

2-2018


Reasoning with Pseudowords: How Properties of Novel Verbal Stimuli Influence Item Difficulty and Linguistic-Group Score Differences on Cognitive Ability Assessments

Paul Agnello

The Graduate Center, City University of New York

How does access to this work benefit you? Let us know!

Follow this and additional works at: https://academicworks.cuny.edu/gc_etds

 Part of the [Cognitive Psychology Commons](#), [Industrial and Organizational Psychology Commons](#), and the [Psycholinguistics and Neurolinguistics Commons](#)

Recommended Citation

Agnello, Paul, "Reasoning with Pseudowords: How Properties of Novel Verbal Stimuli Influence Item Difficulty and Linguistic-Group Score Differences on Cognitive Ability Assessments" (2018). *CUNY Academic Works*.
https://academicworks.cuny.edu/gc_etds/2505

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.

REASONING WITH PSEUDOWORDS: HOW PROPERTIES OF NOVEL VERBAL
STIMULI INFLUENCE ITEM DIFFICULTY AND LINGUISTIC-GROUP SCORE
DIFFERENCES ON COGNITIVE ABILITY ASSESSMENTS

by

PAUL AGNELLO

A dissertation submitted to the Graduate Faculty of Psychology in
partial fulfillment of the requirements of the degree of Doctor of Philosophy,
The City University of New York

2018

© 2018
Paul Agnello
All Rights Reserved

REASONING WITH PSEUDOWORDS: HOW PROPERTIES OF NOVEL VERBAL
STIMULI INFLUENCE ITEM DIFFICULTY AND LINGUISTIC-GROUP SCORE
DIFFERENCES ON COGNITIVE ABILITY ASSESSMENTS

by

PAUL AGNELLO

This manuscript has been read and accepted for the Graduate Faculty in
Psychology to satisfy the dissertation
requirement for the degree of Doctor of Philosophy.

Charles A. Scherbaum

Date

Chair of Examining Committee

Richard Bodnar

Date

Executive Officer

Charles A. Scherbaum

Harold Goldstein

Erin Eatough

Logan Watts

Kenneth Yusko

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

REASONING WITH PSEUDOWORDS: HOW PROPERTIES OF NOVEL VERBAL
STIMULI INFLUENCE ITEM DIFFICULTY AND LINGUISTIC-GROUP SCORE
DIFFERENCES ON COGNITIVE ABILITY ASSESSMENTS

by

Paul Agnello

Advisor: Dr. Charles A. Scherbaum

Pseudowords (words that are not real but resemble real words in a language) have been used increasingly as a technique to reduce contamination due to construct-irrelevant variance in assessments of verbal fluid reasoning (Gf). However, despite pseudowords being researched heavily in other psychology sub-disciplines, they have received little attention in cognitive ability testing contexts. Thus, there has been an assumption that all pseudowords work equally and work equally well for all test-takers. The current research examined three objectives with the first being whether changes to the pseudoword properties of length and wordlikeness (how much a pseudoword resembles a typical or common word in English) led to changes in item difficulty on verbal Gf items. The second objective was whether boundary conditions existed such that changes to pseudoword properties would differentially impact two linguistic sub-groups of participants – those who have English as their dominant language and those who do not have English as their dominant language. The last objective was to index and explore performance on these verbal Gf items when pseudowords were replaced with real words. Hypotheses predicting how pseudoword properties influenced item difficulty, how stimulus type – pseudoword or real word, impacted performance across linguistic sub-groups, and how linguistic sub-group status interacted with pseudoword properties were tested. Four sets of pseudowords were developed – short and wordlike, long and wordlike, short and un-wordlike, and long and un-wordlike, as well

as two sets of real words – short and wordlike, and long and un-wordlike. Sixteen verbal Gf items, adapted from the LSAT, were developed to accommodate the pseudowords or real words and explore these three objectives. While none of the hypotheses were statistically significant, the results did indicate further areas of exploration. Specifically, verbal Gf items were easier when they featured longer pseudowords and more difficult when they featured un-wordlike pseudowords. Additionally, while performance of English-non-dominant participants was fairly balanced across real and pseudoword sets, English-dominant participants performed better on items featuring real words. Similarly, linguistic status interacted with wordlikeness such that English-dominant participants featured a decrease in performance as pseudowords moved from wordlike to un-wordlike. A full discussion of the findings, their implications, limitations of the current study, and directions for future research are included.

Acknowledgements

There are numerous individuals to whom I am grateful for helping me complete this research. First and foremost, none of this would be possible without my advisor, Charles Scherbaum who was a constant source of guidance, patience, optimism, and camaraderie throughout the entire graduate school process. I would also like to thank Harold Goldstein and Erin Eatough for serving on my committee as well as Logan Watts and Kenneth Yusko for serving as outside readers. Due to your questions, comments, and contributions, this research was greatly refined and helped create a perfectly cromulent dissertation. This work also would not have been possible without my cohort, fellow graduate students, family, and friends who always provided support and much-needed relief from the grind of completing a dissertation. Lastly, I would like to thank my Mom and Dad for everything they provided for me leading up to and through graduate school. Their enthusiasm throughout my efforts was a perpetual source of motivation.

Table of Contents

Chapter 1: Introduction	1
Chapter 2: Modern Cognitive Ability Tests.....	14
Chapter 3: On Fluid Intelligence, Working Memory, and Pseudowords	22
Chapter 4: Pseudoword Properties and the Implication for Item Properties	38
Chapter 5: Bilingualism	49
Chapter 6: Pilot Study One	72
Chapter 7: Pseudoword and Word Group Development	76
Chapter 8: Item Development.....	85
Chapter 9: Pilot Study Two.....	91
Chapter 10: Methods.....	94
Chapter 11: Results	105
Chapter 12: Discussion	113
Appendix A: Verbal Fluid Intelligence Items with Pseudowords	125
Appendix B: Linguistic Background Measure	129
Appendix C: Socioeconomic Status Measure.....	130
Appendix D: Demographic Measure	131
References.....	145

List of Tables

Table 1. Pilot Study One Condition Properties.....	132
Table 2. Pilot Study One Pseudoword Properties	133
Table 3. Short and Wordlike Pseudoword Properties	135
Table 4. Short and Un-wordlike Pseudoword Properties	136
Table 5. Long and Wordlike Pseudoword Properties	137
Table 6. Long and Un-wordlike Pseudoword Properties.....	138
Table 7. Short and Wordlike Word Properties.....	139
Table 8. Long and Un-wordlike Word Properties	140
Table 9. Pseudoword and Word Property Means by Stimuli Set	141
Table 10. Pilot Study Two Item Stem and Item Difficulties	142
Table 11. Mean Total Score, Vocabulary Score, and Socioeconomic Status (SES) by Stimuli Set	143
Table 12. Mean Item and Total Scores Overall and by Linguistic Sub-group	144

Chapter 1: Introduction

The study and measurement of intelligence¹ and cognitive processes has an extensive history (Baddeley, 2007; Lubinski, 2004; Neisser, et al., 1996; Spearman, 1927) as well as a central importance for the field of psychology (Brand, 1987). However, contemporary intelligence research remains replete with debates and controversies at both definitional and operational levels that would look very familiar to the pioneers of intelligence research. A consensus definition of intelligence has eluded researchers and theorists for decades (Boring, 1923; Fischer, 1969; Gregory, 2004; Fagan, 2000; Sternberg & Detterman, 1986; van der Maas, Kan, & Borsboom, 2014; Wasserman & Tulsky, 2005). Perhaps as a result of this, there have been numerous efforts to determine what intelligence is and/or exactly what falls under the umbrella of intelligence - leading to continual refinement of the construct over time (Carroll, 1993; Cattell, 1944; Chen & Gardner, 2012; Drasgow, 2003; Haier et al., 1988; Jensen, 1998; Jung & Haier, 2007; Kyllonen & Christal, 1990; Naglieri, Das, & Goldstein, 2012; Schneider & McGrew, 2012; Spearman, 1927; Sternberg, 2012; Thurstone, 1938; van der Maas, Dolan, Grasman, Wicherts, Huizenga, & Raijmakers, 2006; Vernon, 1950).

Somewhat endemic from the definitional evolution, the measurement of intelligence has experienced several long-running challenges including the lack of theory guiding test construction (Ittenbach, Esters, & Wainer, 1996; Scherbaum, Goldstein, Ryan, Agnello, Yusko, & Hanges, 2015; Spearman, 1927), the effective development of assessments for different populations (Goodenough, 1949; Lewis & Sullivan, 1985; Malda, van de Vijver, & Temane, 2010; Mather & Wendling, 2012; Meeker, 1985; Ortiz, Ochoa, & Dynda, 2012; Wolman, 1985),

¹ The current paper makes no distinction between the terms ‘intelligence’ and ‘cognitive ability’ and as such, uses them interchangeably.

and the removal of construct-irrelevant variance from assessments (Agnello, Ryan, & Yusko, 2015; Cattell & Horn, 1966; Fagan & Holland, 2009; Freedle & Kostin, 1997; Hoffman, 1962; Johnston, 1984; Malda et al., 2010; Spearman, 1927; Sternberg, 1981). Although contemporary researchers still grapple with these issues, considerable progress has been made over the years on some of these issues.

Early tests of intelligence such as the Binet-Simon, Wechsler, and Army Alpha & Beta tests were not built from any particular theory of intelligence but rather test items were simply constructed to reflect those typical in educational settings or arbitrarily deemed representative of intelligence by the test designer (Ittenbach et al., 1996; Spearman, 1927). This was perhaps a natural consequence of the lack of agreed-upon intelligence theory to guide assessment construction. With the fairly recent advancement in theoretical understanding, with special emphasis on the Cattell-Horn-Carroll (CHC) and Planning, Attention, Successive, and Simultaneous (PASS) theories of intelligence, it has been possible to either retrofit extant intelligence tests to conform to modern theory or use theory to guide test development (see Flanagan & Harrison, 2012). Alongside developments in aligning tests to theories of intelligence, much work has been done to make the tests and their items more precise in their measurement with much of this work centered around removing construct-irrelevant variance from assessments.

Construct-irrelevant variance, as its name suggests, is any extraneous variance captured by an assessment that is not due to the construct of interest but due to the assessment capturing additional constructs or due to the assessment's methodology (Messick, 1995). As opposed to random error, construct-irrelevant variance systematically impacts assessment scores for individuals or groups of individuals and consequently compromises the construct validity of

assessments and hence the ability to draw inferences from assessment scores (Binning & Barrett, 1989; Haladyna & Downing, 2004; Messick, 1995). Sources of construct-irrelevant variance in intelligence assessments include indeterminable items (e.g., items with no correct answer or items with multiple correct answers with no way to distinguish among them; Freedle, 2003; Hoffman, 1962), cultural content (Cattell, 1963; Freedle, 2003; Freedle & Kostin, 1997; Helms-Lorenz, van de Vijver, & Poortinga, 2003; Malda et al., 2010; Scherbaum et al., 2015) and content not relevant to the intended domain (Helms-Lorenz et al., 2003; Malda et al., 2010; Scherbaum et al., 2015).

Cultural content is any content on a test that reflects aspects of a particular culture, which can include broader cultural differences (e.g., Eastern vs. Western cultures), content tied to a specific culture (e.g., an item featuring Dutch sayings), or even just the language of the test (Helms-Lorenz et al., 2003; van de Vijver, 1997). Test content that reflects different cultures or different cultural elements leads to challenges in creating pure measures of cognitive ability, since this specific cultural knowledge may not be evenly distributed across the population. It is also not uncommon for the cognitive complexity of an intelligence assessment to be confounded with cultural content (Helms-Lorenz et al., 2003). For example, Malda et al., (2010) conducted a study which examined racial group score differences on mathematics items that involved sports. Two identical sets of items were created except that one group of items were embedded in a rugby context whereas the second set of items were embedded in a soccer context. Despite the actual arithmetic being identical across both sets, racial-group differences were found due to Blacks and Whites in South Africa having different exposure to each sport. In a different study, Freedle and Kostin (1997) reported that Blacks performed worse than Whites on analogy items featuring *easier* vocabulary words due to each racial group using those words differently. Racial

group score differences were reduced considerably on items involving more difficult vocabulary due to neither racial group having an appreciable degree of familiarity with those words.

Content not relevant to the intended domain impacts item quality in a similar manner as cultural content. As its name suggests, content not relevant to the intended domain is any content in a test or item that impacts performance on the test, taps into a construct or source of knowledge that individuals vary on, but is content that reflects something other than what the test or item purports to measure. For example, consider the mathematics question of “A golfer enters a hole at even but birdies the next two holes. What score does the golfer now have?” Here, a very simple arithmetic problem is completely obscured by golf-related terminology. If this item were included on a test of golf knowledge, the item’s content would be relevant to the intended domain. However, on a test of mathematics, the requisite golf-terminology acts as a considerable contaminant. This phenomenon can extend to vocabulary, generally. If a test item that is not designed to assess vocabulary contains particularly unusual or difficult vocabulary, then that vocabulary can be viewed as content not relevant to the intended domain.

Both cultural content and content not relevant to the intended domain overlap and can be grouped under the label of construct-irrelevant variance due to differential familiarity, which asserts that observed scores are inflated for some individuals or groups of individuals due to them having knowledge of or familiarity with the test content prior to examination (Agnello et al., 2015; Fagan & Holland, 2007; Ortiz, Ochoa, & Dynda, 2012; Scherbaum et al., 2012). Furthermore, this differential familiarity with test content is incidental to the intended construct of measurement. Construct-irrelevance due to differential familiarity with the test or item content is always a source of contamination, but the issue becomes much more challenging when that differential familiarity or knowledge maps onto demographic faultlines.

While indeterminable items are more easily remedied, construct-irrelevant variance due to differential familiarity has been a more insidious problem, particularly on tests of fluid intelligence (Gf) where the novelty of item content and cognitive demands is critical to the effectiveness of assessment (Agnello et al., 2015; Horn & Cattell, 1966; Fagan & Holland, 2007; Sternberg, 1981). Driven by a need to obtain better and less contaminated measures of Gf, especially across populations varying on socioeconomic, ethnic, cultural, and linguistic-proficiency characteristics, several techniques have been developed in an ongoing effort to eliminate construct-irrelevant variance due to differential familiarity from intelligence assessments. These techniques include the reduction or elimination of verbal content from assessments (Goldstein, Scherbaum, & Yusko, 2010; Hossiep, Turck, & Hasella, 1999; Naglieri, 2005), increased utilization of graphical stimuli in assessment items (Goldstein et al., 2010; Naglieri, 2005; Raven, 2000), creating culturally/linguistically parallel test forms (Malda et al., 2010), incorporating familiarization with test-content into the testing situation (i.e., dynamic testing; Resing, Tunteler, de Jong, & Bosma, 2009), and creating (novel) content for which no group is believed to have prior familiarity (Fagan & Holland, 2009; Goldstein et al., 2010; Sternberg, 1981, 2006). Apart from novel, graphical stimuli, this last technique of creating novel content has primarily consisted of utilizing pseudowords which are strings of letters that resemble a real word in a given language (or possibly languages) but which have no semantic or lexical representation (e.g., ‘contramponist’; Gathercole, 1995; Storkel, Armbrüster, & Hogan, 2006).

Replacing verbal item content with pseudowords has been found to reduce test score variation attributable to differential familiarity (Fagan & Holland, 2009; Sternberg, 2006). Importantly, pseudowords provide a method to reduce differential familiarity effects while

maintaining some level of verbal content. Considering that much of day-to-day life is mediated through language, verbal reasoning invokes unique cognitive mechanisms (e.g., Baddeley, 2012) and neurological pathways (Langdon & Warrington, 2000), and is psychometrically distinct from other forms of reasoning (e.g., Carroll, 1993), maintenance of some level of verbal content is paramount. Compared to alternative techniques used to reduce prior familiarity effects (e.g., culturally-parallel test forms, dynamic testing; Malda et al., 2010; Resing et al, 2009), pseudowords may offer an operational advantage in that cultural or linguistic parallel test forms need not be created and test takers do not need to be trained on content during the test which imposes additional testing time or administration burdens.

As mentioned earlier, prior familiarity of test content is often unevenly distributed across a population leading tests and their items to favor one group of test-takers over another. Evidence of the differential familiarity of domain-irrelevant content having an impact on group scores has appeared when considering the race/ethnicity (Freedle, 2003; Freedle & Kostin, 1997; Roth, Bevier, Bobko, Switzer III, & Tyler, 2001; Malda et al, 2010), gender (Loewen, 1988; Rosser, 1989), and linguistic proficiency (Abedi, 2010; Abedi, Lord, & Plummer, 1997; Kobrin, Sathy, & Shaw, 2007; Sireci, Han, & Wells, 2008; Wolf, Herman, & Dietel, 2010) of test-takers. Unfortunately, ameliorating the effects of domain-irrelevant content while maintaining the verbal content of a cognitive ability assessment has proved difficult. The issue becomes complicated further when attempting to develop tests that assesses verbal Gf across a set of test-takers varied in linguistic backgrounds and proficiencies – an area of increasing need during a time of rapid immigration and globalization. Thus, alongside increasing globalizing comes an increasing need for test developers to be able to create cognitive ability instruments that are flexible across

diverse groups of test-takers while ensuring that the instrument itself measures the construct(s) of interest while minimizing contamination to the greatest extent possible.

The two examples of differential familiarity provided earlier (i.e., cultural content and content not relevant to the intended domain) can be thought of, at their root, as vocabulary issues. In instances where vocabulary is not the main construct of interest as well as instances where it is known that vocabulary can differ considerably across the intended population of assessment, replacing critical words with pseudowords is an effective way to assess some sort of verbal ability in a way that minimizes variance due to differential familiarity of item content. Naturally, high or irrelevant vocabulary demands are a contaminant particularly when trying to assess cognitive ability across test takers that vary in linguistic proficiency. Indeed, early studies examining the impact of bilingualism on cognitive ability routinely showed bilinguals performed much more poorly compared to monolinguals on cognitive ability assessments to such an extent that being bilingual was considered a disadvantage (see Grosjean, 1989; Peal & Lambert, 1962). Following the realization that these studies did not match monolingual and bilingual groups on socioeconomic status and that the cognitive ability assessments used often contained high verbal loads that penalized non-native English speakers, researchers have since corrected these issues and demonstrated that there is no general cognitive ability penalty for bilingualism and perhaps some advantages (Bialystock, Craik, & Luk, 2012; Hakuta, Ferdman, & Diaz, 1987; Peal & Lambert, 1962). However, a natural consequence of this has been a sharp reduction in studies using Gf assessments with any verbal content when attempting to understand cognitive differences between monolinguals and bilinguals. While this is certainly understandable, it does limit opportunities to identify where verbal-content pitfalls may lie and how these issues might

inform test design to create verbal assessments that work effectively across multiple linguistic backgrounds.

While it is perhaps easy to replace verbal with non-verbal cognitive ability or Gf assessments in laboratory research when researching different linguistic subgroups, it is more difficult in other areas such as educational or employment testing. In these contexts, it is often necessary to understand the linguistic proficiency/verbal abilities of test-takers and as a result it is not uncommon for bilingual test-takers for whom English is their non-dominant language to underperform compared to test-takers for whom English is their dominant language (Kena, et al., 2016; Kobrin, Sathy, & Shaw, 2007). Apart from score differences between linguistic groups, oftentimes the assessments yield differential validities for English-dominant and English-non-dominant test takers (Haladyna & Downing, 2004; Olmedo, 1981; Sireci et al., 2008; Wolf et al., 2010).

This shift away from using cognitive ability tests with verbal content perhaps explains why pseudowords have not received closer scrutiny in cognitive ability testing contexts. While pseudowords have been found to reduce score differences between groups, beyond reducing contamination due to prior familiarity, it is not entirely clear how they do so, nor is it entirely clear which properties of pseudowords may be associated with group differences. However, judging by the literature in other fields – namely psycholinguistics and cognitive psychology, there is reason to suspect that properties such as the length of a pseudoword or how much a pseudoword resembles a typical word in a language (henceforth referred to as ‘wordlikeness’) can impact item properties, but may not do so equally across demographic subgroups. Thus, the central goals of the present research were 1) to explore whether altering pseudoword properties grants test developers more control over item difficulty and 2) to understand if and how different

balances of pseudoword properties lead to differential performance across linguistic subgroups – namely individuals for whom English is their dominant language compared to individuals for whom English is their non-dominant language.

Research from cognitive psychology and psycholinguistics indicates that variations in word or pseudoword properties such as length and wordlikeness can place differing demands upon working memory (WM) which offers a pathway for theoretical and empirical predictions of how pseudoword properties may impact verbal Gf item properties (e.g., item difficulty, item discrimination). For example, research has found that longer words and pseudowords can be more difficult to recall in short-term memory (STM) tasks (Baddeley, Thomson, & Buchanan, 1975; Ellis & Beaton, 1993; La Pointe & Engle, 1990; Papagno & Villar, 1992). Similarly, less wordlike pseudowords (i.e., pseudowords low in wordlikeness) are more difficult to reproduce and recall from short-term or working memory (Ellis & Beaton, 1993, Gathercole, Frankish, Pickering, & Peaker, 1999; McNulty, 1965; Rodgers, 1969). Thus, two pseudowords otherwise appearing equally novel do not necessarily impose equal burdens upon cognitive processes. Nevertheless, the literature involving cognitive ability tests that feature pseudowords (e.g., Fagan & Holland, 2007; Scherbaum et al., 2015; Singer, Lichtenberger, Kaufman, Kaufman, & Kaufman, 2012; Sternberg, Ferrari, Clinkenbeard, & Grigorenko, 1996) has not yet examined pseudowords and their properties as closely.

