks, are equal to each other. To be clear, both suggestions made here should be implemented. Test lists and categories that have purposeful manipulations in relevant characteristics should be used to garner more detailed clinical information, AND test lists and categories that are well controlled should be used to more "purely" test memory and fluency (i.e., without the confound of lexical and semantic information).

REFERENCES

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review, 122*(3), 558–569. https://doi.org/10.1037/a0038693

Allen, H.A., Liddle, P. F., & Frith, C. D. (1993). Negative features, retrieval processes and verbal fluency in schizophrenia. *The British Journal of Psychiatry, 163*(6), 769–775. https://doi.org/10.1192/bjp.163.6.769

Almond, N. M., & Morrison, C. M. (2014). Episodic intertrial learning of younger and older participants: Effects of age of acquisition. *Aging, Neuropsychology, and Cognition, 21*, 606–632. https://doi.org/10. 1080/13825585.2013.849653

Anderson, J. R. (1972). FRAN: A simulation model of free recall. *Psychology of Learning and Motivation 5*, 315-378. https://doi.org/10.1016/S0079-7421(08)60444-2

Angwin, A. J., Chenery, H. J., Copland, D. A., Murdoch, B. E., & Silburn, P. A. (2006). Self-paced reading and sentence comprehension in Parkinson's disease. *Journal of Neurolinguistics*, *19*(3), 239–252. https://doi.org/10.1016/j.jneuroling.2005.11.004

Baddeley A. D. (1986). *Working memory*. Oxford: Clarendon Press.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14*(6), 575–589. https://doi.org/10.1016/S0022-5371(75)80045-4

Balota, D., & Neely, J. H. (1980). Text-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology*, *6*(5), 576-587. https://doi.org/10.1037/0278-7393.6.5.576

Ballot, C., Stephanie, M., & Christelle, R. (2021). Word imageability and orthographic neighborhood effects on memory: A study in free recall and recognition. *Memory, 29*(6), 829-834. https://doi-org.ledproxy2.uwindsor.ca/10.1080/09658211.2021.1921216

Barker, M. S., Chapman, S., Mejia, A., Joyce, J. L., Sunderaraman, P., Idnay, B., Kreisl, W. C., & Cosentino, S. (2020). Psycholinguistic properties of words on fluency testing relate to different aspects of subjective cognitive decline. *The Journal of Alzheimer's & Dementia, 16*(S6). https://doi.org/10.1002/alz.046108

Baskak, B., Ozel, E. T., Atbasoglu, E. C., & Baskak, S. C. (2008). Peculiar word use as a possible trait marker in schizophrenia. *Schizophrenia Research, 103*(1–3), 311–317. https://doi.org/10.1016/j.schres.2008.04.025

Bayles, K. A., Salmon, D. P., Tomoeda, C. K., Jacobs, D., Caffrey, J. T., Kaszniak, A. W., & Troster, A. I. (1989). Semantic and letter category naming in Alzheimer's patients: A predictable difference. *Developmental Neuropsychology, 5*, 335-347. https://doi.org/10.1080/87565648909540443

Bayley, P. J., Hopkins, R. O., & Squire, L. R. (2003). Successful recollection of remote autobiographical memories by amnesic patients with medial temporal lobe lesions. *Neuron, 38*(1), 135–144. https://doi.org/10.1016/S0896-6273(03)00156-9

Beck, A.T., Steer, R.A., & Brown, G.K. (1996). *Beck depression inventory (BDI-II): Manual and questionnaire*. The Psychological Corporation.

Bhatarah, P., Ward, G., Smith, J., & Hayes, L. (2009). Examining the relationship between free recall and immediate serial recall: Similar patterns of rehearsal and similar effects of word length, presentation rate, and articulatory suppression. *Memory & Cognition, 37*(5), 689-713. https://doi.org/10.3758/MC.37.5.689

Bialystok, E., Craik, F., & Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology, 34*(4), 859–873. https://doi.org/10.1037/0278-7393.34.4.859

Bokat, C. E., & Goldberg, T. E. (2003). Letter and category fluency in schizophrenic patients: a meta-analysis. *Schizophrenia Research, 64*(1), 73–78. https://doi.org/10.1016/S0920-9964(02)00282-7

Bleuler, E., (1950). *Dementia praecox or the group of schizophrenias.* International Universities Press, New York.

Borkowska, A. R. (2015). Language and communicative functions as well as verbal fluency in children with high-functioning autism. *Journal of Intellectual Disability-Diagnosis and Treatment, 3*(3), 147-153.

Brandt, J. (1991). The Hopkins Verbal Learning Test: Development of a new memory test with six equivalent forms. *The Clinical Neuropsychologist, 5*(2), 125-142. https://doi.org/10.1080/13854049108403297

Brandt, J., & Benedict, R. H. B. (2001). *Hopkins verbal learning test – Revised. Administration manual.* Lutz, FL: Psychological Assessment Resources.

Buchanan, T., & Adolphs, R. (2002). The role of the human amygdala in emotional modulation of long-term declarative memory. In S. Moore & M. Oaksford (Eds.), *Emotional cognition: From brain to behavior* (pp. 9-34). London: Benjamins.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin and Review, 8*(3), 531-544. https://doi.org/10.3758/BF03196189

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*(3), 551-581. https://doi.org/10.1037/0033-295X.106.3.551

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich & A. Markman (Eds.), Cognitive dynamics: Conceptual and representational change in humans and machines (pp. 117-156). Hillsdale, NJ: Erlbaum.

Brébion, G., David, A. S., Bressan, R. A., & Pilowsky, L. S. (2005). Word frequency effects on free recall and recognition in patients with schizophrenia. *Journal of Psychiatric Research, 39*(2), 215–222. https://doi.org/10.1016/j.jpsychires.2004.05.010

Breining, B., & Rapp, B. (2019). Investigating the mechanisms of written word production: insights from the written blocked cyclic naming paradigm. *Reading and Writing*, *32*(1), 65–94. https://doi.org/10.1007/s11145-017-9742-4

Brysbaert, M., & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *Quarterly Journal of Experimental Psychology, 64*, 545– 559. https://doi.org/10.1080/17470218.2010.503374

Burke, D. M., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language, 30*(5), 542–579. https://doi.org/10.1016/0749-596X(91)90026-G

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology, 42*(3), 441-458. https://doi.org/10.1037/xhp0000159

Butters, N., Salmon, D. P., Heindel, W., & Granholm, E. (1988). Episodic, semantic, and procedural memory: Some comparisons of Alzheimer and Huntington disease patients. In R. D. Terry (Ed.), *Aging and the brain* (pp. 63–87). New York: Raven Press

Caldwell, A. R. (2022). TOSTER R package 0.5.1. CRAN.

Caplan, D., Rochon, E., & Waters, G. S. (1992). Articulatory and phonological determinants of word length effects in span tasks. *The Quarterly Journal of Experimental Psychology, 45*(2), 177-192. https://doi.org/10.1080/14640749208401323

Cattell, J. M. K. (1886). The time taken up by cerebral operations. *Scientific American Mind, 11*(42), 220. https://doi.org/10.1037/xhp0000159

Chen, R. Y. L., Chen, E. Y. H., Chan, C. K. Y., Lam, L. C. W., & Lieh-Mak, F. (2000). Verbal fluency in schizophrenia: reduction in semantic store. *Australian and New Zealand Journal of Psychiatry, 34*(1), 43–48. https://doi.org/10.1046/j.1440-1614.2000.00647.x

Cluff, M. S., & Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology, 16*(3), 551–563. https://doi.org/10.1037/0096-1523.16.3.551

Coane, J. H., Balota, D. A., Dolan, P. O., & Jacoby, L. L. (2011). Not all sources of familiarity are created equal: The case of word frequency and repetition in episodic recognition. *Memory & Cognition, 39*(5), 791-805. https://doi.org/10.3758/s13421-010-0069-5

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI* (pp. 535-555). Hillsdale, NJ: Lawrence Erlbaum.

Coltheart, V., & Winograd, E. (1986). Word imagery but not age of acquisition affects episodic memory. *Memory & Cognition, 14*, 174–179. https://doi.org/10.3758/BF03198377

Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *Quarterly Journal of Experimental Psychology, 60*, 1072–1082. https://doi.org/10.1080/17470210701315467

Cortese, M. J., Khanna, M. M., & Hacker, S. D. (2010). Recognition memory for 2,578 monosyllabic words. *Memory, 18*, 595–609. https://doi.org/10.1080/09658211.2010.493892

Cortese, M. J., McCarty, D. P., & Schock, J. (2015). A mega recognition memory study of 2897 disyllabic words. *Quarterly Journal of Experimental Psychology, 68*, 1489–1501. https://doi.org/10.1080/17470218.2014.945096

Cortese, M. J., Watson, J. M., Wang, J., & Fugett, A. (2004). Relating distinctive orthographic and phonological processes to episodic memory performance. *Memory & Cognition, 32*(4), 632-639. https://doi.org/10.3758/BF03195854

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Craik, F., & Lockhart, R. (1990). Levels of processing: A retrospective commentary on a framework for memory research. *Canadian Journal of Psychology, 44*(1), 87–112. https://doi.org/10.1037/h0084237

Craik, F., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. Journal of Experimental Psychology: General, 104(3), 268–294. https://doi.org/10.1037/0096- 3445.104.3.268

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Harcourt Brace Jovanovich.

Croisile, Ska, B., Brabant, M.-J., Duchene, A., Lepage, Y., Aimard, G., & Trillet, M. (1996). comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain and Language*, *53*(1), 1–19. https://doi.org/10.1006/brln.1996.0033

Criss, A. H., & Malmberg, K. J. (2008). Evidence in favor of the early-phase elevated attention hypothesis: The effects of letter frequency and object frequency. *Journal of memory and language*, *59*(3), 331-345. https://doi.org/10.1016/j.jml.2008.05.002

Danguecan, A. N., & Buchanan, L. (2016). Semantic neighborhood effects for abstract versus concrete words. Frontiers in Psychology, 7. https://doi.org/10.3389/fpsyg.2016.01034

Deese, J. (1960). Frequency of usage and number of words in free recall: The role of association. *Psychological Reports, 7*(2), 337-344. https://doi-org.ledproxy2.uwindsor.ca/10.2466%2Fpr0.1960.7.2.337

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis Kaplan Executive Function System: Technical manual.* San Antonio, TX: Psychological Corporation.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1994). *California Verbal Learning Test, Children's Version (CVLT-C)*. San Antonio, TX: The Psychological Corporation.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2017). *California Verbal Learning Test, Third Edition (CVLT-3)*. San Antonio, TX: Psychological Corporation.

DeLosh, E. L., McDaniel, M. A. (1996). The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology, 22*(5), 1136-1146. https://doi.org/10.1037/0278-7393.22.5.1136

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*, 284–298. https://doi.org/10.1037/0278-7393.24.2. 284

Diana, R. A., & Reder, L. M. (2006). The low-frequency encoding disadvantage: Word frequency affects processing demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 805-815. https://psycnet.apa.org/doi/10.1037/0278-7393.32.4.805

Diaz, Sailor, K., Cheung, D., & Kuslansky, G. (2004). Category size effects in semantic and letter fluency in Alzheimer's patients. *Brain and Language, 89*(1), 108–114. https://doi.org/10.1016/S0093-934X(03)00307-9

Dorfman, D., & Glanzer, M. (1988). List composition effects in lexical decision and recognition memory. *Journal of Memory and Language*, *27*(6), 633-649. https://doi.org/10.1016/0749-596X(88)90012-5

Doerksen, S., & Shimamura, A. P. (2001). Source memory enhancement for emotional eords. *Emotion*, *1*(1), 5–11. https://doi.org/10.1037/1528-3542.1.1.5

Dougal, S., & Rotello, C. M. (2007). "Remembering" emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review*, *14*(3), 423–429. https://doi.org/10.3758/BF03194083

Doughty, & Done, D. J. (2009). Is semantic memory impaired in schizophrenia? A systematic review and meta-analysis of 91 studies. *Cognitive Neuropsychiatry, 14*(6), 473–509. https://doi.org/10.1080/13546800903073291

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6*, 241-252.