Given WM's close relationship with cognitive ability generally and Gf in particular (Ackerman, Beier, & Boyle, 2005; Buehner, Krumm, Ziegler, & Pluecken, 2006; Kyllonen & Christal, 1990; Oberauer, Schulze, Wilhelm, & Süß, 2005), it is hypothesized that these findings in the pseudoword/WM literature are applicable to the verbal Gf assessment domain and that by altering pseudoword properties to place greater demand on WM, it may be possible to increase

the difficulty of Gf items featuring pseudowords. Put differently, if longer and less wordlike pseudowords are more difficult to maintain, reproduce, and recall in WM study paradigms and WM is tightly linked with Gf, then Gf items featuring longer and less wordlike pseudowords should be more difficult than items with shorter and more wordlike pseudowords. That being said, while pseudowords have proved to be an effective technique for reducing contamination due to prior-familiarity on cognitive ability tests, it should not be assumed that this reduction will remain constant when pseudoword properties are altered. Boundary conditions may exist in the verbal Gf assessment context such that when pseudowords are embedded in alternative verbal Gf items, the alteration of pseudoword properties may interact with the linguistic proficiency of test-takers, which in the current research can be broadly grouped as English-dominant versus English-non-dominant language groups. It remains to be seen if the alteration of these properties leads to a further reduction in or an increase in construct-irrelevant variance.

Such a claim that group differences may re-emerge is not necessarily without precedent either. The variance associated with prior familiarity that pseudowords are designed to minimize may nevertheless manifest as there is evidence that the learning and memorization of words and pseudowords invokes long-term memory (LTM; Cheung, 1996; Gathercole et al., 1999). Baddeley (2003) states that as we age the relationship between phonological working memory and vocabulary becomes reciprocal. Our phonological WM is critical for early vocabulary development but as our vocabulary increases, it then becomes easier to learn new words. Put differently, the learning of new words is aided by our vocabulary which itself is more representative of crystallized intelligence (Gc) rather than the intended construct of Gf. Similarly, memory span has been found to be greater for words that appear frequently in a language compared to words that appear less frequently (Hulme, Maughan, & Brown, 1991)

suggesting that one's experience with a language as well as the familiarity of words influences in the ease with which they can be maintained in WM. Because the familiarity of words impacts their likelihood of recall, despite their novelty, pseudowords' wordlikeness may allow for the re-entry of differential familiarity into item construction. Furthermore, these studies frequently controlled for the linguistic backgrounds of participants, thus it remains possible that the impact of these differences in word properties manifest more strongly across linguistically diverse populations.

The manipulation of pseudoword properties may offer a pathway to incorporate theory from cognitive psychology and psycholinguistics into item construction and to create assessments that cover a greater portion of the verbal Gf construct while keeping construct-irrelevant variance due to prior familiarity at a minimum. Research has demonstrated that subtle changes to cognitive ability items can produce notable changes in item properties (including difficulty), even if the changes are believed to be isomorphic (Koch, McCloy, Trippe, & Paullin, 2012). Thus, the proposed research explores whether altering pseudoword properties impacts item difficulty as well as understanding if and how different balances of pseudoword properties may lead to differential performance across linguistic groups.

In the present research, four pseudoword properties are considered. The first is the length of pseudowords in terms of number of syllables. The next three properties of pseudowords are different measures of wordlikeness: lexical neighborhood density, lexical neighbor frequency, and sequence probability. Lexical neighborhood density and lexical neighbor frequency are lexically-based measures of wordlikeness. Lexical neighborhood density refers to the number of real words in a language that can be derived from the addition, subtraction, or substitution of one letter or phoneme in a word or pseudoword (Storkel et al., 2006). For instance, the word 'cat'

would have the lexical neighbors of ‘cart’, ‘at’, and ‘cut’ when making a letter addition, subtraction, and substitution, respectively. Research has shown that the greater a word’s or pseudoword’s neighborhood density, the easier it is to recall (Storkel et al., 2006; Thorn & Frankish, 2005). Lexical neighbor frequency refers to how often the individual words that comprise the neighborhood appear in common language usage (Vitevitch, Storkel, Francisco, Evans, & Goldstein, 2014).² Put differently, a (pseudo)word with five infrequent neighbors may behave differently than a (pseudo)word with five frequent neighbors (Vitevitch et al., 2014). The third wordlikeness variable of sequence probability is a sub-lexical, rather than lexical attribute of (pseudo)words and can be further decomposed into two forms: orthographic and phonotactic. Orthographic probability refers to the frequency of written letter and letter combinations in words in a specific language while phonotactic probability refers to the frequency of phoneme and phoneme combinations in words in a specific language (Bailey & Hahn, 2001; Storkel et al., 2006). Individuals more easily pronounce and recall words with high sequence probabilities (i.e., more wordlike) compared to those with low probabilities (Gathercole, 1995; Gathercole et al., 1999; Thorn & Frankish, 2005). Ultimately, and discussed later in the paper, the current research manipulates pseudoword length and lexical neighborhood density while controlling pseudoword lexical neighbor frequency and orthographic probability.

Given that individuals are better able to remember pseudowords the more they resemble common native language words at either phonotactic or lexical levels, and that bilinguals generally feature smaller vocabularies and display disadvantages in word recall in their non-

² ‘Common language usage’ for lexical neighborhood frequency was defined by the MCWord developers as how often a word appears in “1,000,000 presentations of text” with those presentations of text consisting of representative text corpora. See the CELEX information page for more detail: <https://catalog.ldc.upenn.edu/ldc96114>.

dominant language (Izawa, 1993; Oller & Eilers, 2002; Thorn & Gathercole, 2001; Thorn, Gathercole, & Frankish, 2002), there is reason to suspect that items involving pseudowords may be differentially difficult across linguistic sub-groups (e.g., monolinguals vs. bilinguals). Simply put, the less pseudowords characterize common vocabulary the more difficult cognitive ability items are expected to be. Further, the more pseudoword properties differentially draw support from LTM, the more those pseudowords are expected to function differently across groups with differing amounts of relevant information available in LTM, in this case English-dominant and English-non-dominant test-takers.

The current research consisted of a study in which verbal fluid reasoning (Gf) items were constructed that featured pseudowords. The study examined how changes in pseudoword and word properties influenced item difficulty and group scores. Specifically, the study examined how changes in pseudoword length and wordlikeness (i.e., how much a pseudoword resembled a typical word in English) on verbal Gf items influenced the item difficulty and influenced performance for individuals who reported that English was their dominant language compared to individuals who reported that English was not their dominant language.³

³ Described in greater detail in Chapter 5, the current state of bilingualism literature has shifted towards understanding a bilingual's linguistic proficiency in terms of dominant vs. non-dominant languages. This has afforded a greater precision with understanding an individual's linguistic proficiencies compared to other linguistic groupings such as monolingual vs. bilingual or first vs. second language. The current research thus compared individuals for whom English was their dominant language to individuals for whom English was not their dominant language.

Chapter 2: Modern Cognitive Ability Tests

Tests of mental abilities date back to the mid nineteenth century beginning with the work of Francis Galton and Alfred Binet (Benjamin, 2000; Ittenbach et al., 1996; Jensen, 2002) while an understanding of the structure of intelligence dates back to the early twentieth century (Spearman, 1927). As described in the previous chapter, several definitional and operational challenges in contemporary intelligence research have been present since psychology's earliest attempts at studying intelligence. For example, in his 1927 book, Spearman describes difficulties with defining intelligence, criticizes extant intelligence tests for lacking theoretical foundations, and also posits that some results, particularly the emergence of specialized or narrow cognitive abilities, are "due to past experience rather than native aptitude" (p. 242). This blend of challenges at both the construct and measurement levels fits quite well in contemporary discourse on intelligence.

Fortunately, due to modern theoretical, statistical, and technological advances, there has been a recent flurry of interest, refinements, and breakthroughs in the intelligence field (Ackerman, Beier, & Boyle, 2005; Chen & Gardner, 2012; Fagan, 2000; Fagan & Holland, 2009; Flanagan, Alfonso, & Ortiz, 2012; Goldstein et al., 2010; Helms-Lorenz et al., 2003; Jung & Haier, 2007; Lang, Kersting, Hülshager, & Lang, 2010; Malda et al., 2010; Naglieri, Das, & Goldstein, 2012; Sabet, Scherbaum, & Goldstein, 2013; Ortiz, Ochoa, & Dynda, 2012; Scherbaum, Goldstein, Yusko, Ryan, & Hanges, 2012; Schneider & McGrew, 2012; Sternberg, 2004, 2006, 2012; van der Maas et al., 2006; van der Maas et al., 2014; Van Iddekinge & Ployhart, 2009). While a full review of this literature is beyond the scope of the present investigation, within the arena of contemporary cognitive ability assessment, two over-arching trends have been present, 1) greater incorporation of contemporary intelligence and cognitive

theory into test construction (Schneider & Newman, 2015; Wee et al., 2014) and 2) more effective measurement of the intelligence construct (Agnello et al., 2015; Brouwers & van de Vijver, 2015; Scherbaum et al., 2015).

Beginning with the former trend, the most prominent example has been the alignment of cognitive ability tests with the Cattell-Horn-Carroll (CHC) theory of intelligence. The CHC theory of intelligence is rooted in the psychometric perspective and blends Carroll's (1993) three-stratum model of intelligence with Cattell and Horn's research bifurcating intelligence into fluid intelligence (Gf) and crystallized intelligence (Gc) and their respective sub-components (Horn & Cattell, 1966). The CHC theory generally posits that there are three hierarchically-arranged strata of intelligence with *g* as the third (top) stratum, eight broad cognitive abilities including both Gf and Gc within stratum two, and over sixty narrow cognitive abilities comprising stratum one (Carroll, 1993; McGrew, 2009; Schneider & Newman, 2015). The CHC theory is currently the most prominent theory informing the construction of modern psychometric cognitive ability tests (see Flanagan & Harrison, 2012; Keith & Reynolds, 2010; McGrew & Wendling, 2010) and assessment methodology such as the Cross-Battery Assessment (XBA; Flanagan et al., 2012). Not limited to the CHC theory of intelligence, other theories of intelligence such as the Planning-Attention-Successive-Simultaneous (PASS) theory of intelligence are also being used to guide test development as well (Matthews, Riccio, & Davis, 2012; Naglieri & Das, 1990; Naglieri & Otero, 2012; Singer, Lichtenberger, Kaufman, Kaufman, & Kaufman, 2012).

The second major trend in cognitive ability test development has been improved measurement of the intended construct or constructs. Improving construct measurement has tended to consist of either developing more inclusive measures/test batteries with explicit links to

intelligence theory (i.e., improving their construct relevance) or developing techniques for reducing sources of construct-irrelevant variance in assessments.

As previously discussed, by advancing intelligence theory, assessments have benefitted by offering better coverage of relevant intelligence constructs. With the elucidation of both broad and narrow abilities, test developers have been able to develop items that assess the relevant mental abilities and processes to a greater degree. This has allowed cognitive ability tests to increase coverage of the latent construct(s) in terms of both breadth and depth. The XBA (Flanagan et al., 2012) serves as a representative example of this trend. The XBA represents a method rather than a specific instrument and is concerned with assessing cognitive and academic abilities alongside neuropsychological processes and provides guidance on ways to measure a wider range or greater depth of constructs while maintaining utmost precision with regard to the criteria of interest (Flanagan et al., 2012). Through extensive confirmatory factor-analytic work, researchers have been able to identify measures which most effectively assess broad or narrow CHC abilities and then use this information to inform decisions about the assessments to be used to evaluate a given individual or sample of individuals (Flanagan et al., 2012).

The second major theme of improved measurement is the minimization of construct-irrelevant variance in an effort to obtain better measures of the intended construct(s). One of the more prominent sources of construct-irrelevant variance has been the differential familiarity of test-takers with test content (Johnston, 1984; van de Vijver & Poortinga, 1997) and often emerges as construct-irrelevant cultural content or domain-specific content (Fagan & Holland, 2009; Malda et al., 2010; van de Vijver & Poortinga, 1997). Differential familiarity of test content can be extremely subtle to detect (Freedle & Kostin, 1997), and related to Spearman's (1927) criticism, manifests as score variation attributable to experience rather than native ability

(Cattell, 1963; van de Vijver & Tanzer, 2004) which is particularly damaging for tests of Gf. Differential familiarity of test content, when operating as a form of construct-irrelevant variance, reduces the precision of assessments as well as our ability to draw inferences from them (Binning & Barrett, 1989). Furthermore, when differential familiarity of irrelevant test content is unevenly distributed across groups of test takers, the result is assessments that appear to disadvantage those groups lacking familiarity (Fagan & Holland, 2002, 2007, 2009; Freedle & Kostin, 1997; Malda et al., 2010).

Verbal content is the primary medium through which construct-irrelevant variance in the form of differential familiarity is introduced to tests (Fagan & Holland, 2009; Johnston, 1984; Malda et al., 2010). While verbal content is often essential for tests of Gc, it frequently operates as a source of contamination on tests of Gf (Naglieri & Otero, 2012). Cultural, linguistic, gender, and socioeconomic groups differ in access and exposure to vocabulary (e.g., Hoff, 2006; Naglieri & Otero, 2012; Rosser, 1989) and for a test designed to measure Gf rather than Gc, verbal content may unintentionally create a measure of both.

Numerous techniques have proliferated to negotiate the effective measurement of cognitive ability while reducing contamination due to differential familiarity of test content with the most extreme technique being the elimination of verbal content entirely from the test. Indeed, there are currently a number of tests that do not contain any verbal content including the Raven's Progressive Matrices (RPM; Raven, 2000), the Universal Nonverbal Intelligence Test (UNIT; McCallum & Bracken, 2012), and the Bochumer Matrizen-test (BOMAT; Hossiep et al., 1999; Jaeggi, Buschkuhl, Jonides, & Perrig, 2008). These tests have been able to assess Gf while reducing contamination due to differential familiarity by featuring items utilizing graphical stimuli. The graphical stimuli present on these tests are assumed to be novel for all test takers

thus equating familiarity with item content across subgroups, though this assertion is debatable (e.g., van de Vijver & Poortinga, 1997; van de Vijver & Tanzer, 2004). Other assessments, while not eliminating verbal content entirely have reduced it by replacing critical words with graphical stimuli (Agnello et al., 2015; Goldstein et al., 2010).

A second approach towards effectively managing the effects of differential familiarity involves creating culture-specific parallel test forms. Oftentimes, it is incorrect to assume equivalence of psychological constructs across cultures, let alone score equivalence (Fischer, 1969; Sternberg, 2004; van de Vijver & Poortinga, 1997; van de Vijver & Tanzer, 2004). Rather than creating novel content with which no group is familiar, this approach involves creating tests that measure the same construct but are designed to reflect respective cultural norms, language, and idiosyncrasies and has been an effective technique for assessing personality, emotion, and cognitive ability constructs (Brouwers & van de Vijver, 2015; Malda et al., 2010; van de Vijver & Poortinga, 2004). Familiarity of item content is calibrated and equated for the two forms so that test scores are comparable across groups (Malda et al., 2010). By designing two assessments with matched levels of familiarity, group differences in scores are reduced.

Dynamic testing represents a third approach for managing the differential familiarity of test content (Resing & Elliot, 2011; Resing et al., 2009). In response to the criticism of traditional tests assessing previously learned material more effectively than the ability to learn, dynamic testing attempts to examine a test-takers' improvement following training during the test sessions (Resing et al., 2009). Test-takers are given feedback and hints by a trained observer via a graduated-prompt technique during testing. This is done in an attempt to see how this new information is incorporated into existing knowledge bases and to ensure that test-takers do not rely on trial-and-error strategies or ineffective strategies. Dynamic testing has been found to

produce score gains for all test-takers and to reduce score variation attributable to differential familiarity (Resing et al., 2009).

A fourth approach for reducing the contamination caused by differential familiarity with verbal test content involves creating novel verbal stimuli commonly known as pseudowords (alternatively referred to as ‘nonwords’ or ‘nonsense words’ in the literature). Pseudowords are sequences of letters arranged in accordance with orthographic and phonotactic conventions in a given language but for which there are no proper lexical or semantic associations (Storkel, Armbrüster, & Hogan, 2006). In other words, pseudowords look and sound like a real word but are not and have no meaning. Consider the word ‘doppelate’ (Gathercole, 1995) as an example of a pseudoword that sounds like a plausible word in English but in actuality is not. By replacing critical verbal item content with pseudowords, prior familiarity is assumed to be equalized at zero across groups and research has found that doing so creates a test more equipped to assess Gf than Gc and reduces score variation stemming from prior familiarity quite well (Fagan & Holland, 2009; Freedle, 2003).

While all of the above methods have demonstrated effectiveness with reducing contamination due to differential familiarity and show strong potential for assessing Gf, each possesses limitations. By eliminating verbal content from tests and using graphical stimuli only, the test is no longer equipped to assess verbal skills. This can be problematic for several reasons: 1) interaction with and usage of language constitutes a large part of daily mental operations and daily life, 2) basic cognitive structures such as working memory appear to have different sub-structures for auditory/linguistic and visual/spatial inputs (Baddeley, 2009), 3) verbal reasoning and spatial reasoning differentially invoke neurological pathways and hemispheric activity in the brain (Langdon & Warrington, 2000), 4) depending on the criteria, verbal ability may be a much

more important predictor than other cognitive abilities (Lang et al., 2010), and 5) psychometrically, verbal ability emerges as separate from mathematical or visual/spatial abilities (Carroll, 1993). Thus, the elimination of verbal content prohibits the assessment of these verbally-based cognitive abilities and creates situations where Gf may be operationalized in a way that disregards verbal reasoning entirely (e.g., Colom, Rebollo, Palacios, Juan-Espinosa, Kyllonen, 2004), possibly contributing to construct underrepresentation in tests or test batteries (Messick, 1995).

Culture-specific parallel test form development involves rigorous pre-work concerned with creating strata of familiarity for verbal test content which may not always be practical, especially when testing populations with multiple distinct cultures or mixed-culture individuals. Furthermore, this approach necessitates that the culture of test-takers be known prior to testing which is problematic in educational and occupational selection settings. Finally, by having trained individuals provide feedback and hints to test-takers during testing, dynamic testing features administrative and test-duration costs that may be unduly burdensome in typical applied testing situations (e.g., educational and employment testing contexts).

Pseudowords have been used in extant verbal reasoning items whereby a description of a pseudoword is embedded in text describing the pseudoword and then asking test-takers to select which real-life object the pseudoword most represents (Fagan & Holland, 2009; Sternberg, 2006). For instance:

After brewing coffee, she poured some into her vemp.

Vemp most closely means:

- A. Mouth
- B. Hair
- C. Cup
- D. Sink

Comparatively, pseudowords ostensibly involve fewer drawbacks in that they allow for the assessment of verbal ability in a way that minimizes differential familiarity, they do not create a situation where the cultural/linguistic composition of test-takers needs to be known before testing, and they do not impose administrative or test-duration costs as is the case in dynamic testing. Furthermore, pseudowords are compatible with other approaches for reducing differential familiarity in that they can be easily incorporated into culture-specific parallel test and dynamic testing paradigms. In other words, real words can be replaced with pseudowords in extant verbal Gf items allowing researchers to use a known verbal Gf item format as a means to better understand pseudowords. Despite the purported benefits of pseudowords, there is a general drawback in that little is known about them in cognitive ability testing contexts other than that they reduce score variation attributable to differential familiarity. Similarly, the lack of exploration into pseudoword properties in cognitive ability testing contexts suggests that there has been an underlying assumption that all pseudowords are created equal. While their properties have been studied more extensively in cognitive psychology and psycholinguistics (e.g., Cheung, 1996; Gathercole, 1995; Hulme, Maughan, & Brown, 1991; Kaushanskaya & Marian, 2009; Storkel et al., 2006), it is unclear what these properties imply for testing situations, their ramifications for item development, and whether or not boundary conditions exist that limit the effectiveness of the technique. These issues are considered in detail in the next chapter.

Chapter 3: On Fluid Intelligence, Working Memory, and Pseudowords

Neither concerns regarding the differential familiarity of verbal stimuli nor the pseudoword technique are new, as each possesses a history that traces back to some of the earliest experimental psychology research. Early incarnations of pseudowords were developed by Herman Ebbinghaus in his groundbreaking research on human memory in 1885. Ebbinghaus created consonant-vowel-consonant (CVC) nonsense syllables in an effort to develop lists of novel stimuli to be memorized (Benjamin, 2006). The fact that these syllables were novel was paramount as it was believed that the rate of memorization would differ between familiar and unfamiliar stimuli (Benjamin, 2006). It has been well known that the meaningfulness of otherwise novel nonsense syllables impacts their rate of memorization for nearly a century (Glaze, 1928; Hull, 1933). In addition to their lengthy history in cognitive psychology research, pseudowords have been a useful technique in psycholinguistic research for decades. A prominent example is the work by Berko (1958) which examined the morphological knowledge of children by asking them to manipulate pseudowords into alternative forms (e.g., more than one ‘wug’ becomes ‘wugs’). As was the case in memory research, pseudowords were used due to their lack of familiarity, thus allowing researchers to examine whether or not children had learned the rules guiding language rather than just memorized alternate forms of real words.

Spanning the previous few decades, pseudowords have been increasingly incorporated into fluid intelligence (Gf) assessment items and employed extensively in studies assessing properties of short-term memory (STM) and working memory (WM). Despite these trends, there has been no work examining how pseudowords and shifts in their properties behave in a verbal Gf assessment context. However, before a full discussion of pseudowords and their interplay

with cognitive systems/abilities, brief descriptions of Gf, STM, WM, and their interdependence are warranted.