Durda, K., & Buchanan, L. (2008). WINDSOR: Windsor improved norms of distance and similarity of representations of semantics. Behavior Research Methods, 40, 705-712. https://doi.org/10.3758/BRM.40.3.705

Durda, K., & Buchanan, L. (2006). WordMine2 [Online] Available: http://web2.uwindsor.ca/wordmine

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences, 8*(7), 301-306. https://doi.org/10.1016/j.tics.2004.05.003

Elvevåg, B., Weinstock, D., Akil, M., Kleinman, J., & Goldberg, T. (2001). A comparison of verbal fluency tasks in schizophrenic patients and normal controls. *Schizophrenia Research, 51*(2-3), 119–126. https://doi.org/10.1016/S0920-9964(00)00053-0

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence Erlbaum Associates Publishers.

Estes, W. K., & Maddox, W. T. (2002). On the process underlying stimulus-familiarity effects in recognition of words and nonwords. *Journal of Experimental Psychology, 28*(6), 1003-1018. https://doi.org/10.1037/0278-7393.28.6.1003

Fishman, K. N., Ashbaugh, A. R., Lanctôt, K. L., Cayley, M. L., Herrmann, N., Murray, B. J., Sicard, M., Lien, K., Sahlas, D. J., & Swartz, R. H. (2018). Apathy, not depressive symptoms, as a predictor of semantic and phonemic fluency task performance in stroke and transient ischemic attack. *Journal of Clinical and Experimental Neuropsychology, 40*(5), 449–461. https://doi.org/10.1080/13803395.2017.1371282

Forbes-McKay, K. E., Ellis, A. W., Shanks, M. F., & Venneri, A. (2005). The age of acquisition of words produced in a semantic fluency task can reliably differentiate normal from pathological age related cognitive decline. *Neuropsychologia*, *43*(11), 1625–1632. https://doi.org/10.1016/j.neuropsychologia.2005.01.008

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. Walker (Eds.), *New approaches to language mechanisms* (pp. 255-287). Amsterdam: North Holland.

Gilhooly, K. J., & Gilhooly, M. L. (1979). Age-of-acquisition effects in lexical and episodic memory tasks. *Memory & Cognition*, *7*, 214– 223. https://doi.org/10.3758/BF03197541

Gilhooly, K. J., & Watson, F. L. (1981). Word age-of-acquisition effects: A review. *Current Psychological Reviews*, *1*, 269–286. https://doi. org/10.1007/BF02684489

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. Psychological Review, 91(1), 1–67. https://doi.org/10.1037/0033-295X.91.1.1

Glanc, G. A., & Greene, R. L. (2007). Orthographic neighborhood size effects in recognition memory. *Memory & Cognition*, *35*(2), 365-371. https://doi.org/10.3758/BF03193457

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, *13*, 8-20. https://doi.org/10.3758/BF03198438

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology*, *16*(1), 5-16. https://doi.org/10.1037/0278-7393.16.1.5

Goh, W. D., & Pisoni, D. B. (2003). Effects of lexical competition on immediate memory span for spoken words. *The Quarterly Journal of Experimental Psychology, 56*(6), 929-954. https://doi.org/10.1080/02724980244000710

Gollan, T. H., Montoya, R. I., & Werner, G. A. (2002). Semantic and letter fluency in Spanish-English bilinguals. *Neuropsychology*, *16*(4), 562–576. https://doi.org/10.1037/0894-4105.16.4.562

Gomez, R., & white, D. (2006). Using verbal fluency to detect very mild dementia of the Alzheimer type. *Archives of Clinical Neuropsychology*, *21*(8), 771–775. https://doi.org/10.1016/j.acn.2006.06.012

Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology, 61*(1), 23-29. https://doi.org/10.1037/h0040561

Graham, K. S. & Hodges, J. R. (1997). Differentiating the roles of the hippocampus complex and the neocortex in long-term memory storage: Evidence from the study of semantic dementia and Alzheimer's disease. *Neuropsychology*, *11*(1), 77–89. https://doi.org/10.1037/0894-4105.11.1.77

Gregg, V. H., Freedman, C. M., & Smith, D. K. (1989). Word frequency, articulatory suppression and memory span. *British Journal of Psychology, 80*(3), 363–374. https://doi.org/10.1111/j.2044-8295.1989.tb02326.x

Gregg, V. H., Montgomery, D. C., & Castaño, D. (1980). Recall of common and uncommon words from pure and mixed lists. *Journal of Verbal Learning and Verbal Behavior, 19*(2), 240–245. https://doi.org/10.1016/S0022-5371(80)90202-9

Halderman, L. K., & Chiarello, C. (2005). Cerebral asymmetries in early orthographic and phonological reading processes: Evidence from backward masking. *Brain and Language 95*(2), 342-352. https://doi.org/10.1016/j.bandl.2005.02.005

Hamann, S. (2001). Cognitive and neural mechanisms of emotional memory. *Trends in Cognitive Sciences*, *5*(9), 394–400. https://doi.org/10.1016/S1364-6613(00)01707-1

Hargreaves, I. S., & Pexman, P. M. (2014). Get rich quick: The signal to respond procedure reveals the time course of semantic richness effects during visual word recognition. *Cognition, 131*(2), 216-242. https://doi.org/10.1016/j.cognition.2014.01.001

Hargreaves, I. S., Pexman, P. M., Johnson, J. S., & Zdrazilova, L. (2012). Richer concepts are better remembered: number of features effects in free recall. Frontiers in Human *Neuroscience*, *6*, 73. https://doi.org/10.3389/fnhum.2012.00073

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. https://doi.org/10.1037/0033-295X.111.3.662

Hart, S., Smith, C. M., & Swash, M. (1988). Word fluency in patients with early dementia of Alzheimer type. *British Journal of Clinical Psychology, 27*(2), 115-124. https://doi.org/10.1111/j.2044-8260.1988.tb00759.x

Heathcote, A., Ditton, E., & Mitchell, K. (2006). Word frequency and word likeness mirror effects in episodic recognition memory. *Memory & Cognition, 34,* 826-828. https://doi.org/10.3758/BF03193430

Henderson, S. K., Peterson, K. A., Patterson, K., Lambon Ralph, M. A., & Rowe, J. B. (2023). Verbal fluency tests assess global cognitive status but have limited diagnostic differentiation: evidence from a large-scale examination of six neurodegenerative diseases. *Brain Communications, 5*(2). https://doi.org/10.1093/braincomms/fcad042

Henry, J., & Crawford, J. (2005). A meta-analytic review of verbal fluency deficits in schizophrenia relative to other neurocognitive deficits. *Cognitive Neuropsychiatry, 10*(1), 1–33. https://doi.org/10.1080/13546800344000309

Henry, J. D., Crawford, J. R., & Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: A meta-analysis. *Neuropsychologia*, *42*(9), 1212–1222. https://doi.org/10.1016/j.neuropsychologia.2004.02.001

Heidlmayr, K., Hemforth, B., Moutier, S., & Isel, F. (2015). Neurodynamics of executive control processes in bilinguals: Evidence from ERP and source reconstruction analyses. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00821

Heinrichs, R. W., & Zakzanis, K. K. (1998). neurocognitive deficit in schizophrenia: a quantitative review of the evidence. *Neuropsychology*, *12*(3), 426–445. https://doi.org/10.1037/0894-4105.12.3.426

Hicks, J. L., Marsh, R. L., & Cook, G. I. (2005). An observation on the role of context variability in free recall. *Journal of Experimental Psychology 31*(5), 1160-1164. https://doi.org/10.1037/0278-7393.31.5.1160

Hills, T., Jones, M. & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review, 119*, 431–440. https://doi.org10.1037/a002737

Ho, A. K., Sahakian, B. J., Robbins, T. W., Barker, R. A., Rosser, A. E., & Hodges, J. R. (2002). Verbal fluency in Huntington's disease: A longitudinal analysis of phonemic and semantic clustering and switching. *Neuropsychologia*, *40*(8), 1277–1284. https://doi.org/10.1016/S0028-3932(01)00217-2

Hodges, J. R., & Patterson, K. (1995). Is semantic memory consistently impaired early in the course of Alzheimer's disease? Neuroanatomical and diagnostic implications. *Neuropsychologia*, *33*(4), 441-459. https://doi.org/10.1016/0028-3932(94)00127-B

Hodges, J. R., Patterson, K., Oxbury, S., & Funnell, E. (1992). Semantic dementia. Progressive fluent aphasia with temporal lobe atrophy. *Brain, 115*, 1783–1806.

Hodges, J. R., Patterson, K., & Tyler, L. K. (1994). Loss of semantic memory: Implications for the modularity of mind. *Cognitive Neuropsychology, 11*(5), 505-542. https://doi.org/10.1080/02643299408251984

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2012). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

Holmes, S. J., & Ellis, A. W. (2006). Age of acquisition and typicality effects in three object processing tasks. *Visual Cognition*, *13*(7–8), 884–910. https://doi.org/10.1080/13506280544000093

Huber, P. J. (1981). *Robust statistics*. New York: John Wiley & Sons.

Huff, F. J., Corkin, S., & Growdon, J. H. (1986). Semantic impairment and anomia in Alzheimer's disease. *Brain and Language*, *28*(2), 235–249. https://doi.org/10.1016/0093-934X(86)90103-3

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(5), 1217-1232. https://doi.org/10.1037/0278-7393.23.5.1217

Hulme, C., Suprenant, A. M., Bireta, A. M., Tamra, J., Stuart, G., & Neath, I. (2004). Abolishing the word-length effect. *Journal of Experimental Psychology*, *30*(1), 98-106. https://doi.org/10.1037/0278-7393.30.1.98

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513-541. https://doi.org/10.1016/0749-596X(91)90025-F

Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition, 13*, 789–845. https://doi.org/10. 1080/13506280544000066

Jalbert, A., Neath, I., & Surprenant, A. M. (2011). Does length or neighborhood size cause the word length effect? *Memory & Cognition, 39*(7), 1198-1210. https://doi.org/10.3758/s13421-011-0094-z

Johns, B. T., Sheppard, C. L., Jones, M. N., & Taler, V. (2016). The role of semantic diversity in word recognition across aging and bilingualism. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.00703

Johns, B. T., Taler, V., Pisoni, D. B., Farlow, M. R., Hake, A. M., Kareken, D. A., Unverzagt, F., & Jones, M. N. (2018). Cognitive modeling as an interface between brain and behavior: Measuring the semantic decline in mild cognitive impairment. *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Experimentale, 72*, 117–126. https://doi.org/10.1037/cep0000132

Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition, 13*(7-8), 789–845. https://doi.org/10.1080/13506280544000066

Joordens, S., & Hockley, W. E. (2000). Recollection and familiarity through the looking glass: When old does not mirror new. *Journal of Experimental Psychology, 26*(6), 1534-1555. https://doi.org/10.1037/0278-7393.26.6.1534

Juhasz, B. J., Chambers, D., Shesler, L. W., Haber, A., & Kurtz, M. M. (2012). Evaluating lexical characteristics of verbal fluency output in schizophrenia. *Psychiatry Research, 200*(2–3), 177–183. https://doi.org/10.1016/j.psychres.2012.06.035

Juhasz, B. J., Yap, M. J., Raoul, A., & Kaye, M. (2019). A further examination of word frequency and age-of-acquisition effects in English lexical decision task performance:

The role of frequency trajectory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*, 82–96. https://doi.org/10.1037/xlm0000564

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *Boston naming test.* Lea & Febiger.

Katkov, M., Romani, S., & Tsodyks, M. (2014). Word length effect in free recall of randomly assembled word lists. *Frontiers in Computational Neuroscience, 8*, 129. https://doi.org/10.3389/fncom.2014.00129

Kensinger, E., & Corkin, S. (2003). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & Cognition*, *31*(8), 1169–1180. https://doi.org/10.3758/BF03195800

Kertesz, A., Davidson, W., Mccabe, P., Takagi, K., & Munoz, D. (2003). Primary progressive aphasia: Diagnosis, varieties, evolution. *Journal of the International Neuropsychological Society, 9*(5), 710–719. https://doi.org/10.1017/S1355617703950041

Kiang, M., Christensen, B. K., Kutas, M., & Zipursky, R. B. (2012). Electrophysiological evidence for primary semantic memory functional organization deficits in schizophrenia. *Psychiatry Research*, *196*(2–3), 171–180. https://doi.org/10.1016/j.psychres.2012.02.026

Kiang, M., & Kutas, M. (2006). Abnormal typicality of responses on a category fluency task in schizotypy. *Psychiatry Research*, *145*(2–3), 119–126. https://doi.org/10.1016/j.psychres.2005.12.010