Fluid Intelligence and the Connection to Working Memory

To understand the use of pseudowords in cognitive ability testing, it is important to more formally define Gf and WM (given the close connection to Gf and the use of pseudowords in the WM research). Fluid intelligence refers to a general, domain-free reasoning ability that is less dependent upon prior learning or personal experiences, and is thought to be engaged when reasoning with either novel content or familiar content in novel contexts (Cattell, 1943; Horn & Cattell, 1963; Horn & Blankson, 2013). Identifying relationships and patterns (i.e., inductive reasoning), general sequential reasoning (i.e., deductive reasoning), drawing inferences, and extracting implications are a few of the key products of Gf (Cattell, 1943; Horn & Blankson, 2012; Schneider & McGrew, 2012). As a point of contrast, crystallized intelligence (Gc) is believed to be domain-specific knowledge resulting from experiential, educational, and acculturation processes and influences with reasoning consisting of the application of this obtained and known knowledge to solve problems (Cattell, 1943; Horn & Cattell, 1963; Horn & Blankson, 2012; Schneider & McGrew, 2012).

Fluid intelligence is theoretically and empirically a predictor of Gc (Buehner, Krumm, Ziegler, Pluecken, 2006; Cattell, 1943) and from a psychometric perspective, Gf is the second-order ability that many suggest is most closely related to *g* (Buehner et al., 2006; Carroll, 1993; Schneider & McGrew, 2012) with some researchers suggesting that psychometric *g* and Gf are extremely similar or identical constructs (Cattell, 1987; Floyd, Evans, & McGrew, 2003; Gustaffson, 1984). Fluid intelligence has also been identified as a particularly strong predictor of performance across academic (Gustaffson & Balke, 1993; Kuncel, Hezlett, & Ones, 2004) and

employment (Drasgow, 2003; Goldstein et al., 2012; Schmidt, 2002; Schmidt & Hunter, 1998) contexts, remains a popular selection instrument (Goldstein et al., 2012), and predicts both social well-being and mental health (Jensen, 2002). Given that novelty is essential for the assessment of Gf, many of the efforts around reducing contamination due to prior familiarity have centered on assessments of Gf (Fagan & Holland, 2009; Goldstein et al., 2012; Jaeggi et al., 2008; Malda et al., 2010; McCallum & Bracken, 2012; Raven, 2000).

While Gf and *g* are believed to be isomorphic by some, the same can be said of Gf and WM. Similarities appear in definitions of the two constructs with Schneider & McGrew (2012) defining Gf as “the deliberate but flexible control of attention to solve novel, “on-the-spot” problems” (p.111) and with earlier conceptualizations of WM consisting of the focus of executive attention applied to currently activated and accessible LTM contents or stored STM contents (Conway, Cowan, Bunting, Theriault, & Minkoff, 2002; Cowan, 1993; Engle, 2002). More compelling is the empirical literature beginning with the work of Kyllonen and Christal (1990) whose research suggested that reasoning is ‘little more than working memory capacity’ (p. 1) and reported that the two were correlated between .8 and .9. Stauffer, Ree, and Carretta (1996) similarly report a correlation of $r = .96$ between reasoning and WM factors. However, it should be noted that for both the Kyllonen and Christal (1990) and Stauffer et al. (1996) studies, general reasoning was the construct of interest and not necessarily Gf. Therefore, precise estimates of the relationship between WM and Gf remain muddled. Ackerman, Beier, and Boyle (2005) conducted a meta-analysis of studies examining the relationship between reasoning and WM and reported that the conclusion that ‘*g* and WM are isomorphic’ was overblown with the relationship between *g* and WM sharing 22.9% of the variance. When looking at various measures of Gf, Ackerman et al., (2005) reach a similar conclusion – Gf and WM are related but

not isomorphic with 19.8% of the variance shared between constructs. However, the Ackerman et al., (2005) meta-analysis data were re-analyzed in two separate studies. By selecting only latent-variable studies that were clearly examining the relationship between Gf and WM capacity and other related constructs (e.g., processing speed, STM, Gc), Kane, Hambrick, and Conway (2005) report that Gf and WM share approximately 50% of their variance. Similarly, Oberauer, Schulze, Wilhelm, and Süß (2005) re-analyzed the Ackerman et al. (2005) data while correcting for study-inclusion, methodological, and theoretical limitations and reported that WM accounts for 72.3% of the variance in *g*. Of note is that the Oberauer et al., (2005) study conceptualizes WM as a predictor of *g* rather than the two as just related, though Gf is not isolated from *g*.

It is theoretically congruent to view WM as a predictor of Gf as the former essentially represents some of the hardware used in novel problem solving (Baddeley 2012; Oberauer, Süß, Wilhelm, Wittmann, 2008; Süß, Oberauer, Wittmann, Wilhelm, Schulze, 2002), though this is at odds with certain theories of intelligence, namely the CHC). While research has been unclear at times regarding the cognitive ability criterion (e.g., *g* vs. Gf vs. Gc), much contemporary research has been able to identify WM as a strong predictor of Gf. Conway et al., (2002) report a significant path coefficient of .60 between the two constructs. Redick, Unsworth, Kelly, and Engle (2012) reanalyzed the Conway et al. (2002) data and report that WM accounts for 28% of the variance in Gf and in a new empirical study report that WM accounts for 27% of the variance in Gf. Buehner, Krumm, Ziegler, and Pluecken (2006) report that the coordination and the storage and processing components of WM along with sustained attention account for 83% of the variance in Gf and that this relationship holds when controlling for Gc. Other research has shown that WM significantly predicts Gf, Gc, and spatial reasoning (Gv), though the relationships with Gc and Gv become non-significant when Gf is partialled out (Martinez & Colom, 2009). There is

even a burgeoning literature on mental training suggesting that WM can be improved via training and that these improvements lead to better performance on assessments of Gf (Jaeggi, Buschkuhl, Jonides, & Perrig, 2008; Jaeggi, Buschkuhl, Jonides, & Shah, 2011; Jaeggi, Studer-Luethi, Buschkuhl, Su, Jonides, & Perrig, 2010) though these results are not without their shortcomings/criticisms (see Shipstead, Redick, & Engle, 2012).

Recently, research has explored which features of WM are responsible for the relationship between WM and Gf and has suggested that storage capacity (Chuderski, Taradaj, Nęcka, & Smoleń, 2012; Colom, Abad, Quiroga, Shih, & Flores-Mendoza, 2008) and to a lesser extent executive attention operationalized as updating (Colom et al., 2008) are responsible. Specifically, Colom et al. (2008) report that STM plus WM accounts for 70% of the variance in Gf, though they also report that when partialing out STM and processing speed, WM is no longer a significant predictor of Gf; a finding at odds with previous research (Conway et al., 2002; Engle et al., 1999). To summarize, it appears that WM accounts for roughly 30%-80% of the variance in Gf with this amount of variance dependent upon the broadness or narrowness of the WM operationalization and with storage capacity, updating, and coordination being those WM features and functions most responsible.

Another source of evidence for the Gf and WM linkage comes from the neurocognitive literature. Through functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) techniques, neurocognitive researchers have identified the dorsolateral, ventrolateral, anterior, and medial portions of the prefrontal cortex (PFC), the dorsal anterior cingulate cortex (ACC), and lateral and medial parietal lobe regions in the brain as active during reasoning tasks (Braver, Gray, & Burgess, 2007; Duncan et al., 2000; Geake & Hansen, 2005; 2010; Jung & Haier, 2007; Kane & Engle, 2002). Regarding WM, two meta-analyses present

similar findings with dorsolateral, ventrolateral, medial, and anterior PFC, lateral and medial parietal cortex, and ACC brain regions active during WM tasks (Owen, McMillan, Laird, & Bullmore, 2005; Wager & Smith, 2003). While reasoning and WM processes do not feature identical patterns of brain activation, broadly PFC and parietal regions, and specifically Brodmann Areas 6, 7, 8, 9, 10, 32, 40, and 47 are those brain regions engaged during both Gf and WM activities (e.g., Burgess, Gray, Conway, & Braver, 2011; Jung & Haier, 2007). Recent research involving brain lesions has similarly presented the PFC and parietal regions as underpinning executive control and subsequently Gf and WM functioning (Barbey, Colom, Paul, & Grafman, 2014), and that executive control in the form of interference control mediates a large portion of the neural connection between Gf and WM (Burgess, Gray, Conway, & Braver, 2011).

Establishing the theoretical, psychometric, and neurocognitive overlap between WM and Gf constructs greatly informs the following sections and chapters. To be discussed in depth later in this chapter, the literature involving pseudowords has focused on WM but not Gf. Thus to the extent that WM more or less easily accommodates pseudowords as their properties are varied, the more insight can be gleaned on how varying pseudoword properties may impact Gf item properties.

Working Memory Systems

It is necessary to understand contemporary thought around WM to not only attempt to infer how research conducted on WM systems may be reflected in a verbal Gf reasoning context (which is highly dependent upon WM) but to also understand the implications for patterns of cognition between monolingual and bilingual individuals. Just as with the ambiguity between Gf and WM, currently, the literature is somewhat unclear regarding the boundary between STM and WM (Engle, Tuholski, Laughlin, & Conway, 1999). For example, although STM and WM are

believed to be separate both theoretically (e.g., Baddeley, 2012; Kail & Hall, 2001) and empirically (Engle et al., 1999; Kail & Hall, 2001) the two terms are sometimes used to describe the same cognitive construct (compare Baddeley, 2012 and Jonides, Lewis, Nee, Lustig, Berman, & Moore, 2008). In other cases, WM is viewed as a sub-component of STM (Schneider & McGrew, 2012). Despite the confusion, it appears as though the dominant conceptualization is that STM is a sub-component of WM (Baddeley, 2012; Chunderski, 2013) with the demarcation between the two being the difference between simply maintaining information in some active state (i.e., STM) versus actually performing some cognitive operation upon that information (e.g., Baddeley, 2012; Daneman & Carpenter, 1990) or maintaining that information in the face of distracting information via controlled attention (e.g., Cowan, 2005). The current paper treats STM as a sub-system of WM, a view consistent with the perspective of Baddeley (2012) as well as the empirical literature (Engle et al., 1999; Kail & Hall, 2001).

Working memory is broadly viewed as a cognitive system that conjoins LTM, incoming sensory information, and attention in the service of cognition and the production of action (Baddeley, 2012; Cowan, 2005). Operationally, WM is the information that is kept in a fairly activated and retrievable state with this information serving as the material for cognition (Cowan, 2005). Working memory features a small capacity that is limited to approximately four chunks of information and the contents of WM are fairly fragile as they are prone to rapid decay or interference, as well as easily overwritten by new, incoming information (e.g., Jonides et al., 2008). While it may be possible to identify situations where WM is inactive, it generally is a process that is constantly active. For example, reading, maintaining representations of one's surroundings, and complex problem solving all make use of WM to varying degrees (Cowan, 2005; King & Just, 1991; Lewis & Vasishth, 2005). Currently, amidst the multitude of WM

theories (see Baddeley, 2012), two dominant and readily compatible theories of WM exist: WM as a multistore system and WM as a unitary store system.

According to the multistore theory, WM consists of four components: the phonological loop, the visuospatial sketchpad, the central executive, and the episodic buffer (e.g., Baddeley, 2012). The phonological loop and the visuospatial sketchpad are WM sub-systems which store auditory/verbal information and visual/spatial information, respectively (Baddeley, 2012). The central executive is believed to be one's attention and features no capacity for storage beyond the one piece of information in focus. Focusing on information (one discrete piece or an integrated chunk), switching attention, and controlling other sub-systems are just some of the possible functions of the central executive (e.g., Baddeley, 2012; Baddeley & Logie, 1992; Phillips & Hamilton, 2001). The last component is the episodic buffer which is a limited-capacity, multidimensional store. Specifically, the episodic buffer serves to integrate multiple modalities (e.g., visual and auditory) and multiple forms of encoding (e.g., phonological and semantic encoding for auditory information) together as well liaise between the central executive and LTM. As Baddeley (2012) writes, it is the episodic buffer that allows us to think of an ice hockey-playing elephant when prompted – as this involves the central executive's attention on the instruction (whether it be auditory or written) as well as the episodic buffer's extraction and combination of requisite memories into a novel piece of information held in WM. Whereas non-verbal Gf tests tap into the neurological pathways associated with the visuospatial sketchpad effectively, pseudowords tap into the neurological pathways of the phonological loop in a manner that cannot be done with just visual/graphic Gf stimuli nor actual, familiar words (Baddeley, 2012).

The unitary perspective of WM exists without reference to the four components described by Baddeley and colleagues, and views WM as embedded within LTM and functionally as the conjunction of STM plus the focus of controlled attention on incoming as well as retrievable information (Cowan, 2005; Engle et al., 1999; Jonides et al., 2008). Unlike Baddeley's model, unitary perspectives make frequent reference to the concept of 'resting levels of activation' which is a term used to refer to a spectrum of accessibility of mental information (Cowan, 2005). Information with a higher resting level of activation will be more readily accessible – consider recalling your current phone number compared to your phone number growing up – the latter might be retrievable but possibly not as easily as your current phone number. A point of contrast between the two perspectives of WM is that whereas Baddeley's multistore model suggests that WM and LTM are separate systems, unitary store models make less of a distinction between WM and LTM and tend to view WM as activated and attended content that is otherwise LTM (Cowan, 2010; though see Cowan, Rouder, Blume, & Saults, 2012). Long-term memory contents vary in their baseline thresholds of activation with factors such as the frequency of activation (resulting from experience and practice) as well the elapsed time since the most recent activation influencing activation thresholds (e.g., Anderson & Reder, 1999; Lewis & Visishth, 2005). Beyond elaborating WM's situation relative to other cognitive systems, much work has been devoted to understanding the amount of activated material that WM can hold with several sources suggesting three to four chunks of information if rehearsal is not permitted (e.g., Cowan, 2005; Cowan et al., 2012). However, recent work has suggested that focused attention can attend to no more than one chunk of information at a time (Oberauer, 2002) which has resulted in a compromise that suggests while attention can only be focused upon one chunk of information at

a time, WM can hold approximately four chunks in a readily-retrievable state (Cowan et al., 2012; Oberauer, 2002).

The two models of WM are quite compatible (Baddeley, 2012) and critically for the present purposes, both agree that WM is linked to both STM and LTM. Explication of the hardware of the mind in the multistore model elaborates how pseudowords engage the brain in a manner different from graphics or from existing words. By embedding WM within LTM, the unitary store model provides a strong backdrop for interpreting how subtle differences in familiarity of (pseudo)words lead to differences in cognitive performance. By stressing that the majority of contents in LTM are retrievable but at different cognitive costs (i.e., differences in resting levels of activation) the unitary model also dovetails quite well with modern conceptualizations of both cognition and bilingualism (McClelland, 2000; Zhao & Li, 2010). Thus, the two theories of WM appear to have critical application for cognitive processes in a verbal Gf assessment context particularly for test-takers of different linguistic backgrounds and proficiencies. Furthermore, both models are relevant for interpreting results of research utilizing pseudowords.

Pseudowords in the Service of Explicating Cognitive Systems

Given that WM and Gf are tightly linked, it is instructive to look at how pseudowords have been used in WM studies to better understand how pseudowords connect to WM as well as how manipulations to pseudowords may subsequently impact Gf verbal reasoning items. Within the cognitive psychology literature, pseudowords have been involved in clarifying a number of mental capabilities. Pseudowords have been used in assessing the amount of information that can be held in temporary storage in adults (Hulme, Maughan, & Brown, 1991; Papagno & Vallar, 1992) and children (Roodenrys, Hulme, & Brown, 1993). Research has also demonstrated that

the quality of this information is subject to both primacy and recency effects with subjects more likely to repeat the syllables at the beginning and the end of the pseudoword more correctly than those in the middle (Gupta, 2005). The correct repetition of a pseudoword has become a popular test of the phonological loop component of WM (Baddeley, Gathercole, & Papagno, 1998; Gathercole, 1995; Gathercole, Frankish, Pickering, & Peaker, 1999). Furthermore, resulting from the use of pseudowords, the phonological loop has been posited as a cognitive sub-system that is uniquely active in learning new verbal information as opposed to merely storing all incoming verbal information (Baddeley, Papagno, & Vallar, 1988). Pseudowords have also been critical for demonstrating that the phonological loop is active when verbal information is presented visually, not just auditorily (Baddeley et al., 1988).

Pseudoword repetition has been found to be a strong predictor of vocabulary development early in life (Gathercole, 2005; Gathercole & Baddeley, 1990) and the ability to learn new words in one's native language (Gathercole & Baddeley, 1989) and in a second language (Cheung, 1996; Masoura & Gathercole, 1999; Masoura & Gathercole, 2005; Papagno & Vallar, 1995). Pseudoword span (i.e., the number of pseudowords that can be recalled correctly) predicts learning new words better than simple STM span tests (e.g., digit span) and Gf measures for those individuals with smaller vocabularies (though it is not necessarily predictive for those with a larger vocabulary; Cheung, 1996). Likewise, pseudoword repetition is often more predictive of word learning than simpler tests of STM span (Gathercole & Adams, 1994; Gathercole, Willis, Emslie, & Baddeley, 1992). Through pseudowords it has been possible to demonstrate that familiar and novel stimuli differentially activate cognitive sub-systems as well as the importance of subvocalization for reading and rehearsal. Subvocalization is a process whereby individuals (intentionally or unintentionally) move the muscles used in speech

production without producing audible sounds (Carver, 1990). This process is almost always occurring when reading (Baddeley, 2012; Carver, 1990) - particularly so when reading novel materials (Baddeley, 2012; Carver, 1990), and is a critical rehearsal strategy used in STM study paradigms (Baddeley, 2012; La Pointe & Engle, 1990; Papagno, Valentine, & Baddeley, 1991). By having subjects repeat a specific, irrelevant word or sound aloud, the subvocalization process is disrupted as the muscles that would otherwise be silently mimicking the material to be rehearsed are now preoccupied (e.g., repeating the word 'the' aloud while reading material to be remembered). The disruption of subvocalization does not impact the recall of familiar words (though it can when the familiar words are long enough; see La Pointe & Engle, 1990) but drastically reduces the correct recall of novel words in a foreign language (Ellis & Beaton, 1993; Papagno et al., 1991) and pseudowords (Papagno & Vallar, 1992). This cluster of research demonstrates both how the processing of pseudowords is less aided by LTM contents as well as the greater cognitive-processing power needed to hold pseudowords in active WM.

The differences in recall and production performance between familiar words and pseudowords highlights the nature of the concatenation of WM and LTM processes. The familiarity of material often influences the ease with which it is recalled (Anderson, 1981; Henry & Miller, 1991; Hulme et al., 1991) and indicated by findings that memory span is greater for words rather than pseudowords – even when controlling for time of utterance. However, while those findings represent a difference in recall between familiar and unfamiliar stimuli, differences in recall can be found between pseudowords (as well as words in a foreign language) that resemble typical construction in one's native language compared to those that are less typical (Ellis & Beaton, 1991; McNulty, 1965; Papagno & Vallar, 1992; Service & Craik, 1992). These latter findings suggest that not all novel stimuli are equally accommodated in cognition

and that pseudowords that are highly wordlike are done so more easily. Furthermore, more easily reproduce highly wordlike pseudowords compared to pseudowords of low wordlikeness (Gathercole et al., 1999) and the ability to correctly produce low-wordlike pseudowords is more correlated with STM performance than the correct production of highly-wordlike pseudowords (Gathercole, 1995). Taken together, it appears that LTM plays a role in supporting the representation of familiar and novel material in WM though this role is necessarily more limited in the case of the latter. Familiarity with phonological content begins to emerge at a very early age (Jusczyk, Luce, & Charles-Luce, 1994) and with few exceptions generally benefits word recognition, memory, and learning (e.g., Storkel et al., 2006).

A few mechanisms have been identified explaining this beneficial effect with one being that more familiar pseudowords involve less cognitive processing as subjects not only more deftly reproduce highly wordlike pseudowords, but are able to do so more quickly (Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997). Other research has found subjects often associate novel words with known words as a mnemonic (Rodgers, 1969) and more wordlike novel words are recalled more efficiently because it is easier for participants to form associations with them. Papagno et al., (1991) examined differences in learning word pairs consisting of native language words and Russian words across two different linguistic groups – those fluent in Italian and those fluent in English. While Russian words are not pseudowords in the same sense that has primarily been discussed, those that are still pronounceable when transliterated from a Cyrillic to a Latin script effectively function as pseudowords (see Cohen & Aphek, 1980; Hulme et al., 1991; Rodgers, 1969; Service & Craik, 1993, for other research utilizing foreign vocabulary in lieu of pseudowords). When subvocalization was disrupted, Italian subjects were able to recall Italian-Italian word pairs but experienced difficulty recalling Italian-Russian word pairs. However,

English subjects, though recalling fewer English-Russian than English-English word pairs, did not experience a deficit for the former when subvocalization was prevented. It was revealed that English subjects more easily formed associations between the Russian and English words – a strong display of LTM's tenacity for propping up WM contents. When replicating the study but using pseudowords and Finnish words bearing much less resemblance to common English words, the English subjects were less able to form associations and recall performance dropped as expected when subvocalization was prevented.

Beyond familiarity, other properties of pseudowords have been found to influence memorization and recall. Pseudowords have been used to replicate the word length effect (Baddeley, Buchanan, & Thomson, 1975; LaPointe & Engle, 1990), with longer pseudowords being more difficult to recall than shorter ones (Ellis & Beaton, 1993; Hulme et al., 1991; Papagno & Vallar, 1992). Estimates suggest that the capacity of the phonological loop is approximately two seconds worth of information (Baddeley, Buchanan, & Thomson, 1975; Hulme et al., 1991), as such longer pseudowords take up more of this finite capacity. Another property is the similarity of the words or pseudowords. When subjects are asked to recall phonologically similar pseudowords, performance is lower than when the pseudowords are not similar (Papagno & Vallar, 1992). Interestingly, this pattern reverses when moving from a STM study paradigm to a LTM study paradigm (Baddeley, 1966).

Based off of the above literature review, it appears that pseudowords engage WM in a manner that is quite different than words with both multistore and unitary store theories being able to account for this phenomenon. Generally speaking, pseudowords engage WM processes in a way which limits LTM's ability to function in an auxiliary role and which when considered in a verbal reasoning assessment context suggests that pseudowords can greatly minimize

contamination due to prior familiarity from entering the assessment. From a multistore perspective, pseudowords are uniquely handled by the phonological loop. The memorization of pseudowords appears to require the phonological loop, central executive, and episodic buffer to act in unison with the phonological loop holding the sensory trace and the latter two functions embarking on a quest to make this task easier – whether it be by rehearsal, educating LTM support in the form of association words, or some other strategy. From a unitary perspective of WM, the implications are similar. Pseudowords limit the support that LTM offers and based off of the literature, it appears that pseudowords have different baseline levels of activation depending on their wordlikeness. Furthermore, given the lack of LTM representation, pseudowords should be more difficult to chunk in an effort to free up the limited space in WM. As an example, the ability to combine multiple terms into a single chunk is easy if given the terms ‘elephant’ and ‘ice hockey’ but much more difficult if given the terms ‘contramponist’ and ‘ice hockey’. A pathway to examine individual differences in WM in terms of capacity or the efficiency of controlled attention (or both) is plausible given that pseudowords can and do engage WM processes and that people vary in the ability to recall lists of pseudowords. As such, pseudowords offer a way to test individual differences in WM in a manner which minimizes the confounding effects of prior familiarity, a notion which can have implications for the assessment of Gf.

Summary

There is much theoretical (Baddeley, 2012; Cattell 1943), empirical (Ackerman et al., 2005; Buehner et al., 2006; Chuderski et al., 2012; Colom et al., 2008; Kane et al., 2005; Oberauer et al., 2005; 2008), and neurocognitive (Barbey et al., 2014; Burgess et al., 2011; Gray, Chabris, & Braver, 2003) work suggesting that a considerable amount of overlap exists between WM and Gf constructs with the former being critical for the effective execution of the latter

(Buehner et al., 2006; Chuderski et al., 2012; Jaeggi et al., 2008; 2010; 2011; Oberauer et al., 2005). With regard to Gf, pseudowords are an item feature of increasing popularity as they provide a way to assess verbal Gf in a manner that reduces prior familiarity of the verbal item content (Fagan & Holland, 2009; Sternberg, 1981). Furthermore, pseudowords have been used frequently in studies assessing STM and WM properties. They uniquely engage the phonological loop (Baddeley et al., 1988; Gathercole, 1995) and limit support provided by LTM when maintenance, memorization, and recall are required (Hulme et al., 1991). Short-term memory and WM can be differentially taxed depending on the properties of the pseudowords with longer and less wordlike pseudowords requiring greater resources for maintenance, processing, and production (Ellis & Beaton, 1993; Gathercole, 1995; Gathercole et al., 1999; Hulme et al., 1991; Papagno et al., 1991; Papagno & Vallar, 1992; Service & Craik, 1993; Vitevitch et al., 1997). Given that pseudowords can be altered to differentially draw upon STM and WM resources and that WM strongly relates to Gf, it is believed that altering pseudoword properties has implications for assessments of verbal Gf. Specifically, the difficulty of Gf items featuring pseudowords may be altered as pseudoword properties are altered. The following section makes the case for verbal Gf item properties being varied when two pseudoword properties are altered: pseudoword length and lexical neighborhood density.

Chapter 4: Pseudoword Properties and the Implication for Item Properties

It is broadly theorized in this research that the properties of pseudowords have implications for verbal Gf item properties and in particular, item difficulty. Specifically, differences in the length and wordlikeness (e.g., lexical neighborhood density) of the pseudoword are believed to differentially consume cognitive resources whether it be by requiring LTM to exert more effort in an attempt to support WM contents, pseudowords having a lower (or no) resting level of activation and thus requiring more resources to hold in an activated state, or to manipulate in a verbal Gf context. Though examining the properties of pseudowords in a cognitive ability testing context is perhaps novel, a similar exercise involving real words has been the basis for assessing verbal demands in text. Word length and the familiarity/frequency of words are common variables considered when assessing the readability of a body of text (Benjamin, 2012; Drum, Calfee, & Cook, 1981; Klare, 1974-1975). Drum et al. (1981) reports that polysyllabic (i.e., words of two or more syllables) words and words that are either uncommon or possess an irregular orthographic form (i.e., words in a language that feature an atypical spelling) contribute to difficulty in verbal tasks. While these tasks tend to be reading comprehension tasks rather than analytical verbal reasoning, the word properties discussed are quite parallel to pseudoword length (i.e., polysyllabic vs. monosyllabic) and neighborhood density (i.e., uncommon/orthographic form).

Pseudoword length. The word length effect is the finding that lists of longer words are more difficult to memorize and recall than lists of shorter words, even if the total number of words to be recalled is the same (Baddeley, Thomson, & Buchanan, 1975). Initially, it was reported that word length, when operationalized as either spoken duration or number of syllables influences performance on recall (Baddeley, et al., 1975; Hulme, Thomson, Muir, & Lawrence,

1984; Hulme, Maughan, & Brown, 1991; Roodenrys & Hulme, 1993) and this holds whether the words are presented auditorily or visually (Baddeley et al., 1975; Bireta, Neath, & Surprenant, 2006; Campy, 2011; Cowan, Baddeley, Elliott, & Norris, 2003; Hulme, Neath, Stuart, Shostak, Surprenant, & Brown, 2006; Hulme, Surprenant, Bireta, Stuart, & Neath, 2004). However, when operationalized as spoken duration, word length has been found to create the word length effect only when using the original Baddeley et al. (1975) stimuli. While this set of stimuli has consistently produced the word length effect during replication (Cowan, Day, Sauls, Keller, Johnson, & Flores, 1992; Lovatt, Avons, & Masterson, 2000; Mueller, Seymour, Kieras, & Meyer, 2003), studies creating new sets of stimuli have failed to consistently find a spoken-duration word length effect (Lovatt et al., 2000; Neath, Bireta, & Surprenant, 2003; Service, 1998). When operationalizing word length as the number of syllables, the word length effect has received much more support (Baddeley et al., 1975; Cowan et al., 2003; Hulme, et al., 2006; Hulme et al., 2004; Roodenrys, Hulme, & Brown, 1993; Tehan, Hendry, & Kocinski, 2001; but see Jalbert, Neath, Surprenant, 2011 and Jalbert, Neath, Bireta, Surprenant, 2011).

The word length effect however has received challenges and no cognitive-based theory to date either accommodates or predicts all of the word length effect findings (or lack thereof). Specifically, some studies report that word length is irrelevant for learning lists of words (Papagno & Vallar, 1992) or that articulatory suppression during learning creates a situation where lists of longer words may actually be easier to memorize than lists of shorter words (Romani, McAlpine, Olson, Tsouknida, & Martin, 2005). Long words may also be more likely to be recalled correctly when they are unique in the list of stimuli (i.e., when a long word is included in a list of short words; Hulme et al., 2006) or in instances of cued recall in a LTM study paradigm (Tehan & Tolan, 2007). While findings in support of the syllable-based word

length effect in forward serial recall generally remain fairly strong, studies requiring backward serial recall find no clear evidence in support of a word length effect (Surprenant, Brown, Jalbert, Neath, Bireta, & Tehan, 2006). However, it is critical to note that many instances where the word length effect does not hold are specific to words and do not necessarily apply to pseudowords.

One of the more notable challenges to the word length effect has come from the work of Jalbert and colleagues. Their position is that prior word length effect studies, while controlling for word characteristics such as familiarity, frequency, concreteness, part of speech, and imageability, never controlled for lexical neighborhood density which is hypothesized as contributing to the observed word length effect. Resulting from several studies it was reported that lexical neighborhood density accounted for the word length effect, that articulatory suppression removes the neighborhood density effect in the same way it does for the word length effect, and most intriguing, that longer words from larger neighborhoods are more easily recalled than shorter words from smaller neighborhoods. This is found to be the case for both words and pseudowords (Jalbert, Neath, & Surprenant, 2011; Jalbert, Neath, Bireta, & Surprenant, 2011).

Apart from serial-recall studies, research involving other tasks has reported word length effects. In lexical decision tasks where subjects are shown a word or pseudoword and asked to confirm whether the stimuli is real or not, and in naming tasks, where subjects are required to pronounce the stimuli aloud, word length effects are found with longer words requiring more processing time than shorter words (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; New, Ferrand, Pallier, Brysbaert, 2006; Yap & Balota, 2009; but see Bijeljac-Babic, Millogo Farioli, & Grainger, 2004 who only found the effect in children). These findings are typically more robust for pseudowords compared to words as well. While results are not necessarily consistently found and effect sizes may be somewhat modest in non-linguistically and non-

cognitively impaired populations (Barton, Hanif, Björnström, Hills, 2014; New et al., 2006), when effects are found they are generally supportive of longer words having inhibitory effects on the desired response. One notable exception is found in New et al., (2006) where response time and word length demonstrated a curvilinear pattern. Response times decreased as word length increased in length from three to five letters, remain flat from five to eight letters, and increased from eight to thirteen letters. These findings also warrant some hesitation in interpreting the work of Jalbert and colleagues as all of the long and short pseudowords used were between four and six letters.

Given the general empirical support of the word length effect in the memory literature, the sensitivity of the word length effect regarding the specific experimental task used, and the presence of a word length effect in non-memory research, it is justifiable to postulate that the word length effect may manifest in responses to verbal reasoning items. With Jalbert and colleagues' findings in mind, the following hypothesis was proposed:

Hypothesis 1: Items containing longer pseudowords will be more difficult than items containing shorter pseudowords when controlling for wordlikeness (operationalized as lexical neighborhood density, lexical neighbor frequency, and orthographic probability).

Wordlikeness. Not all words in a given language are equally prototypical of words in that language. For instance, consider the difference between the words 'line' and 'foxy'. The former appears considerably more typical of a word in English for a number of reasons, including featuring letters more frequently represented in the orthographic corpus, having more words that are spelled and pronounced similarly, as well as perhaps just simply sounding more like a typical word – all features which would lead the more wordlike word to have a greater number of lexical neighbors and a higher orthographic probability. Similarly, the structure of words influences how

wordlike they appear, such as the difference between the words ‘knee’ and ‘keen’. While the former appears more frequently in the English language, the latter more closely resembles a word in English orthographically and features a pronunciation that is more typical of the orthographic form. Hence, wordlikeness can be thought of as the extent to which a specific written or sound sequence represents a typical written or sound sequence in a given language (Bailey & Hahn, 2001). While subjective ratings of wordlikeness are popular in research (Bailey & Hahn, 2001; Cheung, 1996; Gathercole, Willis, Emslie, & Baddeley, 1991), there exist a handful of objective ratings as well. Common objective measures of wordlikeness include phonotactic/orthographic (sequence) probability, lexical neighborhood density, and lexical neighbor frequency.

Sequence probability refers to several operationalizations in the literature and is best broadly conceptualized as the frequency that individual letters/sounds or letter-/sound-clusters occur in a given language (Bartolotti & Marian, 2014; Ellis & Beaton, 1993; Storkel, Armbrüster, & Hogan, 2006). Two forms of sequence probability are orthographic probability and phonotactic probability which refer to written (visual) and phoneme (auditory) frequencies, respectively (Ellis & Beaton, 1993; Storkel et al., 2006).

Sequence probabilities have two approaches for calculation: relative approaches and absolute approaches. Both of these approaches are similar with the principal difference being absolute approaches consider the position of a sound, letter, or letter-cluster in a word (Ellis & Beaton, 1993). For instance, a relative bigram (i.e., two adjacent letters) approach will consider how often a specific bigram appears in words in a given language. For instance the bigram ‘st’ will appear in the words ‘start’, ‘stop’, ‘past’, ‘list’, etc. An absolute approach considers the frequency that the ‘st’ bigram appears as the third and fourth letters in a word, hence ‘past’ and

‘list’ would be considered whereas ‘start’ and ‘stop’ would not. Generally, words with higher phonotactic or orthographic probabilities feature sequencing more typical in a given language and are considered more wordlike than words with lower probabilities (Frisch, Large, & Pisoni, 2000). Additionally, humans are very sensitive to phonotactic probabilities with infants starting to display preferences for common sound sequences by nine months of age (Jusczyk, Cutler, & Redanz, 1993; Jusczyk, Luce, & Luce, 1994). Further, children and adults more easily recall pseudowords that obey a novel set of phonological rules after only a brief exposure to the words during an incidental learning task (Majerus, Van der Linden, Mulder, Meulemans, & Peters, 2004).

Lexical neighborhood density has multiple operationalizations in the literature but generally is viewed as the number of words that can be created following a one-letter or one-phoneme alteration to a word or pseudoword (Bailey & Hahn, 2001). While some conceptualizations of lexical neighborhood density consider only words of the same length to the target word or pseudoword (Jalbert et al., 2011a, 2011b; Medler & Binder, 2005), a more flexible and typical conceptualization is one that allows for letter or phoneme additions and deletions as well as substitutions, thus including words with one letter/phoneme more or less than the target word in the lexical neighborhood (Bailey & Hahn, 2001, Luce, 1986; Luce & Large, 2001). Words or pseudowords with more lexical neighbors (i.e., having denser neighborhoods) are generally rated as more wordlike than words or pseudowords with few lexical neighbors (i.e., sparse neighborhoods).

Finally, lexical neighbor frequency considers not the amount of total neighbors a word has, but how often those neighbors occur in common language use (Jalbert et al., 2011; Vitevitch, Storkel, Francisco, Evans, & Goldstein, 2014). In other words, neighborhoods of equal

sizes are still not necessarily equal in terms of cognitive accommodation. Vitevitch et al., (2014) created pseudowords that had neighborhood sizes of just one. However, there were two pools of pseudowords, those with one neighbor that appeared frequently in language usage (e.g., doctor) and those with one neighbor that appeared infrequently (e.g., daytime). These pseudowords were paired with novel pictures. In a recall phase where participants had to recall the correct name after presentation of the novel picture, participants more effectively remembered pseudowords from the more frequent neighborhood pool. Other studies report similar results with words from high-frequency neighborhoods being more easily recalled than words from low-frequency neighborhoods (Roodenrys, Hulme, Lethbridge, Hinton, & Nimmo, 2002). However, research has generally considered neighborhood frequency much less often than sequence probability or neighborhood density.

One central difficulty when designing comparison groups of pseudowords is that it is impossible to have complete control over length, sequence probability, lexical neighborhood density, and lexical neighbor frequency due to all four of these properties being correlated, particularly the three properties of wordlikeness (Roodenrys et al., 2002; Storkel, 2004; Vitevitch, Luce, Pisoni, & Auer, 1999). Shorter words will naturally have larger neighborhoods on average compared to longer words due to there being more short words present in language (Bard & Shillcock, 1993; Pisoni, Nusbaum, Luce & Slowiczek, 1985; Zipf, 1935). Similarly, shorter words are generally used more frequently in language (Zipf, 1935), so neighbor frequencies may naturally differ between shorter and longer words.

Regarding wordlikeness, lexical neighborhood density and phonotactic probability are highly correlated, with one study reporting that the two were correlated $r = .91$ in their pool of words (Roodenrys et al., 2002). Other research generally reports a correlation of $r = .61$

(Vitevitch et al., 1999), making it difficult to disentangle the two. When research has compared sequence probability to lexical neighborhood density, lexical neighborhood density is generally found to be the more important attribute as it correlates more strongly with subjective ratings of wordlikeness (Bailey & Hahn, 2001), and in studies of word recall, after controlling for lexical neighborhood density, the impact of sequence probability generally becomes non-significant (Andrews, 1992; Janse & Newman, 2013; Roodenrys & Hinton, 2002; Storkel et al., 2006). Extant research has not featured direct comparisons of lexical neighborhood density to neighborhood frequency. However, in Roodenrys et al. (2002), neighborhood density had stronger effects than neighborhood frequency.

Given the difficulties around being able to develop groups of pseudowords where only one property varies, and given lexical neighborhood density's general dominance in the literature, the proposed studies will only consider lexical neighborhood density as a wordlikeness variable. Orthographic probability and lexical neighbor frequency will be controlled for in the groups of pseudowords, but will not be directly manipulated.

While sequence probability has produced more inconsistent results in word recall and word learning studies, lexical neighborhood density has had fairly consistent effects. In studies involving serial recall tasks, lists featuring wordlike (pseudo)words are recalled better than lists of un-wordlike (pseudo)words (Allen & Hulme, 2006; Jalbert et al., 2011; Roodenrys & Hinton, 2002; Roodenrys et al., 2002; Thorn & Frankish, 2005). In word learning studies where pseudowords are paired with novel objects, neighborhood density is positively related to word learning whether the sample consists of adults (Storkel et al., 2006) or preschoolers (Storkel, Bontempo, Aschenbrenner, Maekawa, & Lee, 2013). One exception to this may be in the case of long-term recall. There is evidence suggesting that while pseudowords embedded in dense

neighborhoods are learned more effectively following immediate testing, after a period of forgetting (e.g., one week), participants are more likely to correctly remember sparse neighborhood density/novel object pairings (Storkel, Bontempo, & Pak, 2014; but see Storkel & Lee, 2011 as a counterexample involving preschooler participants). In repetition tasks where a word or pseudoword is presented and participants are asked to repeat the word correctly as quickly as possible, lexical neighborhood density has an inhibitory effect when the stimuli are real words but a facilitating effect when pseudowords are the stimuli (Andrews, 1992; Vitevitch & Luce, 1998).

The bulk of the studies just described have considered phonetically-based as opposed to orthographically-based neighborhoods. Unlike sequence probability, there appears to be little explicit attention given to possible modality-based differences in neighborhood conceptualization (i.e., neighborhoods based on phonemes for auditory presentation vs. neighborhoods based on letters for visual presentation). However, the limited research that considers the effects of orthographically-based neighborhood density on the serial recall of visually presented lists reports similar findings to those described above. Whether consisting of real words or pseudowords, lists are recalled better when they feature dense neighborhood (pseudo)words compared to sparse neighborhood (pseudo)words (Jalbert et al., 2011a, 2011b).

However, it should be noted that the current study ultimately used an alternative conceptualization to neighborhood density called the Levenshtein distance (Yarkoni, Balota, & Yap, 2008). While the term Levenshtein distance simply refers to a mathematical term, in this context it specifically represents the mean number of transformations (i.e., letter addition, deletion, substitution, and transposition) needed to generate a given word's twenty closest orthographic neighbors. Unique to the Levenshtein distance calculation is the inclusion of the

transposition which means that flipping two letters in a word is counted as one alteration rather than two alterations in the form of two substitutions. As an example, a neighbor that is one transposition alteration away for the word ‘trail’ is ‘trial’. While explained more in depth in Yarkoni et al. (2008), software exists that computes the Levenshtein distance for a target word (or pseudoword) by comparing it to a corpus of 62,400 words. Levenshtein distances have one as their lowest possible value (meaning that a (pseudo)word’s twenty closest orthographic neighbors are all one alteration away) and an upper bound that varies depending on the number of letters in a given word or letter string. Based on the scant literature (Suarez, Tan, Yap, & Goh, 2011) as well as insights drawn from two pilot studies conducted by the author (see chapters 6, 7, and 9 in this paper for more information), a typical range for Levenshtein distance is between one and four.

The reasons for using the Levenshtein distance as opposed to alternative measures such as neighborhood density or neighborhood density frequency are 1) it represents the usage of a novel or scarcely used measure of wordlikeness, and 2) it provides a more granular measure of wordlikeness/neighborhood density and can provide wordlikeness information on groups of words that all share the same number of neighbors (Suarez et al., 2011). To display the difference between Levenshtein distance and neighborhood density, consider the words ‘grain’ and ‘again’ as well as the pseudoword ‘agtop’ which have lexical neighborhoods of 5, 0, and 0⁴ (or 12, 1, and 1, by an alternative metric⁵), respectively, but Levenshtein distances of 1.5, 1.9, and 2.35, respectively. Hence, by neighborhood density, ‘again’ and ‘agtop’ are equally unwordlike whereas by Levenshtein distance ‘agtop’ is clearly less wordlike than ‘again’. This

⁴ Neighborhood density obtained via MCWord; Medler & Binder, 2005.

⁵ Neighborhood density obtained via ClearPond; Marian, Bartolotti, Chabal, & Shook, 2012.

added level of granularity becomes critical when developing pseudoword stimuli that are designed to be less wordlike as often they will have neighborhood densities of one or zero but Levenshtein distances with considerably more variance. Lastly, it should be noted that unlike neighborhood density, where higher values indicate more wordlike, for the Levenshtein distance metric, lower values indicate pseudowords are more wordlike.

Empirical evidence has strongly favored the notion that more wordlike words (in the form of a denser lexical neighborhood) are processed more quickly as well as learned and recalled more accurately than less wordlike pseudowords. Further, the one possible disadvantage where wordlike words are more difficult to recall in a long-term recall context is not relevant in a Gf assessment context. Hence the following hypotheses were proposed:

Hypothesis 2: Items containing un-wordlike (operationalized as lexical neighborhood density) pseudowords will be more difficult than items containing wordlike pseudowords when controlling for pseudoword length, lexical neighbor frequency, and orthographic probability.

Hypothesis 3: Pseudoword properties will interact such that items featuring long and un-wordlike pseudowords will be more difficult than the other item type combinations.

Chapter 5: Bilingualism

Chapter five is organized as follows: first a description of bilingualism and issues unique to studying bilinguals followed secondly by contemporary theories describing the development, arrangement, and navigation of bilingual semantic and lexical knowledge, and third, a description of cognitive domains where performance differences are observed between bilinguals and monolinguals, with hypotheses offered at the end.

As described in the preceding chapters, long-term memory (LTM) and working memory (WM) are inexorably linked. As an example of the impact of LTM on WM processes, it is easier to remember lists of known words compared to lists of pseudowords (Papagno, Baddeley, & Valentine, 1989) and it is easier to remember lists of highly wordlike pseudowords compared to un-wordlike pseudowords (Gathercole, Willis, Emslie, & Baddeley, 1991). Put differently, the greater the presence a (pseudo)word has in LTM, the easier it is to recall in serial recall tasks. While the impact of LTM on WM can depend on the task (i.e., lexical decision tasks vs. serial recall tasks vs. long-term word learning), the larger point remains the same: the two memory processes rely on one another for better or worse.