Klaver, P., Fell, J., Dietl, T., Schür, S., Schaller, C., Elger, C. E., & Fernández, G. (2005). Word imageability affects the hippocampus in recognition memory. *Hippocampus*, *15*(6), 704–712. https://doi.org/10.1002/hipo.20081

Korkman, M., Kirk, U., & Kemp, S. (2007). NEPSY - Second Edition (NEPSY - II) [Database record]. APA PsycTests. https://doi.org/10.1037/t15125-000

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

Laisney, M., Matuszewski, V., Mézenge, F., Belliard, S., Sayette, V., Eustache, F., & Desgranges, B. (2009). The underlying mechanisms of verbal fluency deficit in frontotemporal dementia and semantic dementia. *Journal of Neurology, 256*(7), 1083–1094. https://doi.org/10.1007/s00415-009-5073-y

Lakens, D., Scheel, A., Isager, P. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259-269. https://doi.org/10.1177/2515245918770963

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. The Psychological Review., 104(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology, 30,* 261-273. https://doi.org/10.1111/j.1469-8986.1993.tb03352.x

Lau, M. C., Goh, W. D., & Yap, M. J. (2018). An item-level analysis of lexical-semantic effects in free recall and recognition memory using the megastudy approach. *Quarterly Journal of Experimental Psychology*, *71*(10), 2207–2222. https://doi.org/10.1177/1747021817739834

Leiphart, J., Rosenfeld, J. P., & Gabrieli, J. D. (1993). Event-related potential correlates of implicit priming and explicit memory tasks. *International Journal of Psychophysiology, 15*(3), 197–206. https://doi.org/10.1016/0167-8760(93)90003-8

Leeson, McKenna, P. J., & Laws, K. R. (2005). Storage and access procedures in schizophrenia: Evidence for a two phase model of lexical impairment. *Journal of Clinical and Experimental Neuropsychology, 27*(6), 700–710. https://doi.org/10.1080/13803390490918507

Lezak, M., Howieson, D., Bigler, E., Tranel, D. (2004). Neuropsychological assessment (5th Ed.). *Oxford University Press.*

Libben, G., & Jarema, G. (2002). Mental lexicon research in the new millennium. *Brain and Language, 8*(1-3), 2-11. https://doi.org/10.1006/brln.2002.2654

Libon, D. J., Swenson, R., Ashendorf, L., Russell, B., & Bowers, D. (2013). Edith Kaplan and the boston process approach. *The Clinical Neuropsychologist, 27*(3), 1223-1233. https://doi-org.ledproxy2.uwindsor.ca/10.1080/13854046.2013.833295

Lohnas, L. J., & Kahana, M. J. (2013). Parametric effects of word frequency in memory for mixed frequency lists. *Journal of Experimental Psychology, 39*(6), 1943-1946. 1943–1946. https://doi.org/10.1037/a0033669

Lovatt, P., Avons, S. E., & Masterson, J. (2000). The word-length effect and disyllabic words. *The Quarterly Journal of Experimental Psychology, 53*(1), 1-22. https://doi.org/10.1080/713755877

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing, 19*(1), 1-36. https://doi.org/10.1097%2F00003446-199802000-00001

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 122–147). The MIT Press.

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

Lynott, D., Connell, L., Brysbaert, M., Brand, K., & Carney, J. (2020). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, *52*, 1271-291. https://doi.org/10.3758/s13428-019-01316-z

MacLeod, C. M., & Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 132-142. https://psycnet.apa.org/doi/10.1037/0278-7393.22.1.132

Macmillan, M. B., Neath, I., Surprenant, A. M. (2021). Re-assessing age of acquisition effects in recognition, free recall, and serial recall. *Memory & Cognition, 49*, 939-954. https://doi.org/10.3758/s13421-021-01137-6

Malmberg, K. J., & Murnane, K. (2002). List composition and the word-frequency effect for recognition memory. *Journal of Experimental Psychology, 28*(4), 616-630. https://doi.org/10.1037/0278-7393.28.4.616

Malmberg, K. J., Zeelenberg, R., & Shiffrin, R. M. (2004). Turning up the noise or turning down the volume? On the nature of the impairment of episodic recognition memory by Midazolam. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition, 30*, 540-549. https://doi.org/10.1037/0278-7393.30.2.540

May, R. B., Cuddy, L. J., & Norton, J. M. (1979). Temporal contrast and the word frequency effect. *Canadian Journal of Psychology*, *33*(3), 141-147. https://doi.org/10.1037/h0081712

Mayr, U. (2002). On the dissociation between clustering and switching in verbal fluency: Comment on Troyer, Moscovitch, Winocur, Alexander and Stuss. *Neuropsychologia.*, *40*(5), 562–566. https://doi.org/10.1016/S0028-3932(01)00132-4

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*(4), 724-760. https://doi.org/10.1037/0033-295X.105.4.734-760

McLeod, C. M., & Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of Experimental Psychology*, *22*(1), 132-142. https://doi.org/10.1037/0278-7393.22.1.132

McNamara, T. P. (1994). Theories of priming: II. Types of primes. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(3), 507-520. https://doi.org/10.1037/0278-7393.20.3.507

McNamara, P., Obler, L. K., Au, R., Durso, R., & Albert, M. L. (1992). Speech monitoring skills in Alzheimer’s disease, Parkinson’s disease, and normal aging. *Brain and Language*, *42*(1), 38–51. https://doi.org/10.1016/0093-934X(92)90055-J

Merritt, P. S., DeLosh, E. L., & McDaniel, M. A. (2006). Effects of word frequency on individual-item and serial order retention: Tests of the order-encoding view. *Memory & Cognition*, *34*(8), 1615-1627. https://doi.org/10.3758/BF03195924

Mendez, M. F., Clark, D. G., Shapira, J. S., & Cummings, J. L. (2003). Speech and language in progressive nonfluent aphasia compared with early Alzheimer's disease. *Neurology, 61*(8), 1108–1113. https://doi.org/10.1212/01.wnl.0000090563.97453.90

Meschyan, G., & Hernandez, A. (2002). Age of acquisition and word frequency. *Memory & Cognition, 30*, 262–269. https://doi.org/10. 3758/BF03195287