Information stored in LTM impacts task performance in a number of ways, for instance a larger vocabulary would impact verbal WM by providing denser lexical neighborhoods for words (or pseudowords). Additionally, greater vocabulary offers more options to choose from if using keyword strategies (i.e., supplanting a pseudoword with a known word) in memorization/word-learning tasks. Another example of the link between WM and LTM is research suggesting that lists of words are more easily memorized if they share a semantic category (e.g., the category of ‘animals’; Mandler, 1967; Nott & Lambert, 1968). Given that LTM supports verbal WM constantly, it is prudent to consider how this impacts bilinguals who

are less proficient when operating in their second or non-dominant language in verbal tasks.

While bilinguals do not necessarily have less experience with language generally (i.e., a bilingual individual has been speaking approximately just as long as a monolingual individual of the same age), they almost always have a smaller vocabulary, less practice, and less experience with their non-dominant language compared to native or monolingual speakers of that language (Bialystok & Luk, 2012; Bialystok et al., 2010; Oller & Eilers, 2002, Perani, Abutalebi, Paulesu, Brambati, Scifo, Cappa, & Fazio, 2003). Thus, the LTM support provided to bilinguals in their non-dominant language is typically less than that provided to them in their dominant language or that provided by LTM for monolinguals (though not always).

Bilingualism

While the term ‘bilingual’ is one that is easily understood colloquially, it is a term fraught with ambiguity operationally. No two bilinguals are necessarily alike, varying across important characteristics such as the two languages known, whether those two languages are alphabetic or not (e.g., English vs. Japanese) and if yes, whether the alphabet is shared or not (e.g., Roman vs. Cyrillic alphabets), how early in life second language learning began, the developmental context of second language acquisition (e.g., a second language spoken both at home and work/school vs. work/school only), and the speaker’s proficiency in not only their second language but their first language as well. The variance encompassed by the term ‘bilingual’ has ensured that bilinguals as a population are resistant to easy sub-categorization for research purposes. This difficulty with classifying bilinguals has led to constant revisions to classifications of bilinguals and theories of bilingualism (see Basnight-Brown, 2013). Nevertheless, contemporary research has made notable strides towards understanding the bilingual mind and the impact bilingualism

has on thought, cognitive abilities, and even aging (Bialystok, 2005; Bialystok, Craik, & Luk, 2012; Kroll, Dussias, Bice, & Perrotti, 2015).

Originally, it was believed that bilinguals were essentially two monolinguals in one and that lexical information was held and accessed separately within the bilingual mind (see Grosjean, 1989 for a review). However, recent research shows that the bilingual mind has unitary lexical and semantic storages with non-selective access to both languages (Brenders et al 2010; Lopez & Young, 1974; Marian & Spivey, 2003; Van Heuven, Dijkstra, & Grainger, 1998). Put differently, languages are not stored in different parts of the brain nor in neurological isolation from one another and speaking in one language does not ‘shut off’ the other. Both of a bilingual’s languages are active when engaging with language across activities such as reading, listening, and planning speech production (Kroll, Dussias, Bice, & Perrotti, 2015) with this effect even holding across modalities (i.e., even if the second language is a language such as American Sign Language; Monford, Wilkinson, Villwock, Piñar, & Kroll, 2011; Shook & Marian, 2012). As such, one’s first language exerts a strong influence on cognitive processes in their second language, and with time and proficiency, the reverse can quickly become true as well (Duyck & Brysbaert, 2004; Kroll & Stewart, 1994; Marian & Spivey, 2003; Wu & Thierry, 2010). Given the integrated, non-partitioned cognitive structure of language(s), this suggests that the LTM of both monolingual and bilingual speakers is constantly offering or attempting to offer support for the cognitive task at hand, but due to differences in experience and proficiency between monolingual and bilingual speakers, the support offered may vary considerably.

Other streams of research have provided more clarity around other issues associated with bilingualism. Specifically, while there is a clear critical period in development for learning a first language, after which learning any language as a first language will be nigh-impossible, there is

no such critical period for learning a second language (Singleton, 2005). Second languages can be learned at any time in life with the qualifier that developmentally, earlier second language exposure and practice leads to better proficiency (Birdsong, 2005). That being said, while mastery (i.e., near native-like control of a language) of a second language is typically observed when an individual begins learning that second language at a young age, mastery of a second language has been observed for relatively late-learners (i.e., post-puberty; Birdsong, 2005, but see Abrahamsson & Hyltenstam, 2008, for nuance). A second but related finding, is that while one's age of acquisition of a second language was initially considered the relevant focus for understanding second-language proficiency, contemporary research on bilingualism has shifted from age of acquisition to favoring actual proficiency of each language. In other words, it was believed that outside of situations where an individual grew up in a bilingual home/began learning two languages from birth, an individual's first language was always their dominant language, with age of acquisition serving as a useful index for understanding where an individual stood in terms of second-language proficiency (Peal & Lambert, 1962). Currently, the second language is not always assumed to be the weaker, or less-proficient language. As such, current research uses terminology such as 'dominant language' and 'non-dominant language' or merely assigns the label of 'first language' to whichever language an individual is more proficient (Francis & Baca, 2014).⁶ This distinction is notable as it is not uncommon for bilingual individuals to be exposed to and use their first language exclusively in the home from birth until school age, and then ultimately develop greater vocabulary/proficiency in a second language as

⁶ The current research adheres to the 'dominant' and 'non-dominant' conceptualization of bilingual proficiency, thus that terminology is preferred throughout. However, the terms 'first' and 'second' language will be used when describing previous research that used those terms for clarity.

this second language is used both more often and across more contexts. Thus for researchers attempting to understand how linguistic status/ability impacts verbal fluid reasoning, simply knowing the first and second languages of bilinguals is not enough. Thus the current research adopts ‘dominant’ and ‘non-dominant’ labels for languages.

Models of Bilingualism. Current models of cognition are gravitating away from computer-metaphor models and towards more sophisticated connectionist models (alternatively referred to as parallel processing models; McClelland, 2000). Traditional computer-metaphor models depict cognition as the serialized conjoining of sensory inputs, short-term storage, and long-term storage, with each of these as relatively discrete phases/components of cognition typically with buffers and feedback loops between components allowing for interaction between these components (Benjamin, 2006; Roediger III, 1980). Somewhat ironically, improvements in computer power and processing have facilitated the shift away from computer-metaphor models towards connectionist models of cognition (McClelland, 2000; McClelland & Cleeremans, 2009; Smith, 2009). These connectionist models blur the boundaries between short-term and long-term storage and further explicate the mechanisms behind the formation and retrieval of associations and memories. These models attempt to understand and view the brain as a connection of neurons with a loose ‘blank slate’ starting point, whereby Hebbian learning provides the mechanism through which neurons begin to become increasingly interconnected allowing for the development of concepts, memories, and ability to draw associations and make inferences (Hebb, 1949; McClelland, 2000; Smith, 2009). Over time, collections of neurons that have been activated together become more strongly interconnected and the information embedded in the patterns of activation becomes easier to retrieve. This dovetails nicely with unitary storage

models of WM where by WM is embedded within LTM and contents are more easily retrieved from LTM into WM when they have higher resting-levels of activation.

This shift in theories of cognition also coincides with a similar trend in the bilingualism literature whereby connectionist models of language development and bilingual development are growing in popularity (Dijkstra & Rekke, 2010; Li & Zhao, 2013; Rohde & Plaut, 2003).

Connectionist models of language posit that over repeated exposure to a language, lexical storage develops and becomes connected to semantic storage,⁷ or that words begin to be learned and then paired with their definitions and physical referents. Similar processes occur for the development of a lexicon for a second language which becomes paired not only with the semantic storage but with the lexicon of the first language as well (Li & Zhao, 2013). Given that less-familiar (in terms of experience/exposure) information has fewer neuronal connections and has greater activation costs in models of cognition, similar patterns are found in models of bilingualism, particularly when comparing dominant to non-dominant language processing (Kroll et al., 2015; Kroll & Stewart, 1994; Takano & Noda, 1995; Zhao & Li, 2010). The reduced availability of neuronal supports as well as the greater costs of maintaining less familiar information in an activated state perhaps in part explains test-score differences on verbal reasoning between test-takers completing a test in their dominant vs. those completing it in their non-dominant language (Kobrin et al., 2007; Kena et al., 2016; Takano & Noda, 1995).

As opposed to sequential or serial processing where operations are carried out rapidly one after another, connectionist models allow for simultaneous operations, thus connectionist models view cognitive processing as concurrently occurring across the brain and as a simultaneously

⁷ The term ‘storage’ here refers to the collection of relevant neurons, not necessarily a discrete region in the brain, which would not necessarily be compatible with connectionist models.

top-down and bottom-up process (Smith, 2009). These models attempt to mirror biological neuronal functioning where cognition is viewed as the product of patterns of activation across nodes (Smith, 2009). Connectionist models of language typically model Hebbian interactions across nodes that are themselves arranged across several linguistic maps or levels. Depending on the model these can be linguistic maps consisting of letters, words, and language as a whole (e.g., the BIA model, Dijkstra & Van Heuven, 2010), or lexical-semantic and phonological maps (e.g., the Lex-Dev model, Li & Zhao, 2013; Zhao & Li, 2010). Nodes within and between maps are linked via connections that respond to Hebbian learning rules. These connections adjust across simulation iterations when the model is trained on a set of inputs such as a number of words in a first and second language along with the frequencies of word occurrence in a realistic, bilingual setting. Connectionist models have been able to accurately generate model results that correspond with the extant empirical results of the developmental trajectories of both languages in bilinguals across various developmental contexts (Zhao & Li, 2013). However, most relevant for the present research, given the Hebbian learning rules that connectionist models obey, a less familiar, less practiced language would have less iterations in a model and typically feature weaker connection weights. Broadly put, a bilingual's non-dominant language can be viewed of as a 'weaker' language in terms of overall proficiency, practice, and the quality and quantity of neuronal connections – whether that be conceptualized as the connection strength between nodes or resting levels of activation.

Indeed, pre-connectionist models posited similar ideas. The Revised Hierarchical Model (RHM) is a non-connectionist model that views the bilingual mind as having two lexical storages each with varying connections to one another as well as to a semantic storage (Kroll & Scholl, 1992; Kroll & Stewart, 1994). Initially, there is one lexical storage that emerges to connect to a

semantic storage. Over time, a lexical storage for a second language begins to develop though the development of this second language is mediated through the lexical storage of the first language. In other words, early on in second language development, the lexical storage of the second language has access to the semantic storage only through the first language. With enough exposure and practice, the second language can directly access the semantic storage, though this access will never be as strong as the first language's. Similar to a connectionist model, the RHM posits that a second language is housed in the brain in a weaker or less strongly connected fashion than the first language. One discrepancy is that a connectionist model does not assume that the less-dominant language is the second language. However, the RHM also suggests something of a cognitive cost when operating in one's second language given that second language processing relies on/is mediated by first language storage.

A more recent model that uses neuroimaging techniques to understand how bilinguals effectively deploy and switch between two languages also suggests something of a cognitive cost for bilinguals when operating in a non-dominant language. The inhibitory control hypothesis (Abutalebi & Green, 2007; Green, 1998) posits that both of a bilingual's lexica are in competition with each other and the bilingual effectively deploys each language through the mechanism of inhibition. Research has identified that the costs of switching between languages, as measured by reaction time, are greater when switching from a second, or non-dominant language to a first, or dominant language (Abutalebi & Green, 2007; Kroll & Stewart, 1994). While it may seem counterintuitive at first to experience greater difficulty when moving from a less to a more proficient language, this switch cost is viewed as the result of greater resources devoted to inhibiting the more proficient language and thus creating difficulties un-inhibiting the dominant language. In line with connectionist models, the inhibitory control hypotheses suggests

that both languages are active during linguistic processing, and that whichever language is not currently in use needs to be inhibited. This process of inhibition demands greater resources when using the non-dominant language (Abutalebi & Green, 2008).

In addition to suggesting that language processing and production in all languages are served by common neuroanatomical structures (Abutalebi & Green, 2007; Abutalebi & Green, 2008), the inhibitory control hypothesis also examines the neurocognitive network underlying cognitive control and task-switching. For monolinguals in non-linguistic tasks, cognitive control and task-switching tasks activate the prefrontal cortex, the anterior cingulate cortex (ACC), the parietal cortex, and the basal ganglia (Abutalebi & Green, 2007). These regions are also broadly utilized in language switching tasks for bilinguals as well, corroborated by research involving aphasics who experience difficulties effectively switching languages after lesions in the left prefrontal cortex and posterior parietal cortex areas (Luk, Green, Abutalebi, & Grady, 2011). Other streams of research show that when navigating two languages, the ACC acts as a conflict monitor that sends signals to the prefrontal cortex when the wrong language is chosen (Abutalebi, Della Rosa, Green, Hernandez, Scifo, Keim, Cappa, & Costa, 2012; Abutalebi & Green, 2008). As multiple languages are constantly active, so is the ACC, with proficient bilinguals displaying some cognitive-structural differences in the ACC compared to monolinguals (Abutalebi et al., 2012).

The finding that there are some cognitive-structural differences between monolingual and bilingual brains, coupled with theories of bilingualism that suggest that non-dominant languages are integrated into cognitive processes in a manner that incurs a cognitive cost (whether that be by a non-dominant language featuring less well-developed nodal connections, generally lower resting levels of activation, reliance on a dominant language, or greater resources needed to

inhibit the dominant language), implies there is reason to suspect that monolinguals and bilinguals or bilinguals operating in their dominant language vs. bilinguals operating in their non-dominant language may perform differently on assessments of verbal Gf. However, while it is clear that there are differences between bilinguals and monolinguals and between dominant and non-dominant linguistic ability, it is unclear what these differences mean in a verbal Gf assessment context. While these group differences may be subtle and may not manifest across common or familiar day-to-day activities, they may manifest during more rigorous, demanding, or novel cognitive activities such as a verbal Gf assessment. The following section explores these differences in greater detail in an attempt to predict how linguistic status may impact performance on verbal Gf items involving pseudowords.

Cognitive Differences between Bilinguals and Monolinguals

Bilingualism necessitates the usage of cognitive skills that monolinguals either do not need to use or, at a more general level, do use but for different purposes (e.g., executive control). As just described, needing to learn, manage, and deploy two languages effectively is neither an easy nor an automatic process and over time, bilingualism can lead to neurocognitive-structural changes not found in monolinguals (Abutalebi et al., 2012; Bialystok, Craik, Grady, Chau, Ishii, Gunji, & Pantev, 2005). Thus given differences between the two groups in brain activity, it is reasonable to wonder if linguistic status impacts performance across cognitive tasks. Indeed, bilinguals and monolinguals *do* show differences in performance across a range of certain types of cognitive tasks including word recognition, cognitive/executive control, foreign word learning, word recall, and STM/WM tasks (Bialystok et al., 2012; Kroll et al., 2015).

Overall, bilinguals generally experience an advantage relative to monolinguals on foreign-language word learning, cognitive/executive control tasks, and possibly proactive

interference tasks. With regard to foreign-language word learning, a bilingual advantage compared to monolinguals is found when word learning consists of novel words that are phonologically unfamiliar to both bilinguals and monolinguals and these novel words are either paired with images (Kaushanskaya, Yoo, & Hecke, 2012) or dominant-language translations (Kaushanskaya & Marian, 2009a; Kaushanskaya & Marian, 2009b; van Hell & Mahn, 1997). Moving beyond the bilingual/monolingual distinction, experience with more languages also provides similar effects as reported in Papagno and Vallar (1995) which compared trilinguals to bilinguals and found better learning for the trilingual group. These findings hold whether the dominant language of linguistic groups were the same (Kaushanskaya & Marian, 2009a; Kaushanskaya & Marian, 2009b; Kaushanskaya et al., 2012; Papagno & Vallar, 1995) or different (van Hell & Mahn, 1997). However, all of these experiments involved auditory presentation of the novel words and all the novel words adhered to a novel phoneme-to-orthography mapping. In other words, the stimuli used were akin to learning a new language rather than learning new words in a known language, thus it can be said that generally bilinguals show an advantage to learning words in a new language, though it is unclear if such an advantage exists for reasoning with pseudowords that are presented visually/orthographically.

A second cognitive area where bilinguals tend to experience an advantage compared to monolinguals is the area of executive control. While executive control is typically viewed as a sub-dimension of WM (Cowan, 2005; Hambrick, Kane, & Engle, 2005; Oberauer et al., 2008), and while it is referred to by different names in the bilingual literature (e.g., cognitive control, conflict monitoring, working memory; Morales, Calvo, & Bialystok, 2013), operationally speaking, bilinguals experience an advantage on executive control tasks that feature quick responses of conflict monitoring or incongruence tasks. One executive control task is the Flanker

task which involves the presentation of an arrow that points either left or right and is flanked on either side by two arrows (i.e., flanked by four arrows total). During congruent trials, the flanking arrows point in the same direction as the center arrow and during incongruent trials, the flanking arrows point in the opposite direction, with the participant needing to indicate the direction of the central arrow. There may also be control trials where the center arrow is flanked by dashes (Costa, Hernandez, Sebastian-Galles, 2008). Results of the Flanker task indicate that the reaction times for bilinguals are faster than those of monolinguals, particularly on conflict trials (i.e., those trials where a congruent trial is followed by an incongruent trial or vice-versa) while no group differences appear in terms of error rates or accuracy (Abutalebi et al., 2012; Costa et al., 2008). However, it is not anticipated that this bilingual advantage would carry over to verbal Gf assessments as these executive control tasks tend to require extremely rapid responses to non-verbal stimuli – something very different than the cognitive demands made by verbal Gf items.

The Simon task is also an executive control task that taps into both conflict monitoring and the storage component of WM (Morales, Calvo & Bialystok, 2013). In the Simon task, participants are given a set of objects and a set of rules, such as when a star appears, press the ‘m’ key on a keyboard and when a tree appears, press the ‘z’ key. During the Simon task, the rules for which objects are associated with which key do not change but the location of the objects do, with congruent trials representing trials where the object appears on screen on the same side as the correct key and incongruent trials representing instances where the object appears on screen on the opposite side of the correct key. Working memory load can be manipulated by the number of object-to-key rules that must be remembered (Morales et al., 2013; Schroeder & Marian, 2012). Research indicates that bilinguals outperform monolinguals in

terms of both accuracy and reaction time on the Simon task, whether participants are drawn from elderly (i.e., sixty plus years of age; Bialystok, Craik, Klein, Viswanathan, 2004; Schroeder & Marian, 2012) or from child (i.e., four to seven years of age; Martin-Rhee & Bialystok, 2008; Morales et al., 2013) populations. However, evidence demonstrating a bilingual advantage on the Simon task for non-elderly adult populations is more mixed (see: Bialystok, 2006; Bialystok et al., 2005) while other researchers have questioned the bilingual advantage for children due to failure to control for the SES of linguistic groups (Morton & Harper, 2007). Though the Simon task draws greater similarity to verbal Gf items in that it attempts to impact WM load directly, the bilingual advantage is not necessarily expected to carry over to verbal Gf items for the same reasons that the Flanker-task advantage is not expected to – the quick-response, non-verbal nature of the task limits the applicability of the findings to verbal Gf assessment contexts.

A third type of executive control task is a switch task. In these tasks, participants are given stimuli that have certain properties and the correct response depends on those properties. For instance, Albutalebi et al., (2012) used a language switching task where pictures appeared in either blue or green with the color determining which language participants were supposed to use to name the picture. For monolingual participants, the same set of pictures were used but the colors determined whether monolinguals were required to produce a noun or verb associated with the picture. Gold, Kim, Johnson, Kryscio, and Smith (2013) used a color-shape switch task where stimuli were either squares or circles and were either red or blue. Subjects would have to remember which key to press to indicate whether it was a certain shape or color and trials would switch between color or shape as the criterion. In both studies, there is no difference in accuracy between monolinguals and bilinguals, but bilinguals produce faster reaction times (Abutalebi et al., 2012; Gold et al., 2013). Similar to results from the Simon task, group differences are less

clear for younger adults, but the bilingual advantage is found for older adults (i.e., Sixty years of age, Gold et al., 2013). Thus bilinguals may perform better than monolinguals in testing contexts that involve conflict monitoring or executive control such as the Flanker, Simon, or switch tasks. However, performance differences between linguistic groups across these tasks are not always found for young-to-middle-age adult groups. That in conjunction with these tasks not resembling typical verbal Gf items suggests that any executive control advantages held by bilinguals may not manifest as better performance compared to monolinguals across verbal Gf items.

Lastly, there is some early evidence that bilinguals may be able to better manage proactive interference compared to monolinguals in memory tasks (Bialystok & Feng, 2009). For the proactive interference task, participants were presented four lists of words, one at a time. After the list was presented, there would be a brief filler task followed by participants having to recall the words. The first three lists contained words from the same semantic category (e.g., animals, sports, colors, etc.), and the last list contained words from a new semantic category. While bilinguals and monolinguals performed similarly on lists one and four, both of which had new semantic groups, bilingual children showed better recall of words in lists two and three compared to monolingual children. The better recall of words in lists two and three was due to fewer intrusions of semantically related words from previous lists. However, similar to the Simon and switch tasks, the reduction of proactive interference was considerably less robust for adult participants (Bialystok & Feng, 2009). Thus, bilinguals may feature an advantage compared to monolinguals on tasks that require minimizing interference from earlier tasks/items or rapid updating of information, but the findings are currently too scant to speculate about a clear impact for verbal Gf performance – particularly if the verbal Gf items do not involve heavy repetition or semantic overlap of stimuli.

Bilingualism can be quite advantageous and certainly not the limitation it was originally believed to be in early research (see Hakuta & Diaz, 1985). However, there are still a number of cognitive tasks on which monolinguals do outperform bilinguals. As described earlier, recall and repetition tasks increase in difficulty as words move farther and farther away from dominant-language typicality (e.g., a common word in the dominant language → uncommon word in the dominant language → pseudoword in the dominant language → novel word in a foreign language) due to there being increasingly less support available and provided by LTM. Thus, repetition and recall tasks, where testing involves the non-dominant language of a bilingual, represent those tasks where monolinguals typically outperform bilinguals. Research involving pseudoword repetition reports that monolinguals perform better than bilinguals when the pseudowords are constructed to resemble the non-dominant language of bilinguals/the language of monolinguals (Engel de Abreu, 2011; Kaushanskaya & Yoo, 2013). Kaushanskaya and Yoo (2013) had Korean (dominant)/English (non-dominant) bilinguals repeat a two-, four-, or six-syllable pseudoword either immediately after pseudoword presentation (i.e., a STM task) or after a brief filler task (i.e., a WM task). Performance was always worse in the non-dominant language and declined as word length increased. However, this decline was more precipitous in the more cognitively demanding WM task. Given that these tasks involve both (pseudo)word length and a tighter connection to WM, the findings that individuals completing the task in their dominant language outperform those completing the task in their non-dominant language appear to have greater implications for typical verbal Gf items. However, it should be noted that these tasks involve audible, pseudoword reproduction, thus these findings alone would not necessarily ascribe an advantage to individuals completing verbal Gf items in their dominant language.

In word recall tasks where participants are given lists of words to remember and subsequently recall, the results are similar. Bilinguals perform more poorly in their non-dominant language compared to either their performance in their dominant language or to monolinguals (Fernandes, Craik, Bialystok, Kreuger, & 2007; Francis & Baca, 2014; Glanzer & Duarte, 1971). A study by Glanzer and Duarte (1971) featured bilingual participants memorizing lists containing words drawn from both of their languages. It was reported that bilingual participants recalled more dominant- than non-dominant-language words. Francis and Baca (2014) reported that in free recall paradigms, while participants recalled fewer words in their non-dominant than their dominant language, neither of these numbers were significantly different from the number of words recalled by monolinguals. However, in serial recall paradigms where not only words but their order was necessary for correct recall, non-dominant language recall of bilinguals was worse than both dominant language recall and recall by monolinguals. The addition of a serial-recall component however suggests that bilinguals performing this task in their non-dominant language suffer a greater performance decrement compared to those performing the task in their dominant language once a more demanding WM-load component is added. However, it should be noted that though significant, the observed differences were fairly small. Regarding modality of presentation, worse recall in one's non-dominant language has held across visual (Francis & Baca, 2014) and auditory (Fernandes et al., 2007) modalities of presentation.

Working memory span also shows some differences when shifting from a dominant to a non-dominant language. When assessed via the traditional Daneman and Carpenter (1980) design (i.e., reading span), where subjects are given a series of sentences of which they must confirm the accuracy of its logic/syntax, and then a to-be-remembered word follows each sentence, bilinguals remember less words in their non-dominant than their dominant language or

compared to monolinguals (Service, Simola, Metsaheimo, & Maury, 2002; van den Noort, Bosch, & Hugdahl, 2006; Vejnovic, Milin, & Zdravkovic, 2010). Furthermore, when splitting bilinguals into more and less proficient groups, while monolinguals still perform the best, bilinguals more proficient in the non-dominant language outperform bilinguals less proficient in the non-dominant language, even if both groups report learning and speaking the non-dominant language for equivalent amounts of time (Vejnovic et al., 2010). However, there are variations of the WM span task that are less linguistically demanding such as the operation span task where rather than verify the logic of sentences, participants verify the logic of a mathematical operation which is then followed by a to-be-remembered word or number. Operation span tasks and other non-verbal WM span tasks such as the backwards digit span task typically yield smaller differences or equal performance between bilinguals and monolinguals (Engel de Abreu, 2011; Ratiu & Azuma, 2015; Sanchez, Wiley, Miura, Colflesh, Ricks, Jensen, & Conway, 2010). These findings have a few implications for anticipating verbal Gf performance differences across linguistic subgroups. First, given verbal Gf's heavy reliance on WM, it is believed that individuals reasoning in their non-dominant language will suffer similar performance decrements compared to those reasoning in their dominant language. Second, linguistic group differences are minimized when dealing with numeric memoranda rather than verbal memoranda – even if the numeric memoranda are presented in a non-dominant language. Given that verbal Gf items typically avoid any numeric information, it is believed that this observed performance decrement on WM span tasks will translate to verbal Gf tasks.

Lastly, a study by Takano and Noda (1993) involved participants completing thinking tasks (i.e., non-verbal tasks such as simple mathematics, card rotation, maze-tracing, or surface estimation tasks). However, while performing these thinking tasks, a linguistic task would co-

occur in which participants would sporadically hear a sentence that they had to verify the logic of. Typically, these were simple yes/no sentences such as ‘apples can be blue’. Performance on the thinking task was higher when the linguistic task was in the dominant as opposed to non-dominant language. These results held whether the non-dominant language of participants was English or Japanese. An important note about the handful of studies just described is that the results held across numerous languages, including studies where the bilinguals had varied dominant languages, and across both more and less proficient bilinguals.

Summary. It is generally believed that performing tasks in the non-dominant language consumes more cognitive resources (Service et al., 2002; Takano & Noda, 1993). Such a perspective gels with the theories of bilingualism described earlier suggesting that the non-dominant-language lexicon is more weakly connected to semantic storage (Kroll & Scholl, 1994) or has lower levels of baseline resting activation (Smith, 2009; Zhao & Li, 2013). Corroborating this is the neurocognitive research suggesting cognitive processing is more effortful in the non-dominant language compared to the dominant language as evidenced by higher levels of brain activity (Wang et al., 2007).

That being said, bilinguals do experience some advantages compared to monolinguals, and across all of the cognitive tasks just described, group differences are typically fairly small. However, these tasks are quite different from cognitive tasks typical of a Gf assessment. A bilingual advantage is found on executive control tasks involving conflict monitoring and quick responses. Fluid intelligence assessments typically do not require an extremely quick response (i.e., measurement at the millisecond level), even in speeded test contexts. Bilinguals, and particularly young to middle-aged adult bilinguals rarely featured an advantage on accuracy in cognitive control tasks (Abutalebi et al., 2012; Costa et al., 2008; Gold et al., 2013; but see

Bialystok et al., 2004 for counter evidence on the Simon task) and it is unclear if the bilingual advantage on reaction times is meaningful for Gf items. Similarly, pseudoword repetition tasks do not mirror Gf items as they require the actual pronunciation of pseudowords. While pseudoword repetition may be a useful task for predicting language development, particularly for children (Gathercole, 1995), it is less useful for predicting cognitive ability as pronunciation abilities and productive vocabulary are outside of the realm of Gf assessment. Additionally, there have been some difficulties replicating some of the bilingual advantage findings (Paap, Johnson, Sawi, 2014; Paap, Johnson, Sawi, 2015). Word recall tasks (e.g., Fernandes et al., 2007; Francis & Baca, 2014; Glanzer & Duarte, 1971; and the proactive interference task of Kaushanskaya & Yoo, 2013) tend to be more reflective of STM/WM tasks and involve the rapid presentation of to-be-remembered stimuli (e.g., one item every second). Short-term memory tasks feature a weaker correlation with Gf (Ackerman et al., 2005; Engle et al., 1999), and the rapid presentation of stimuli is unlike Gf contexts.

Word learning tasks appear to be somewhat related to reasoning items involving pseudowords as a novel sequence of letters needs to be processed and understood in some manner. Across word learning studies involving a novel phonology, bilinguals outperformed monolinguals (Kaushanskaya & Marian, 2009a; Kaushanskaya & Marian, 2009b; Kaushanskaya et al., 2012; van Hell & Mahn, 1997). Research by Kaushanskaya et al., 2012 reported that phonological familiarity of pseudowords only mattered for familiar referents (e.g., a known image such as an elephant or banana rather than a novel image such as an alien) and when comparing bilingual to monolingual word learning, bilinguals only featured an advantage when unfamiliar phonological pseudowords were paired with familiar referents. However, the referenced word learning tasks involve mapping a novel pseudoword, generated from a novel or

unfamiliar phonology with a referent (either a ‘translation’ or an image). There is no processing or reasoning task accompanying the learning process.

Working memory span tasks start to more closely resemble Gf tasks as both concurrent storage and processing demands are required. Notably, though bilinguals and monolinguals perform the same on non-verbal versions (Engel de Abreu, 2011; Ratiu & Azuma, 2015; Sanchez et al., 2010), as these tasks involve more verbal content, the performance of bilinguals in non-dominant language begins to drop compared to monolinguals (Service et al., 2002; van den Noort et al., 2006; Vejnovic et al., 2010). However, as with word recall tasks, WM span tasks also diverge from verbal Gf assessments as they also feature the rapid and temporary presentation of stimuli.

The methods employed by Takano and Noda (1993; 1995) may be the most informative when attempting to understand how dominant or non-dominant language usage may impact bilingual performance on Gf tasks. Across mathematical or spatial reasoning tasks, participant performance was best when uninterrupted by language tasks, decreased when interrupted by language tasks in the dominant language of participants, and decreased the most when interrupted by a language task in the non-dominant language of participants. Interestingly, performance on the thinking task declined even when the thinking task consisted of a fairly simple mathematics task such as the addition of two two-digit numbers. The thinking tasks did not consist of briefly presented stimuli, nor was it particularly time-bound as it entailed completing approximately twelve simple addition problems within three and a half minutes. However, ultimately, the thinking task was non-verbal in nature, leaving the question of how the properties of pseudowords interact with linguistic status unanswered.

There was also one further area left unexplored by the research thus far. Namely, understanding the potential gaps in performance between verbal Gf items involving pseudowords and verbal Gf items involving real words. As it currently stands, there is no research comparing pseudowords to comparable real words. In other words, to better understand how pseudowords operate in a cognitive ability testing context, a useful reference point is that same context and items but involving real words. Verbal Gf items involving real words reintroduces some degree of contamination due to prior familiarity – namely the introduction of vocabulary (i.e., Gc) in a Gf test. Due to the Gf/Gc disparity in item content, one would expect the items utilizing pseudowords to behave differently than those featuring real words. The real words can also be calibrated to match the pseudowords on all of the properties discussed thus far – namely length and lexical neighborhood density to provide for a strong set of comparison stimuli. Additionally, as the current research was also concerned with how performance changes for individuals of differing linguistic backgrounds and that these two groups differ on vocabulary (Izawa, 1993; Oller & Eilers, 2002), it is believed that this prior familiarity may manifest as score differences across these linguistic groups. Thus to understand how pseudowords minimize contamination due to prior familiarity, the introduction of parallel items with real words is a key lever.

When reviewing the above literature, it should be noted again that differences in performance on cognitive tasks between monolingual and bilingual groups though significant tend to be small – typically manifesting as differences in milliseconds (Abutalebi et al., 2012; Costa et al., 2012; Gold et al., 2013; Schroeder & Marian, 2012), a WM capacity difference of around one word (van den Noort et al., 2006; Vejnovic et al., 2010), or a small percentage difference in words learned (Kaushanskaya & Marian 2009b; Kaushanskaya et al., 2012). In the Takano and Noda (1993; 1995) studies where thinking performance declined more so when

performing a concurrent linguistic task in a second or non-dominant language compared to a first or dominant language, while the total number of completed thinking items drops considerably, the error rates increase only marginally. However, in a context such as an employment testing context, small performance differences may have large consequences (Abelson, 1985).

Additionally, judging by the extant literature it appears that outside of cognitive tasks involving quick reactions or mapping new phonological forms to known semantic entries, monolinguals have an advantage compared to bilinguals using their non-dominant language. When this was considered alongside research suggesting cognition in the non-dominant language requires more resources, coupled with research described earlier in Chapter 4 suggesting that longer (pseudo)words and less wordlike (pseudo)words demand greater processing power, the following hypotheses were proposed:

Hypothesis 4a: Linguistic status will interact with word-type such that English-non-dominant bilinguals will perform significantly worse on items featuring real words compared to items featuring pseudowords.

Hypothesis 4b: Linguistic status will interact with word-type such that monolingual and English-dominant bilinguals will perform significantly better on items featuring real words compared to English-non-dominant bilinguals and on items featuring pseudowords.

Hypothesis 5: Linguistic status will interact with pseudoword length such that English-non-dominant bilinguals will experience a greater decline in performance on items involving long pseudowords compared to monolinguals or English-dominant bilinguals.

Hypothesis 6: Linguistic status will interact with pseudoword wordlikeness such that English-non-dominant bilinguals will experience a greater decline in performance on

items involving un-wordlike pseudowords compared to monolinguals or English-dominant bilinguals.

There are a few final points worth mentioning regarding the offered hypotheses.

Bilinguals generally suffer performance detriments compared to monolinguals on cognitive tasks that do not involve conflict monitoring and that are fairly time-bound as in to-be-remembered stimuli appeared and disappeared rather quickly or there was limited time to perform a cognitive task. Further, considering that 1) linguistic group differences are small on cognitive tasks to begin with, 2) that the pseudoword technique has been developed to and successfully does minimize group differences on cognitive ability tests (Fagan & Holland, 2009; Goldstein et al., 2009; Sternberg, 2006), 3) that in the proposed paradigm, pseudoword production (i.e., pronunciation) was not part of the task, and 4) that in the proposed paradigm, items did not have strict time limits, it was believed that English-dominant bilinguals will not fare significantly worse than monolinguals on the proposed verbal Gf items. While there are likely verbal reasoning tasks that exist where these groups may show performance differences, this current paradigm was not expected to be one of them, thus the relevant linguistic comparison groups were understood to be monolingual and English-dominant bilinguals compared to English-non-dominant bilinguals.

Chapter 6: Pilot Study One

The first of two pilot studies was conducted to gather data on how study design and verbal reasoning item structure influences item difficulty and participant strategy usage to identify possible pitfalls around using these verbal reasoning items, as well as early feedback on pseudowords. As the first pilot study was a preliminary look at how study/item features, apart from pseudowords, impact item difficulty, a basic type of item structure was used that allowed for easier permutations. Verbal true/false Gf items that contained a minimal amount of text were developed. Items featured either four or five pseudowords and participants were asked true false questions about relationships between pseudowords. Consider the example below:

Pseudoword 1 is darker than Pseudoword 2.
 Pseudoword 2 is lighter than Pseudoword 3.
 Pseudoword 4 is darker than Pseudoword 2.
 Pseudoword 4 is lighter than Pseudoword 3.

1. Pseudoword 4 is darker than pseudoword 1: True or False?
2. Pseudoword 3 is lighter than pseudoword 1: True or False?

These items involved only the presentation of pseudowords and a few words describing the relationship between the pseudowords – as mentioned, the minimal amount of text made it easier to create several item types that differed slightly as well as to disentangle the effects of item/study structure on participant performance.

Eight conditions each containing five item stems with four ensuing items for a total of twenty verbal reasoning items were developed. The items in each condition varied on a number of structural properties. The first was number of statements in the item stem– each statement compared one pseudoword to another, with some conditions having three item-stem statements and others having four statements. Following the item stems were four true or false questions about the relationships between pseudowords. The second structural property was the number of

pseudowords in the item stem. For conditions where three statements were presented, four pseudowords were present while conditions that had four statements in the stem, the first item featured four pseudowords while the remaining four stems featured five pseudowords. The number of pseudoword relationships in each statement varied between one or two (e.g., “pseudoword 1 is darker than pseudoword 2, and slower” as an example of a statement with two relationships). The items also differed on the number of response options. Most conditions had two, true and false, with conditions four and five adding ‘cannot be determined’ as the third response option.

Conditions 6 through 8 were fashioned after condition 1, but differed in unique ways. Condition 6 attempted to increase working memory load by presenting the item stems and the true/false items separately. As the pilot study was computerized, item stems and items were presented on different screens with no back button. Conditions 7 and 8 were concerned with minimizing the possibility that in an effort to reduce the cognitive demands of the items, participants would only memorize the first letter of each pseudoword. Condition 7 featured instructions saying that following completion of the twenty items, there would be a re-test on whether or not pseudowords appeared in the items, thus reducing incentive to only memorize the first letter. Following completion of the twenty items, participants were presented with several pseudowords, some of which had been included in one of the five item stems and some which had not and asked to indicate whether they appeared or not. Condition 8 attempted to eliminate the strategy of only memorizing the first letter by featuring two or more pseudowords per item stem that started with the same letter. Lastly, while items differed across conditions, each stem had two items that were repeated throughout all conditions to serve as way to isolate the impact

of the changes of the conditions on item difficulty. The features of the eight conditions and the mean and standard deviation of difficulty for the ten repeated items are presented in Table 1.

For pilot study one, 190 participants completed the study with four participants providing unusable data that were removed, for a final sample of 186. The informed consent, items, and debrief were presented on a computer. Item difficulty was operationalized as the percentage of participants who correctly answered the item, thus higher numbers reflect easier items. The mean difficulty of items across all conditions was .814 with a standard deviation of .191. The results by condition suggest a few findings. The first was that items with fewer statements and pseudowords were easier. Secondly, items with three response options were more difficult than only just true/false options. Third, the number of relationships in the item stem did not noticeably impact difficulty. Fourth, following completion of the verbal reasoning items, the pilot study included a question asking participants about their strategy usage. From the responses, the strategy of memorizing only the first letter of a pseudoword was sporadically, but not generally adopted. A notable exception was condition 6 where participants were much more likely to report using the first-letter strategy to memorize the relationships due to the items being on a different screen from the item stems. Apart from condition 6, this strategy was attempted across conditions at very low rates and it did not guarantee success on the items.

Overall, the results from pilot study one suggested that items for the main study should have four rather than three statements. Each item should have more than two or three response options to minimize potential ceiling effects, and that items need to appear on the same screen as the relationships to minimize unwanted alternative strategy usage by participants. Additionally, the number of relationships contained within a particular statement can vary between one and two and both the re-test component and having two pseudowords per item stem starting with the

same letter can be used to both potentially increase cognitive load and reduce unwanted alternative strategy usage. The pseudowords used in pilot study one and their properties are presented in Table 2.

Chapter 7: Pseudoword and Word Group Development

Pilot study one served as a way to inform the broader item structure but did not necessarily inform pseudoword creation. The true/false nature of the items led to a mean difficulty of .814 which was both too easy and obscured the influence of pseudoword properties on item difficulty. Further, the pseudowords in item stems in pilot study one were one, two, or three syllables. Since no concerning patterns were observed across different syllable-lengths and the current research endeavored to have a more stringent test of (pseudo)word length by featuring two conditions of length – one and three syllables, a new batch of pseudowords had to be created.

Pseudoword Development and Properties

Pseudoword Generation. There exist a handful of online resources for either generating pseudowords, getting the properties of pseudowords, or both. The current research used a combination of pseudowords generated by MCWord⁸ (Medler & Binder, 2005) as well as pseudowords generated by the author. MCWord is a free, online repository of lexical and sublexical information for words in the English language. In addition to providing lexical and sublexical information, MCWord also allows the user to generate pseudowords according to user-entered specifications (e.g., desired length) as well as varying approximations of English orthography (e.g. bigram vs. trigram pseudoword generation). The ‘generate pseudowords’ function of MCWord (referred to as ‘nonwords’ by the program) allows the user to specify the desired length by entering a range of number of letters and lets the user specify how pseudowords are to be generated. Options for this include consonant strings, random letter strings, and either constrained or unconstrained unigram-, bigram-, or trigram-based strings. The unigram, bigram, and trigram methodologies involve determining how often letter combinations appear in a lexicon on an averaged frequency-per-

⁸ MCWord is available for use online at: <http://www.neuro.mcw.edu/mcword/>

million basis of word usage (i.e., how often does the bigram ‘do’ appear in a word per million words). Unigram, bigram, and trigram refer to approaches that consider one, two, or three letters, respectively. A constrained bigram approach considers bigrams that only appear in the same position in words of the same length (e.g., the bigram frequency of ‘do’ in the word ‘dogs’ will factor in the word ‘does’ but not the words ‘idol’ or ‘double’) whereas an unconstrained approach does not consider where in the word a bigram appears nor the length of the word.

The proposed research employed both constrained bigram- and trigram-based string methods of generation as the other options produce a poor ratio of pronounceable to un-pronounceable letter sequences. The creation of the finalized lists of pseudowords involved generating approximately one-hundred thousand pseudowords. Following this, there was an iterative process of whittling this pool down based on the relevant pseudoword properties of length and wordlikeness (i.e., Levenshtein distance or lexical neighborhood density).

Pseudoword Length. The fact that the user can specify the letter-length of generated pseudowords aided in developing pseudowords of specific syllable-lengths. Shorter pseudowords were generated setting the string range from four to six letters while longer pseudowords were initially generated setting the string range from eight to ten letters. These ranges were based off of the work by New et al. (2006), who reported differences in reactions to words of different lengths as well as based off of practical constraints. Three-letter pseudowords were not considered as it is not possible to create pronounceable, three-letter pseudowords that have a sufficiently high Levenshtein Distance (LD; i.e., un-wordlike) and are not already a word. Indeed, although four letter pseudowords were generated, none were ultimately used for the same reasons. Due to difficulties around generating longer pseudowords that had low LDs (i.e., wordlike), the letter constraints were relaxed to include pseudowords with seven letters. While not ideal, and in contrast somewhat to the insights

provided by New et al. (2006), the number of letters is simply a proxy for the desired measure of pseudoword length which is the number of syllables.