Monsch, A. U., Bondi, M. W., Butters, N., Salmon, D. P., Katzman, R., & Thal, L. J. (1992). Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type. *Archives of neurology, 49*(12), 1253-1258. https://doi.org/10.1001/archneur.1992.00530360051017

Morris, P. E. (1981). Age of acquisition, imagery, recall, and the limitations of multiple-regression analysis. *Memory & Cognition, 9*, 277– 282. https://doi.org/10.3758/BF03196961

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*(2), 165-178. https://doi.org/10.1037/h0027366

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(4), 863–883. https://doi.org/10.1037/0278-7393.21.4.863

Murphy, K. J., Rich, J. B., & Troyer, A. K. (2006). Verbal fluency patterns in amnestic mild cognitive impairment are characteristic of Alzheimer's type dementia. *Journal of the International Neuropsychological Society, 12*(4). https://doi.org/10.1017/S1355617706060590

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review, 111*(3), 721-756. https://doi.org/10.1037/0033-295X.111.3.721

Neath, I., Bireta, T. J., & Surprenant, A. M. (2003). The time-based word length effect and stimulus set specificity. *Psychonomic Bulletin & Review, 10*(2), 430-434. https://doi.org/10.3758/BF03196502

Nelson, D. L., Kitto, K., Galea, D., McEvoy, C. L., & Bruza, P. D. (2013). How activation, entanglement, and searching a semantic network contribute to event memory. *Memory & Cognition, 41*(6), 797–819. https://doi.org/10.3758/s13421-013-0312-y

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review, 105*(2), 299–324. https://doi.org/10.1037/0033-295X.105.2.299

Nunnally, J. C., & Bernstien, I. H. (1994). Psychometric theory (3rd Ed.). *McGraw-Hill*

Oldfield, R. C. (1966). Things, words, and the brain. *The Quarterly Journal of Experimental Psychology, 18*(4), 340-535. https://doi.org/10.1080/14640746608400052

Ochsner, K. N. (2000). Are affective events richly recollected or simply familiar? The experience and process of recognizing feelings past. *Journal of Experimental Psychology, 129*(2), 242–261. https://doi.org/10.1037/0096-3445.129.2.242

Ozubko, J. D., & Joordens, S. (2007). The mixed truth about frequency effects on free recall: Effects of study list composition. *Psychonomic Bulletin & Review, 14*(5), 871-876. https://doi.org/10.3758/BF03194114

Page, M. P., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review, 105*(4), 761-781. https://doi.org/10.1037/0033-295X.105.4.761-781

Paivio, A. (1991). Dual coding theory: Retrospect and current Status. *Canadian Journal of Psychology, 45*(3), 255–287. https://doi.org/10.1037/h0084295

Paivio, A., Walsh, M., & Bons, T. (1994). Concreteness effects on memory: When and why? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1196–1204. https://doi.org/10.1037/0278-7393.20.5.1196

Paulsen, Romero, R., Chan, A., Davis, A. V., Heaton, R. K., & Jeste, D. V. (1996). Impairment of the semantic network in schizophrenia. *Psychiatry Research, 63*(2-3), 109–121. https://doi.org/10.1016/0165-1781(96)02901-0

Pérez, M. A. (2007). Age of acquisition persists as the main factor in picture naming when cumulative word frequency and frequency trajectory are controlled. *Quarterly Journal of Experimental Psychology, 60*, 32–42. https://doi.org/10.1080/ 17470210600577423

Perry, R. J., Watson, P., & Hodges, J. R. (2000). The nature and staging of attention dysfunction in early (minimal and mild) Alzheimer's disease: Relationship to episodic and semantic memory impairment. *Neuropsychologia, 38*(3), 252–271. https://doi.org/10.1016/S0028-3932(99)00079-2

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Archives of Neurology, 56*(3), 303-308. https://doi.org/10.1001/archneur.56.3.303

Petersen R. C., Doody R., Kurz A., Mohs R. C., Morris J. C., Rabins P. V., et al. (2001). Current concepts in mild cognitive impairment. *Archives of Neurology, 58* 1985–1992. https://doi.org/10.1001/archneur.58.12.1985

Pexman, P., Hargreaves, I., Siakaluk, P., Bodner, G., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review, 15*(1), 161–167. https://doi.org/10.3758/PBR.15.1.161

Phelps, E. A., LaBar, K. S., & Spencer, D. D. (1997). Memory for emotional words following unilateral temporal lobectomy. *Brain and Cognition, 35*(1), 85–109. https://doi.org/10.1006/brcg.1997.0929

Popov, V., & Reder, L. M. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review*, *127*(1), 1-46. https://psycnet.apa.org/doi/10.1037/rev0000161

Portocarrero, J., Burright, R., & Donovick, P. (2007). Vocabulary and verbal fluency of bilingual and monolingual college students. *Archives of Clinical Neuropsychology, 22*(3), 415–422. https://doi.org/10.1016/j.acn.2007.01.015

Preston, K. A. (1935). The speed of word perception and its relation to reading ability. *Journal of General Psychology, 13*, 199-203.

Price, C. C., & Grossman, M. (2005). Verb agreements during on-line sentence processing in Alzheimer's disease and frontotemporal dementia. *Brain and Language, 94*(2), 217–232. https://doi.org/10.1016/j.bandl.2004.12.009

R Core Team (2021). *R: A language and environment for statistical computing.* (Version 4.1) [Computer Software].

Rabovsky, M., Sommer, W., Abdel Rahman, R. (2012). The time course of semantic richness effects in visual word recognition. *Frontiers in Human Neuroscience, 6*, https://doi.org/10.3389/fnhum.2012.00011

Randolph, C., Tierney, M. C., Mohr, E., & Chase, T. N. (1998). The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): Preliminary clinical validity.

Journal of Clinical and Experimental Neuropsychology, 20(3), 310–319. https://doi.org/10.1076/jcen.20.3.310.823

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgements in a continuous recognition paradigm. *Journal of Experimental Psychology, 26*(2), 294-320. https://doi.org/10.1037/0278-7393.26.2.294

Reder, L. M., Paynter, C., Diana, R. A., Ngiam, J., & Dickison, D. (2007). Experience is a double-edged sword: A computational model of the encoding/retrieval trade-off with familiarity. *Psychology of Learning and Motivation, 48*, 271-312. https://doi.org/10.1016/S0079-7421(07)48007-0

Reitan, R.M. and Wolfson, D. (1993) *The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation* (2nd Ed.), Neuropsychology Press, Tucson.