Based on the work of Jalbert and colleagues (Jalbert et al., 2011a; Jalbert et al., 2011b), where no word length effect was found for pseudoword groups having a short/long difference of only one syllable, the proposed research endeavored to create short/long conditions that differed by more than one syllable. It was additionally believed that the length conditions used by Jalbert and colleagues, where length conditions not only differed by only one syllable but were all the same number of letters would be an insufficient test of (pseudo)word length in the present research. However, it is almost, if not completely impossible to create lists of pseudowords of one and four syllables that can then be matched on wordlikeness properties. Thus, ultimately for this research, a short pseudoword was one syllable and five to six letters in length and a long pseudoword was three syllables and seven to ten letters in length.

Levenshtein Distance. The LD calculator used in this research was created and provided by the Institute for Language and Speech Processing. The calculator is freely available online directly at: <http://speech.ilsp.gr/iplr/leven.rar> or from the parent web page: <http://speech.ilsp.gr/iplr/downloads.htm> and by scrolling down to the sub-section entitled “A C program that calculates orthographic and phonological distances” in the software sub-section. The DOS-command-prompt styled program allows the user to specify a text file containing the list of (pseudo)words for which LDs are desired, a text file that serves as the corpus against which the LDs will be calculated, and a text file name and path where the output file containing the LD values will be created. The file serving as the corpus contained 62,400 words ranging from three to twelve

letters. This program specifically calculates the orthographic Levenshtein distance of a target (pseudo)word's twenty closest neighbors, referred to as OLD20 (Yarkoni et al., 2008).⁹

Just as there does not appear to be consensus on the boundary between dense and sparse neighborhood densities, there exists no agreed-upon LD boundaries. An iterative process of pseudoword generation and subsequent LD determination was performed in an effort to understand the greatest amount of separation possible between low and high LD groups that would still allow for comparable pseudoword groups to be developed. For illustration, the differences in neighborhood density when moving from an LD of one to two to three are quite large. By definition, a word with an LD value of one will have at least twenty true neighbors and would be a common, highly-wordlike word whereas an LD value of three would indicate that the twenty closest neighbors of a word in the corpus are an average of three letter-transformations away. These words may have one true neighbor or, more commonly, zero neighbors.

When considering the LD of pseudowords ranging from five to ten letters in length that are still pronounceable as words in English, LD statistics range from one to about three-and-a-half (with shorter pseudowords almost never reaching this higher number while still remaining pronounceable). Given this, it was decided that LD groups of wordlike and un-wordlike should differ by an LD of about one. Ultimately, after several iterations, wordlike, high-density LD groups had a mean LD of 1.84 while un-wordlike, low-density LD groups had a mean of 2.76, a difference of 0.92. The wordlike LD group had a mean of 2.75 neighbors, ranging from one to nine while the un-wordlike LD group had a mean of 0.08, ranging from zero to one.

⁹ While the present research will use the OLD20 statistic as the measure of Levenshtein distance, the paper will refer to it using the acronym 'LD' to reduce terms.

Orthographic Neighborhood Frequency. As described earlier, neighborhoods of an equal size do not necessarily have equal implications for cognitive processes. In addition to providing neighborhood densities, MCWord also provides orthographic neighborhood frequencies. The orthographic neighborhood frequency is defined as the averaged frequency per million of a (pseudo)word's orthographic neighbors. As was the case with neighborhood density and Levenshtein distance, there is no agreed-upon boundary between high and low neighborhood frequency. Thus, a goal when generating pseudoword groups was to ensure that both short and long words of low LD (high density neighborhoods) had comparable mean neighborhood frequencies as well as short and long words of high LD (low density neighborhoods). Following the creation of pseudoword groups, high density LD groups had a mean lexical neighbor frequency of 4.12 while low density LD groups had a mean lexical neighbor frequency of 1.60.

Orthographic Probability. Among the various conceptualizations of sequence probability, a bigram/biphone conceptualization has been the most frequently used in research (Bailey & Hahn, 2001; Bartolotti & Marian, 2014). While reasons for this are not explained directly in research, judging from the iterative process used to develop the current pool of pseudowords, a unigram/uniphone approach is too lenient while a trigram approach is too restrictive when either generating pseudowords or creating matched groups of pseudowords that differ on some other property. Being the standard sequence probability-conceptualization used in research, this research also used a bigram conceptualization. Further, given that the proposed research consists of written/visually presented stimuli, it did not consider phonotactic probabilities. This is consistent with research suggesting that stronger sequence probability effects are found when they are modality-consistent (i.e., using orthographic probability for visually presented stimuli and phonotactic probability for auditorily presented stimuli; Bailey & Hahn, 2001).

MCWord, in addition to providing neighborhood and neighborhood frequency statistics also provides orthographic probability statistics. The form of orthographic probability selected for pseudoword classification was constrained bigrams due to it both facilitating easier pseudoword generation and being the preferred methodology/conceptualization in the extant literature (e.g., Bailey & Hahn, 2001). This means the orthographic probability values for each (pseudo)word consist of the summation of frequencies-of-occurrence for bigrams (i.e., two-letter pairings in a word – ‘dogs’ has the bigrams of ‘do’, ‘og’, and ‘gs’) in words, as opposed to unigram (one letter) or trigram (three letters) approaches. The fact that it is a constrained rather than unconstrained value means it only considers words of the same length and takes a relative bigram approach, meaning the bigram must be in the same position of the word. Thus, when considering the first bigram ‘do’, the word ‘does’ influences the bigram/orthographic probability for ‘dogs’ whereas the words ‘odor’ and ‘doorman’ do not. The current research utilized the constrained bigram values of orthographic probability. Following the creation of pseudoword groups, wordlike groups had a mean orthographic probability 1,231.14 while un-wordlike groups had a mean orthographic probability of 657.56.

Overall, four sets of stimuli were generated, with each set featuring twenty-four pseudowords. Specifically, short and wordlike (i.e., one syllable, low LD) pseudowords are presented in Table 3, short and un-wordlike (i.e., one syllable, high LD) pseudowords are presented in Table 4, long and wordlike (i.e., three syllables, low LD) pseudowords are presented in Table 5, and long and un-wordlike (i.e., three syllables, high LD) pseudowords are presented in Table 6. Additionally, the means of the properties for each pseudoword group are presented side-by-side in Table 9.

Word Group Development and Properties

As two hypotheses consider how the verbal Gf reasoning items function when the pseudowords are replaced with real words, matching sets of real words were created. As the current

study is primarily focused on the pseudoword technique and how alteration of their properties influence item characteristics across two linguistic groups of participants, the generation of real word groups was a secondary concern behind the generation of pseudowords. Thus, the parameters for word group construction were designed to mirror those for the sets of pseudoword, rather than the reverse. Additionally, real words have certain properties that pseudowords do not – namely, part of speech, the frequency with which the word appears in text – by definition this value is zero for pseudowords, and the concreteness (alternatively called ‘imageability’) of the word. As an example of concreteness, the noun ‘cat’ more readily comes to mind than the noun ‘theory’ and both of these nouns come to mind more readily than the adjective ‘isomorphic’. Thus, the real-word groups were designed to be matched to each other on these properties as well.

Word Generation. The part of speech of a word influences how easily that word is memorized with nouns being the part of speech that is most easily memorized (McDonough, Song, Hirsh-Pasek, Golinkoff, & Lannon, 2011). Additionally, given the present research’s focus on individuals for whom English is the non-dominant language, nouns are typically learned earlier and more quickly in second-language acquisition (Gentner, 1982; McDonough et al., 2011), thus the proposed research limited the pool of real words to nouns (including words which are not primarily used as a noun, but do feature a secondary usage as a noun, consider the word ‘haunt’ as an example). The total pool of words considered was the same pool of 62,400 words used to generate LD statistics. These words were run through www.easydefine.com (Fellbaum, 1998), a web service that provides definitions to lists of words and in the process identifies the part of speech of the words, to serve as the first step in whittling down the pool of over sixty-two thousand words.

Regarding word length, the parameters were nouns of one syllable, between four and six letters long, as well as nouns of three syllables between seven and ten letters long. One of the benefits

of pseudowords is that they can be created across a larger range of lexical neighborhood densities than real words. As evidence for this, it was impossible to create groups of short words that were un-wordlike (i.e., embedded in a sparse neighborhood) where the mean LD matched that of the short and un-wordlike pseudowords. Similarly, it was also impossible to create a group of long words that were sufficiently wordlike (i.e., embedded in a dense lexical neighborhood) where the mean LD matched that of the respective pseudoword set. Extant words in the English language at these mean lexical neighborhood densities for the short and long groups, respectively, either do not exist, do exist but are not a noun, or not enough of them exist to provide for a full condition. Ultimately, two sets of nouns were developed - short and wordlike words, and long and un-wordlike words. These two sets of words were matched to their respective pseudoword groups across not only length and lexical neighborhood density (defined as LD), but also the two word properties to be controlled - lexical neighbor frequency and orthographic probability.

Concreteness ratings for words were provided by Brysbaert, Warriner, and Kuperman (2014) which provided the concreteness ratings for forty-thousand generally known English-word lemmas with lemmas representing the root word for various inflected forms – as an example, the word ‘run’ is the lemma for ‘runs’, ‘running’, ‘ran’, etc. Out of the final pool of forty-eight words that were included in the present research, there were nine total words that did not feature a concreteness value – four in one condition, five in the other condition. Additionally, one word ‘decanter’ featured the concreteness value for the lemma ‘decant’. The mean concreteness value for the short, dense (i.e., wordlike) stimuli set was 4.11 while the mean concreteness value for the long, sparse (i.e., un-wordlike) set was 3.95.

As with pseudowords, each set of real words featured twenty-four words and their properties are presented in tables. Specifically, short, dense (i.e., wordlike) words are presented in Table 7 and

long, sparse (i.e., un-wordlike) words are presented in Table 8. Additionally, the word-group means appear alongside the pseudoword-group means in Table 9.

Chapter 8: Item Development

The first pilot study provided information that guided the calibration of the verbal Gf item type to be used in both pilot study two and the main study. From the first pilot study data, it appeared that verbal Gf items should contain more than four pseudowords, information critical to solving the items should not be separated from the actual items themselves (i.e., item stems should not appear on a different screen from the items), at least one pair of pseudowords should start with the same letter to minimize first-letter strategy usage, and items should feature at least three if not more response options as a mean overall difficulty of .814 suggested a potential ceiling effect. While the first pilot study informed several test and item design elements, it was not sufficient to identify the ideal verbal Gf item type. The current study's selection of a verbal Gf item format was further guided by several concepts drawn from extant item types as well as the literature. The first concept was that the items should be reflective of Gf verbal reasoning. Thus the cognitive operations needed to solve an item were not primarily visually or spatially based as in the Raven's Progressive Matrices or a mental rotation task nor should they ultimately depend upon numerical, mathematical, or other non-verbal abilities.

Second, an ideal verbal Gf item will be mostly to entirely non-entrenched. The concept of non-entrenchment broadly refers to item content being disconnected from real-world, crystallized knowledge (Sternberg, 1981), in the service of reducing contamination due to prior familiarity. While it is perhaps arguable that full non-entrenchment is impossible, extant item types do vary on the degree to which they incorporate real-world, crystallized knowledge. For instance, a non-entrenched item type proposed by Sternberg (1981) features rudimentary navigational concepts (i.e., the concepts of north and south, and the equator) while the graphical reasoning items in the Ravens Progressive Matrices present virtually no real-world or

crystallized knowledge. Ultimately, when non-entrenched items contain real-world information, this information should either be common knowledge to the intended pool of test-takers (i.e., rudimentary navigational knowledge may be a contaminant for young children but perfectly acceptable for adults), or somewhat incidental to correctly solving the problem (e.g., in the Sternberg (1981) item types, the test-taker can ostensibly answer the items correctly so long as they recognize that ‘north’ and ‘south’ represent different things even if they are unsure exactly what they represent).

Third, despite being a verbal Gf reasoning item, the item content could not be overly-verbose. Items were to remain a test of verbal Gf reasoning abilities with novel stimuli and not unintentionally operate as a test of long-form reading comprehension or as a test of vocabulary – with both of these item types being reflective of verbal Gc reasoning. A challenge with presenting novel information (i.e., pseudowords) in Gf reasoning items is that, depending on the item format, there is a need to contextualize the novel information somewhat, which can require an appreciable amount of verbiage. This is why the item types included in the work by Fagan and Holland (2009) and Sternberg (1981) contain moderate amounts of text. Thus, the Gf verbal reasoning items used in the present research did not mirror traditional verbal Gc reasoning items where a lengthy passage is presented with test-takers subsequently answering items about that passage.

Fourth, the item format and logic needed to easily accommodate the inclusion of pseudowords. Items structured for pseudowords to be easily interchanged with other pseudowords led to an efficient way to generate parallel test forms that only differed in the type of pseudowords included (e.g., long and un-wordlike vs. short and un-wordlike).

Fifth, the verbal Gf item format needed to involve reasoning with pseudowords as opposed to what is being termed reasoning around pseudowords. For instance, the item formats proposed by Fagan and Holland (2009) involved participants being trained on the meaning of pseudowords. This was accomplished by presenting short bodies of text that used pseudowords in sentences. Following the training period, participants were given questions that asked which real-world concepts the pseudowords were most like. While this is a rather useful method for measuring word learning, the ability to draw inferences, and/or verbal comprehension abilities while reducing contamination due to prior familiarity, the pseudowords themselves are almost incidental to correctly solving the problem – the critical information was embedded in the text with the pseudoword serving as a simple label. Hence this format involved maintaining pseudowords in memory to an extent but did not necessarily ask the participant to perform reasoning that directly involved the pseudowords. The condition of reasoning with pseudowords, rather than around them, also necessitated not providing participants with paper and pencil to ensure that all information presented was accommodated, manipulated, and solved in a purely cognitive manner.

Identifying an extant verbal Gf item type that met these criteria was somewhat difficult but was accomplished by utilizing item types frequently found in the Law School Admissions Test (LSAT). While the LSAT has a handful of item types, some of which involve tracing logic of arguments or legal knowledge, the LSAT also contains reasoning items that are more divorced from real-world, entrenched knowledge and issues. A LSAT item type that is typical of this last case appears as follows:

Passage for Question 1¹⁰

A university library budget committee must reduce exactly five of eight areas of expenditure — G, L, M, N, P, R, S, and W — in accordance with the following conditions:

1. If both G and S are reduced, W is also reduced.
2. If N is reduced, neither R nor S is reduced.
3. If P is reduced, L is not reduced.
4. Of the three areas L, M, and R, exactly two are reduced.

Question 1

If both M and R are reduced, which one of the following is a pair of areas neither of which could be reduced?

1. G, L
2. G, N
3. L, N
4. L, P
5. P, S

By replacing letters in the above example item with pseudowords, the result was a verbal Gf item that involved reasoning with pseudowords, as opposed to around them, and in a problem structure that did not feature excessive text (though it should be noted that this sample item was higher on vocabulary demands than what was considered ideal for the current research). This problem type also contains neither domain-specific knowledge nor does it require spatial, numerical, or mathematical abilities to solve. The LSAT item type also addressed several issues raised from the pilot study data as the item type could easily accommodate more than four pseudowords and items could easily have more than two or three response options.

Verbal Gf item skeletons were developed to mirror the LSAT items with a few modifications. Typical LSAT items feature six to eight pieces of information that test-takers need to sequence, but in the current research that number was reduced to five. Unlike the LSAT, the

¹⁰ Item presented is a public-domain item located on the Law School Admission Council website at the following web address: <http://www.lsac.org/jd/lsat/prep/analytical-reasoning>. Website last retrieved: 02/02/2016.

current research did not provide participants with paper or writing utensils and the problem-solving was exclusively cognitive. Based on the research showing that working memory (WM) can maintain about four chunks of information (Baddeley, 2012; Cowan, 2005; Cowan et al., 2012), item stems containing five pseudowords were developed, as sequencing those five pseudowords amounted to four chunks of information. The item types developed for this study were designed to feature less text than the typical LSAT items but were nonetheless lightly-contextualized for two primary reasons: 1) there is literature suggesting that non-native speakers may need or rely on contextual information more than native speakers when interpreting otherwise de-contextualized information (Carrell, 1984), and 2) the item context was employment-themed given the interest in the application of this strategy to employment tests. Unlike the LSAT where each item stem features a unique contextualization (e.g., an item about swimmers finishing a race followed by an item about seating arrangements at a dinner table), item stems for the proposed research were designed to align across the same context. This was done such that while some additional reading was required up front, the item stems contained a minimal amount of text for contextualization. The context provided to participants in the directions is “The questions you will be answering involve imaginary products at an imaginary store. You will be given rules about their relationships and will have to determine the ordering of their relationships”. Further text was provided at the start of the practice section that read “You will now be asked to read sentences describing relationships between different products. The products in the sentences will be imaginary products with nonsense words as names”. Four item stems were constructed with each stem featuring four items. Each item featured four response options, typically multiple choice, with one correct and three incorrect options. One item was

created as a ‘check all that apply’ item with only one correct solution. The item stems and items are presented in Appendix A.

The current research necessitated groups of pseudowords and real words that were essentially orthogonal by length and wordlikeness as well as an item structure that would allow for a cleaner examination of the influence of pseudoword properties on item difficulty and linguistic sub-group performance. Pilot study one informed item structure with the results suggesting a different item format (i.e., not true/false items) needed to prevent ceiling effects, item stems with four relationships and five pseudowords, and that presentation of item stems and item questions appear on the same screen to reduce unwanted strategy usage. Following pilot study one, groups of pseudowords were developed that were orthogonal on length and wordlikeness - operationalized as Levenshtein distance (while being matched on other properties such as lexical neighborhood frequency, part of speech, and concreteness). Lastly, a new item type needed to be developed – with the verbal reasoning items from the LSAT being modified to combine insights from the literature, desirable features of extant verbal Gf item types, and accommodating the structural insights from pilot study one.

Chapter 9: Pilot Study Two

Prior to the main study, a second pilot study was conducted to troubleshoot the proposed LSAT-styled item stems and items, and assess their difficulty. For the second pilot study, participants were seated at a computer where the informed consent, directions, practice items, items under consideration, and demographic questions were presented. Two versions were created, with each condition being identical except the order of presentation of the item stems was changed as well as the pseudowords within that particular item stem and items. The total number of participants across both versions was sixty. Item difficulty was operationalized as in the first pilot study as the percentage of participants who correctly answered the item. The overall mean difficulty of all items was 0.507. This number is in the mid-range of the difficulty scale suggesting that these item types are neither too hard nor too easy for participants. This number also suggests there will be sufficient variability in the data and that there should be no ceiling or floor effects. Items associated with item stem one produced a mean difficulty of 0.588, items for item stem two produced a mean difficulty of 0.483, items for item stem three produced a mean difficulty of 0.342, and items for item stem four produced a mean difficulty of 0.617. Item stem two featured an item with a mean difficulty of 0.083 which was examined and replaced for the main study. The mean difficulty for item stem three of 0.342 is lower than ideal. This item stem will be examined to determine the drivers of this low mean – particularly the second item associated with item stem three as participants scored below chance (i.e., 25%). The fourth item associated with item stem four will also be examined and modified/replaced as participants scored below chance. The mean difficulties for item stems overall and individual items are presented in Table 10.

Regarding participant strategy usage, as with condition 8 in pilot study one, each item stem in pilot study two contained two pseudowords that began with the same letter. None of the sixty participants who completed the second pilot study reported memorizing only the first letter to aid problem-solving. Only two participants, representing three percent of the data, did report attempting to shorten the pseudowords. This suggests that the current item structure is sufficient for mitigating strategies that influence the ability to accurately assess the impact of pseudoword length on item difficulty.

Readability. Given that the items under consideration were designed to be administered across participants with varying levels of English proficiency and that a central goal of item generation was to remain a verbal Gf item while not featuring a considerable amount of text nor being overly verbose, readability analyses were performed across the items to determine their reading difficulty. The website <https://readability-score.com> provides readability statistics across a number of readability conceptualizations. The current research considered two conceptualizations: the Flesh-Kincaid grade level index and the SMOG index.

The Flesch-Kincaid grade level index is a measure that considers the ratio of total words divided by total sentences and the ratio of total syllables to total words in a given text. These values are multiplied by constants and added together to achieve a score that approximates the grade-level of a text. For instance a Flesch-Kincaid grade level of ‘7’ indicates that the difficulty of the analyzed text is commensurate with a seventh-grade reading level.

The SMOG index also provides grade-level output by considering the complexity of sentences and the number of polysyllabic words (i.e., words of three or more syllables) in a given body of text. Unlike the Flesch-Kincaid index, the SMOG index requires at least thirty sentences,

making it unsuitable for determining readability at the item stem or item levels. However, it was applied to the instruction and practice-question text.

For each index computed, only the item skeleton was entered. Specifically, each item stem with the relationships as well as the questions were entered. Possible response options (i.e., “a. Boust, Trawns, Brench, Knills, Murnt”, the first response option of problem 1 in Appendix A) were removed. Additionally, all pseudowords were removed – as the purpose of the present research is to consider pseudoword length effects, this creates an issue since each grade-level index considers syllables. By removing pseudowords, the following grade level indices provide the grade-level difficulty of the item skeletons. Following the readability analyses, the Flesch-Kincaid grade levels were 6.9 for item stem 1, 5.0 for item stem 2, 5.8 for item stem 3, and 2.8 for item stem 4. All of the item stems featured a grade level at or below eighth grade suggesting that there were no excessive language demands with the LSAT-styled item stems and items. The readability grade levels for the directions and practice items (with pseudowords and answer stems removed) was 6.8 for the Flesch-Kincaid and 10.3 for the SMOG index for a mean grade level of 8.6. As with the item stems, the directions and practice items featured a grade level at or below eighth grade. Overall, the reading level of the instructions, practice items, and verbal Gf item stems fell within the reading ability of the intended college sample.