Richardson, J. T. E. (1975). Concreteness and imageability. *Quarterly Journal of Experimental Psychology, 27*(2), 235–249. https://doi.org/10.1080/14640747508400483

Robert, C., & Rico Duarte, L. (2016). Semantic richness and aging: The effect of number of features in the lexical decision task. *Journal of Psycholinguistic Research, 45*(2), 359–365. https://doi.org/10.1007/s10936-015-9352-8

Rohrer, D., Salmon, D. P., Wixted, J. T., & Paulsen, J. S. (1999). The disparate effects of Alzheimer's disease and Huntington's disease on semantic memory. *Neuropsychology., 13*(3), 381–388. https://doi.org/10.1037/0894-4105.13.3.381

Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of*

Experimental Psychology, 28(6), 1019-1034. https://doi-org.ledproxy2.uwindsor.ca/10.1037/0278-7393.28.6.1019

Rosen, W. G. (1980). Verbal fluency in aging and dementia. *Journal of Clinical and Experimental Neuropsychology, 2*(2), 135-146. https://doi.org/10.1080/01688638008403788

Rosselli, M., Ardila, A., Salvatierra, J., Marquez, M., Luis, M., & Weekes, V. A. (2002). A cross-linguistic comparison of verbal fluency tests. *International Journal of Neuroscience, 112*(6), 759–776. https://doi.org/10.1080/00207450290025752

Rosselli, M., Ardila, A., Araujo, K., Weekes, V. A., Caracciolo, V., Padilla, M., & Ostrosky-Solí, F. (2000). Verbal fluency and repetition skills in healthy older spanish-english bilinguals. *Applied Neuropsychology, 7*(1), 17–24. https://doi.org/10.1207/S15324826AN0701_3

Russo, R., & Grammatopoulou, N. (2003). Word length and articulatory suppression affect short-term and long-term recall tasks. *Memory & Cognition, 31*(5), 728-737. https://doi.org/10.3758/BF03196111

Sandoval, T. C., Gollan, T. H., Ferreira, V. S., & Salmon, D. P. (2010). What causes the bilingual disadvantage in verbal fluency? The dual-task analogy. *Bilingualism: Language and Cognition, 13*(2), 231–252. https://doi.org/10.1017/S1366728909990514

Sass, K., Fetz, K., Oetken, S., Habel, U., & Heim, S. (2013). Emotional verbal fluency: A new task on emotion and executive function interaction. *Behavioral Sciences, 3*(3), 372–387. https://doi.org/10.3390/bs3030372

Sailor, K. M., Zimmerman, M. E., & Sanders, A. E. (2011). Differential impacts of age of acquisition on letter and semantic fluency in Alzheimer's disease patients and healthy older adults. *The Quarterly Journal of Experimental Psychology: QJEP, 64*(12), 2383–2391. https://doi.org/10.1080/17470218.2011.596660

Schacter, Daniel L., & Tulving, E. (1994). Memory systems 1994. Cambridge, Mass.: MIT Press.

Scheffel, L., Duffy, J. R., Strand, E. A., & Josephs, K. A. (2021). Word fluency test performance in primary progressive aphasia and primary progressive apraxia of speech. *American Journal of Speech-Language Pathology*, *30*, 2635–2642. https://doi-org.ledproxy2.uwindsor.ca/10.1044/2021_AJSLP-21-00058

Schriefers, H. (1992). Lexical access in the production of noun phrases. *Cognition*, *45*(1), 33-54. https://doi.org/10.1016/0010-0277(92)90022-A

Schrank, F. A., & Wendling, B. J. (2018). The Woodcock–Johnson IV: Tests of cognitive abilities, tests of oral language, tests of achievement. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 383–451). The Guilford Press.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2018). The Glasgow norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods, 51*(3), 1258– 1270. https://doi.org/10.3758/s13428-018-1099-3

Seidenburg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*(4), 521-568. https://doi.org/10.1037/0033-295X.96.4.523

Service, E. (1998). The effect of word length on immediate serial recall depends on phonological complexity, not articulatory duration. *The Quarterly Journal of Experimental Psychology, 51*(2), 283-304. https://doi.org/10.1080/713755759

Sherman, E. M. S., Brooks, B. L. (2015). *Child and Adolescent Memory Profile (ChAMP).* Psychological Assessment Resources.

Sheslow, D., & Adams, W. (2003). *Wide range assessment of memory and learning* (2nd ed.). Lutz, FL: Psychological Assessment Resources, Inc.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM – retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145-166. https://doi.org/10.3758/BF03209391

Sommers, M. S. (1996). The structural organization of the mental lexicon and its contribution to age-related declines in spoken-word recognition. *Psychology and Aging, 11*(2), 333-341. https://doi-org/10.1037/0882-7974.11.2.333

Spek, A., Schatorjé, T., Scholte, E., & van Berckelaer-Onnes, I. (2009). Verbal fluency in adults with high functioning autism or Asperger syndrome. *Neuropsychologia*, *47*(3), 652–656. https://doi.org/10.1016/j.neuropsychologia.2008.11.015

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). A compendium of neuropsychological tests: Administration, norms, and commentary (*3rd ed.*) *Oxford University Press.*

Sumby, W. H. (1963). Word frequency and serial position effects. *Journal of Verbal Learning and Verbal Behavior*, *1*(6), 443-450. https://doi-org.ledproxy2.uwindsor.ca/10.1016/S0022-5371(63)80030-4

Taler, V., Davidson, P. S., Sheppard, C., & Gardiner, J. (2021). A discourse-theoretic approach to story recall in aging and mild cognitive impairment. *Aging, Neuropsychology, and Cognition*, *28*(5), 762–780. https://doi.org/10.1080/13825585.2020.1821865

Taler, V., & Johns, B. (2022). Using big data to understand bilingual performance in semantic fluency: Findings from the Canadian Longitudinal Study on Aging. *PloS One, 17*(11). https://doi.org/10.1371/journal.pone.0277660

Taler, V., Johns, B. T., & Jones, M. N. (2020). A large-scale semantic analysis of verbal fluency across the aging spectrum: Data from the Canadian longitudinal study on aging. *The Journals of Gerontology: Series B, 75*(9), e221–e230. https://doi.org/10.1093/geronb/gbz003

Taler, V., Johns, B., Young, K., Sheppard, C., & Jones, M.N. (2013). A computational analysis of semantic structure in bilingual fluency. *Journal of Memory and Language, 69*, 607–618. https://doi.org/10.1016/j.jml.2013.08.004

Taler, V., & Phillips, N. A. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Neuropsychology, 30*(5), 501–556. https://doi.org/10.1080/13803390701550128

Tehan, G., & Humphreys, M.S. (1988). Articulatory loop explanations of memory span and pronunciation rate correspondences: A cautionary note. *Bulletin of Psychonomic Society, 26*, 293–296. https://doi.org/10.3758/BF03337662

Tehan, G., & Tolan, G. A., (2007). Word length effects in long-term memory. *Journal of Memory and Language, 56*, 35-48. https://doi.org/10.1016/j.jml.2006.08.015

Tóth, O., Pesthy, O., Farkas, K., Guttengéber, A., Komoróczy, E., Réthelyi, J. M., Szuromi, B., & Németh, D. (2022). Intact fluency in autism? A comprehensive approach of verbal fluency task including word imageability and concreteness. *Autism Research, 15*(4), 677–686. https://doi.org/10.1002/aur.2672

Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology, 12*, 242-248. https://doi.org/10.1080/17470216008416732

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology*, *11*(1), 138–146. https://doi.org/10.1037/0894-4105.11.1.138

Troyer, A. K., Moscovitch, M., Winocur, G., Leach, L., & Freedman, and M. (1998). Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. *Journal of the International Neuropsychological Society*, *4*(2), 137–143. https://doi.org/10.1017/S1355617798001374

Tsakiris, M. (2012). *Sex differences and autism: Brain function during verbal fluency and mental rotation*. *PloS One*, *7*(6). https://doi.org/10.1371/journal.pone.0038355

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26*(1), 1–12. https://doi.org/10.1037/h0080017

Vakil, E. (2012). Neuropsychological assessment: Principles, rationale, and challenges. *Journal of Clinical and Experimental Neuropsychology, 34*(2), 135-150. https://doi-org.ledproxy2.uwindsor.ca/10.1080/13803395.2011.623121

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1-67. https://www.jstatsoft.org/v45/i03/

Vandenberghe, R. R., Vandenbulcke, M., Weintraub, S., Johnson, N., Porke, K., Thompson, C. K., & Mesulam, M. M. (2005). Paradoxical features of word finding difficulty in primary progressive aphasia. *Annals of Neurology, 57*(2), 204–209. https://doi.org/10.1002/ana.20362

Venneri, A., McGeown, W. J., Hietanen, H. M., Guerrini, C., Ellis, A. W., & Shanks, M. F. (2008). The anatomical bases of semantic retrieval deficits in early Alzheimer's disease. *Neuropsychologia, 46*(2), 497–510. https://doi.org/10.1016/j.neuropsychologia.2007.08.026

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological science*, *9*(4), 325-329. https://doi.org/10.1111/1467-9280.00064

Vogel, A. P., Chenery, H. J., Dart, C. M., Doan, B., Tan, M., & Copland, D. A. (2009). Verbal fluency, semantics, context and symptom complexes in schizophrenia. *Journal of Psycholinguistic Research, 38*(5), 459–473. https://doi.org/10.1007/s10936-009-9100-z

Vogel, A., Vogel, A., Gade, J., & Stokholm, G. (2005). Semantic memory impairment in the earliest phases of Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders, 19*(2-3), 75–81. https://doi.org/10.1159/000082352

Ward, G., Woodward, G., Stevens, A., & Stinson, C. (2003). Using overt rehearsals to explain word frequency effects in free recall. *Journal of Experimental Psychology, 29*(2), 186-210. https://doi.org/10.1037/0278-7393.29.2.186

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191- 1207. https://doi.org/10.3758/s13428-012-0314-x

Watkins, M. J., LeCompte, D. C., & Kyungmi, K. (2000). Role of study strategy in recall of mixed lists of common and rare words. *Journal of Experimental Psychology, 26*(1), 239-245. https://doi.org/10.1037/0278-7393.26.1.239

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV) Administration and Scoring Manual.* San Antonio, TX: Pearson.

Wechsler, D. (2011). *The test of premorbid functioning (TOPF).* San Antonio, TX: The Psychological Corporation.

Windmann, & Kutas, M. (2001). Electrophysiological correlates of emotion-induced recognition bias. *Journal of Cognitive Neuroscience*, *13*(5), 577–592. https://doi.org/10.1162/089892901750363172

Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review, 1*, 89–106. https://doi-org.ledproxy2.uwindsor.ca/10.3758/BF03200763

Wong Gonzalez, D. (2018). The relationship between semantic and episodic memory: Exploring the effect of semantic neighborhood density on episodic memory (Publication No. 7589). [Doctoral dissertation, University of Windsor] https://scholar.uwindsor.ca/etd/7585

Wright, C.E. (1979). Duration differences between rare and common words and their implications for the interpretation of word frequency effects. *Memory & Cognition*, *7*, 411–419. https://doi.org/10.3758/BF03198257

Yap, M. J., & Balota, D. A. (2015). Visual word recognition. In A. Pollastek & R. Treiman (Eds.), *The Oxford Handbook of Reading* (pp. 26-43). Oxford University Press.

Yap, M., Tan, S., Pexman, P., & Hargreaves, I. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review, 18*(4), 742–750. https://doi.org/10.3758/s13423-011-0092-y

Zakzanis, K. K., McDonald, K., & Troyer, A. K. (2011). Component analysis of verbal fluency in patients with mild traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology*, *33*(7), 785–792. https://doi.org/10.1080/13803395.2011.558496

VITA AUCTORIS

NAME:	Brette E. Lansue
PLACE OF BIRTH:	Windsor, ON
YEAR OF BIRTH:	1994
EDUCATION:	General Amherst High School, Amherstburg, ON, 2012
	University of Windsor, B.A., Windsor, ON, 2018
	University of Windsor, M.A., Windsor, ON, 2020