Chapter 10: Methods

Operationalizing Bilingualism

As mentioned in Chapter 5, there exists great ambiguity regarding how to classify bilinguals, with differences in language combinations and their (dis)similarities, age of acquisition, and proficiency presenting something of a moving target for researchers. Given the lack of precision in the terms ‘bilingual’ or ‘English as a second language’, it is not believed that anyone identifying as simply bilingual or having English as their second language is stringent enough. Or put differently, it is not believed that the proposed changes to pseudowords would manifest as performance differences on verbal Gf items between native speakers of English and ESL individuals who have been speaking English primarily since entering school. Thus, the bilingual comparison group was defined as anyone who is bilingual and reports a language other than English as their dominant language, and/or reports having started learning English at or after the age of ten.

Main Study

The main study was concerned with three related phenomena. The first phenomenon was examining the impact that varying two (pseudo)word properties, length and wordlikeness (operationalized as Levenshtein distance), had on item difficulty. For greater interpretability, two additional pseudoword and word properties were controlled – lexical neighbor frequency and orthographic probability, as these strongly co-vary with length and lexical neighborhood density (i.e., wordlikeness). These variations were addressed by hypotheses 1, 2, and 3 with hypothesis 1 examining the impact of length, hypothesis 2 examining the impact of wordlikeness, and hypothesis 3 examining the impact of both length and wordlikeness on item difficulty. The second phenomenon was the interaction between stimuli type – novel or not novel

(operationalized as pseudowords and real words) and linguistic sub-group on verbal Gf performance. Hypothesis 4a examined the influence of word type (pseudo vs. real) on non-English-dominant participant performance while hypothesis 4b examined how linguistic sub-groups performed across both pseudo and real words. The last phenomena was the interaction between pseudoword properties and linguistic subgroup performance. Hypothesis 5 examined how pseudoword length influenced performance for English-dominant compared to English-non-dominant speakers while hypothesis 6 did the same with wordlikeness instead of length. The dependent variable for all analyses was performance on the verbal GF reasoning items. For hypotheises 1, 2, 3, 5, and 6, the dependent variable was participant total score. For hypotheses 4a and 4b, rather than total score across all verbal Gf items, the dependent variables were split apart by word type - items featuring real words and items featuring pseudowords.

Participants

Participants were sampled from a large, northeastern public university. A power analysis was conducted to determine the required sample size. The main study featured four pseudoword conditions (short, wordlike; short, un-wordlike; long, wordlike; and long, un-wordlike) and two of real words (short, wordlike and long, un-wordlike), though these two real word conditions were a within-person design. Further, each condition needed a sufficient number of English-dominant and English-non-dominant language participants. Given the smaller group differences associated with the Gf item types to be tested along with the modest differences in performance between linguistic groups described in Chapter 5, the anticipated effect size was small to medium. Following a power analysis, the planned sample size was approximately forty participants per condition for a total sample size of two hundred participants (or approximately twenty participants per linguistic subgroup, per condition).

Manipulations

Following the iterative process of pseudoword generation described in Chapter 7, orthogonal sets of pseudowords were developed. Regarding length, all short pseudowords were one syllable long. As an additional check on length, the mean number of letters per pseudoword was considered and short wordlike and short un-wordlike conditions had a mean pseudoword length of 5.58 and 5.79 letters, respectively. Long wordlike and long un-wordlike pseudoword sets each consisted of pseudowords that were three syllables and had a mean length of 8.17 and 8.54 letters, respectively. For wordlikeness, short wordlike and long wordlike stimuli sets had a mean LD of 1.81 and 1.87, respectively. This indicated that for both sets of wordlike pseudowords, the nearest twenty real words for each pseudoword were on average 1.8 letter-transformations away. For short un-wordlike and long un-wordlike pseudoword groups, the mean LD was 2.76 and 2.77, respectively, indicating that the nearest twenty neighbors for each pseudoword were on average approximately 2.7 letter-transformations away.

The two conditions of real words were designed to match their respective pseudoword conditions on both length and LD. The short wordlike condition for real words featured a mean length of 5.50 letters and a mean LD of 1.80. The long un-wordlike condition of real words featured a mean length of 8.38 letters and a mean LD of 2.79.

Regarding the first pseudoword property that was controlled, lexical neighbor frequency, the short wordlike and long wordlike stimuli sets had means of 5.74 and 2.51, respectively. For clarification, this indicated that for the short wordlike set, the lexical neighbors (i.e., real words) of the pseudowords, occurred on average 5.74 times per million words.¹¹ The mean lexical

¹¹ MCWord derives its frequency from the CELEX database, another repository of word properties. In order to determine lexical neighbor frequency, representative samples of written and spoken text were gathered which amounted to 17,900,000 instances of word use.

neighbor frequency of short un-wordlike and long un-wordlike pseudoword groups was 0.51 and 2.70, respectively. The value of 2.70 for long un-wordlike groups was higher than the value for short un-wordlike groups, as well as higher than the value for long wordlike groups. However, this was due to one pseudoword ‘spriously’ having one neighbor that occurs very frequently (64.85 times per million; the lexical neighbor ‘seriously’). All other pseudowords in this set had no neighbors and thus a lexical neighbor frequency of zero. Given that it was only one pseudoword, it was not believed that the mean value placed the long un-wordlike stimuli set in a similar tier of wordlikeness as the long wordlike set regarding the frequency of pseudoword neighbors. Furthermore, the lexical neighborhood density (i.e., the number of neighbors) for the long wordlike set ranged from one to six with a mean of 2.79 whereas the long un-wordlike set ranged from zero to one with a mean of 0.08, suggesting the two sets were sufficiently disparate. The stimuli sets containing real words that were short wordlike and long un-wordlike featured a mean lexical neighbor frequency of 5.31 and 0.04, respectively. These values matched their respective pseudoword stimuli set.

The last word property to be controlled was orthographic probability. As a refresher, the conceptualization of orthographic probability used in the present research was constrained bigrams. Bigram probabilities consist of the frequency with which two-letter strings occur in words (on a per-million rate), with each of these values being summed to form a word’s orthographic probability. Alternatives include unigram and trigram approaches, though bigram approaches are the preferred approach (e.g., Bailey & Hahn, 2001). The bigram approach was also constrained rather than unconstrained meaning the frequencies are calculated for that word considering only the relative position of the bigram (e.g., ‘do’ in ‘dogs’ is the first bigram) and only in words of the same length. The mean orthographic probabilities for the short wordlike and

long wordlike conditions were 1,165.08 and 1,297.21, respectively. The mean orthographic probabilities for the short un-wordlike and long un-wordlike conditions were 506.40 and 808.72, respectively. While the last two values were more disparate than ideal, this disparity was unsurprising given that in order to achieve sufficiently high LD statistics for the short un-wordlike pseudowords, more atypical bigrams were featured in the spelling of the pseudowords. Regarding the two sets of real words, short wordlike and long un-wordlike conditions had mean orthographic probabilities of 1,013.62 and 484.36, respectively. The former value matched its respective pseudoword set whereas the latter value of 484.36 was considerably lower than the 808.72 mean orthographic probability of the long un-wordlike pseudowords. Given that orthographic probability has been found to be subordinate to lexical neighborhood density in terms of impacting cognitive processes, these disparities were not expected to influence item performance above and beyond lexical neighborhood density. However, they were too large to comfortably be ignored, thus orthographic probability was entered as a covariate when examining item performance. Given that all of the other properties were either matched or orthogonal, no other properties were entered as a covariate or explicitly controlled for in the analyses. All of the pseudoword and real word stimuli-set properties are listed in Table 6 and the overall means of the six sets of stimuli are presented in Table 9.

Demographic and Control Measures

The current research featured two variables that were controlled across participant groups (as opposed to orthographic probability, which was controlled for across stimuli groups) – the vocabulary of the participant and the socioeconomic status (SES) of the participant. As described earlier, a central challenge with verbal Gf is the introduction of language which creates a scenario where it is difficult to disentangle reasoning ability from linguistic knowledge (or Gf

from Gc). Given that the current research is interested in how manipulations to pseudowords impacts reasoning across different sub-groups of test-takers, vocabulary was included as a control variable to help parse out a participant's linguistic proficiency allowing for a cleaner interpretation of how reasoning ability was impacted. Regarding SES, early research on bilingualism's impact on cognitive ability reported that bilingualism negatively influenced cognitive ability (see Grosjean, 1989; Peal & Lambert, 1962). The belief was generally that learning a second language either interfered with learning the first language or created interference when reasoning (Grosjean, 1989; Peal & Lambert, 1962). However, a central flaw in this research was sampling. Peal and Lambert (1962) identified that the monolingual and bilingual groups used in research were not always matched on other characteristics associated with performance on cognitive ability tests – namely SES. The result of matching monolingual and bilingual groups on SES led to findings that suggested there was no cognitive ability penalty associated with bilingualism.

Vocabulary. The vocabulary measure used was the Verbal Ability assessment from the O*NET Ability Profiler. The Verbal Ability assessment is a public-domain assessment that includes nineteen synonym/antonym questions to be completed within an eight-minute timespan. Specifically, each question featured four words and test-takers were asked to identify which two words are most nearly the same in meaning or which two words are most nearly the opposite in meaning. Two practice items from the O*NET Ability Profiler Verbal Assessment are presented below to demonstrate a synonym and antonym item, respectively:

Practice Item 1:

- A. big
- B. large
- C. dry
- D. slow

Practice Item 2:

- A. witty
- B. sad
- C. tired
- D. happy

The vocabulary measure was included as a control to understand performance differences between linguistic sub-groups as well as to see if these differences extended beyond just basic knowledge of the English language lexicon.

Linguistic background measure. The linguistic background of participants was assessed by several questions. The first question assessed if a participant is bilingual or not. If they indicated they were bilingual, there were several follow up questions that assessed the age when a participant began learning English as well as which language they felt more comfortable using. The full linguistic background measure is included in Appendix B.

Socioeconomic Status. Given that early research involving bilinguals produced unreliable results due to failure to control for participants' socioeconomic status (see Peal & Lambert, 1962), the present research asked questions to determine a participant's SES. The SES questions were included as part of the demographics section, included at the end of the study. They are available in Appendix C.

Demographic measure. Participants were asked to self-identify themselves across a range of demographic characteristics including age, race/ethnicity, and gender. These characteristics, the question, and the response options are presented in Appendix D.

Procedure

The current study was administered via computer. Participants were randomly assigned to conditions. To counterbalance the conditions, the order that the item stems were presented was randomized and each of the six conditions had two versions that were identical except the pseudowords in the stem were changed so that subset of pseudowords did not appear in one item stem exclusively. Upon entering the research lab, participants were seated at a computer and the experimenter or an assistant provided a brief description of the study. Following this, the

participants completed an informed consent form, presented on the computer. Once the informed consent form was completed, participants began the study. The first part of the study consisted of the presentation of the four LSAT-styled item stems and their corresponding sixteen verbal Gf items involving pseudowords or words. The six computerized conditions were as follows: 1) short pseudowords, wordlike, 2) short pseudowords, un-wordlike, 3) long pseudowords, wordlike, 4) long pseudowords, un-wordlike, 5) short and wordlike real words and pseudowords, and 6) long and un-wordlike real words and pseudowords. Prior to beginning the sixteen verbal Gf items, three practice items were presented to ensure participants were familiar with the items as well as the logic governing their completion. Once the verbal Gf items were completed, participants began an incidental re-test of pseudoword memory which consisted of the presentation of a list of pseudowords that were and were not included in the verbal Gf questions and asking participants whether or not each pseudoword in the list was included in the previous questions. This was done as a further check to explore any linguistic sub-group differences. To discourage first-letter strategy usage, each set of pseudowords in an item-stem included two pseudowords that began with the same letter. Data from the pilot study suggested that the duplicate initial-letter feature was an effective deterrent to this possible strategy. This section also asked participants to provide information including rating how much the pseudowords resembled real words in English and what strategies they used to answer the verbal Gf questions.

The next section consisted of participants completing the nineteen vocabulary questions from the O*NET Ability Profiler within eight minutes as well as the linguistic background, SES, and demographic questions, presented in Appendices B, C, and D, respectively. The linguistic background measure assessed participants' linguistic background and specifically asked if they speak a language other than English, and if so what additional language(s) do they speak, if

English is their first or second language, if English is their second language, at what age did they begin to learn English, how many years they have been speaking English, which language they feel is their dominant language, and lastly, how often English is spoken in their home. The end of the demographic section represented the end of the study, after which participants were thanked and debriefed.

Analyses

Hypotheses 1 and 2 which assessed the impact that pseudoword length and wordlikeness have on item difficulty, respectively, were assessed via analysis of covariance (ANCOVA). Due to the orthogonal nature of both the pseudoword and real word conditions, length and wordlikeness were categorical variables. Due to orthographic probability not being as orthogonal across conditions as desired, it was entered as a covariate in the ANCOVA. The respective word property (i.e., length or wordlikeness (LD)) was entered as the independent variable. With overall performance (defined as percentage correct) across the sixteen items as the dependent variable, Hypotheses 1 and 2 were considered supported if the results were in the intended direction and the F statistic was significant at the $p < .05$ level.

Hypothesis 3 posited that pseudoword properties would interact such that the condition featuring long, un-wordlike pseudowords would be the most difficult condition. Hypothesis 3 was tested via t-test where the mean performance for the other three conditions was compared to this condition. This hypothesis was considered supported if the t statistic was significant at the $p < .05$ level.

Hypothesis 4a and 4b considered performance differences between English-dominant and English-non-dominant participants across verbal Gf items involving pseudowords and items involving real words. Given that only two sets of real words were able to be created, the structure

of the study changed. Rather than complete all four item stems each with a specific type of pseudoword (e.g., long, un-wordlike), item stems alternated between pseudowords and real words while adhering to that word-property condition. Conditions were counterbalanced so that specific pseudowords or real words did not always appear in the same stem or the same stem-position. To illustrate, consider the breakdown presented below for condition 6:

Item stem sequence	Word type	Condition
Stem 1	Pseudowords	Long, un-wordlike
Stem 2	Real words	Long, un-wordlike
Stem 3	Pseudowords	Long, un-wordlike
Stem 4	Real words	Long, un-wordlike

Following completion of the sixteen items, scores were computed for each half of the assessment – pseudowords and real words. For hypothesis 4a, only the performance of non-dominant English speakers was considered with word type (real or pseudo) entered as the independent variable. Hypothesis 4a was assessed via t-test and results were considered supported if the performance of the real word condition was lower than the pseudoword condition and the *t* statistic was significant at the $p < .05$ level. Hypothesis 4b considered the difference in performance between dominant English speakers and non-dominant English speakers on items involving real words. Hypothesis 4b was assessed via ANCOVA with the control variables of vocabulary and SES entered as covariates. The condition codes for linguistic group with English-as-dominant coded as ‘0’ and English-as-non-dominant coded as ‘1’ were entered as the independent variable. Hypothesis 4b was considered supported if the English-dominant group performed better than the English-non-dominant group and *F* statistic for linguistic group was significant at the $p < .05$ level.

For hypotheses 5 and 6, the impact that linguistic status had on the change in performance as word properties change was examined. Specifically, hypothesis 5, posited that while both English-dominant and English-non-dominant speakers would experience a decline in performance as pseudoword length moved from short to long, this decline would be larger for non-dominant English speakers. Hypothesis 6 was similar except it considered the impact of linguistic performance on the decline in performance as pseudowords moved from wordlike to un-wordlike. Both hypotheses were assessed via ANCOVA in which performance as entered as the dependent variable, orthographic probability, SES, and vocabulary were entered as covariates, pseudoword condition (either length or wordlikeness) and linguistic subgroup entered in the second step, and the interaction term of pseudoword condition \times linguistic group entered as the third step. Hypotheses 5 and 6 was considered supported if the results are in the intended directions and the F statistic for the pseudoword condition \times linguistic group interaction term was significant at the $p < .05$ level.

Chapter 11: Results

Data Cleaning. Overall, 411 participants completed the study, this number exceeded the estimated number of participants due to there being considerably fewer English-non-dominant participants in the sample. Prior to analyses, data were checked to ensure that participants completed the study, took the study seriously (i.e., took at least twenty minutes total), and did not use a strategy to solve the reasoning problems that would invalidate their responses (e.g., a strategy that involves truncating pseudowords – using the first letter or assigning numbers to pseudowords as examples). A final sample of 368 participants remained following removal of participants.

Demographics. The final sample was 51.9% female with 1.4% of participants electing not to report gender. Regarding race/ethnicity, 23.4% of the sample identified as White, non-Latino or Hispanic, 12.0% as Black, non-Latino or Hispanic, 16.0% as Latino or Hispanic, 32.1% as East-Asian, 14.7% as South-Asian, 0.3% as Native Hawaiian or Pacific Islander, 0.5% as Native American or Alaskan Native, and 6.3% as Other. The mean age for participants was twenty-one years of age and ranged from eighteen to forty-three. English-dominant participants were slightly younger than English-non-dominant participants with each group having a mean age of 20.77 and 22.05 years, respectively.

Regarding dominant language spoken, 59.8% of participants were coded as English dominant, while 40.2% of participants were English-non-dominant (i.e., participants who chose a language other than English as the one they felt most comfortable speaking or participants who reported learning English at or after the age of ten). Including English, participants reported speaking a total of fifty different languages with the most common being Chinese and Spanish, after English.

Control Variables. The vocabulary measure adapted from the O*Net ability profiler featured nineteen questions and participants were given a point for each correct answer. The mean score on the vocabulary measure was 8.98. The mean vocabulary score for English-dominant participants was 9.99 while the mean score for English-non-dominant participants was 7.32. This difference in vocabulary scores between linguistic groups was statistically significant following a t-test ($t = -10.62, p < .05$).

The SES measure was a composite measure of three of the four questions in Appendix C. Question 3 in Appendix C which asks participants if they or anyone in their household owns various electronic devices was scored such that participants received a point for answering ‘yes’ to any of the categories. Thus scores on question 3 had a possible range of zero to four. Following aggregation of this question, all SES questions were significantly correlated with each other with the exception of question 2a and the total for question 3. A reliability analysis was performed which yielded a Cronbach’s alpha of .570. Only the removal of question 3 led to an improved alpha of .614. A factor analysis was then performed to extract a component from the SES questions. Question 3 was not included in this process and the resulting extracted component was used as the SES variable. Linguistic groups were found to differ on SES with English-dominant participants reporting a higher SES than English-non-dominant participants. This difference was found to be significant via t test ($t = -6.91, p < .05$).

The two control variables of vocabulary and SES were significantly correlated with each other ($r = .225, p < .05$). Though the correlation was significant at the $p < .05$ level, including both of these variables in further analyses was warranted as they are logically related to the outcome of interest (total score on the verbal Gf questions, Peal & Lambert, 1962), conceptually distinct, and the correlation was not high enough to suggest redundancy. Additionally, both

control variables were significantly correlated with the total score on the verbal Gf questions at the $p < .05$ level (vocabulary: $r = .273$; SES: $r = .108$). Lastly, mean differences in the control variables across conditions were explored. There were no significant differences nor notable discrepancies across conditions. The means and standard deviations for the control variables across conditions and overall are presented in Table 11.

Data Preparation. For hypotheses 1 through 3, the dependent variable was the total score across the sixteen verbal Gf reasoning items. Item responses were recoded as '1' or '0' for correct or incorrect, respectively and total scores were created by summing the recoded scores across the sixteen items. Prior to conducting analyses, the distribution of the total score was examined. The overall distribution for total score was normal and following Levine's test of equality of error variances, which was not significant, it was determined that the distribution of total scores for English-dominant and English-non-dominant participants were equal. The mean difficulty for each item and the total score overall and by linguistic sub-group are presented in Table 12. Item difficulties ranged from .247 to .897 and the reliability of the scores on the items was $\alpha = .762$. The total score had a mean of 8.89 and a standard deviation of 3.52. The means and standard deviations for the total score across conditions and overall are presented in Table 11.

For hypotheses 4a through 6, subsets of the total score were calculated. These hypotheses were concerned with performance on items involving real words and pseudowords and was explored via a within-person design. Participants in these conditions received sixteen items, half of which had real words and half of which had pseudowords, thus each total score subset is out of eight, rather than sixteen. These two scores are referred to as real total and pseudo total in the ensuing sections.

Hypothesis Testing

Hypothesis 1 posited that items containing longer pseudowords will be more difficult than items containing shorter pseudowords when controlling for orthographic probability. An ANCOVA was performed with total score as the dependent variable, pseudoword length as the independent, and orthographic probability entered as the covariate. Following the ANCOVA, the effect of pseudoword length on total score was not significant ($F = .703, p = .403, df = 238, \eta^2 = .003$). Additionally, the results were not in the intended direction as the mean total score for items involving longer pseudowords (9.46) was higher than the mean total score for items involving shorter pseudowords (8.91). This outcome, that items involving shorter pseudowords are more difficult than those involving longer pseudowords is returned to in the discussion section. Thus, Hypothesis 1 was not supported.

Item-level results were also examined to see if individual items were sensitive to changes in word properties. First a t-test was performed across all items excluding data collected from conditions with real words. Length was found to be significant for only one item (item 11, $t = -2.974, p < .05, df = 231.4, d = .385$). A MANCOVA was performed including the covariates of participant vocabulary and SES to further investigate any patterns. As with the t-test, the MANCOVA was performed at the item level excluding conditions involving real words. Two items featured significant differences in performance based on pseudoword length (item 3, $F = 4.246, p < .05, df = 1, \eta^2 = .018$; item 11, $F = 11.298, p < .05, df = 1, \eta^2 = .046$). As with total score overall, for both items 3 and 11, performance was better when items featured longer rather than shorter pseudowords. There were five items where performance was better when those items featured shorter pseudowords. However, these differences were small and not significant.

Hypothesis 2 posited that items containing un-wordlike pseudowords will be more difficult than items containing wordlike pseudowords when controlling for orthographic probability. An ANCOVA was performed with total score as the dependent variable, pseudoword wordlikeness as the independent, and orthographic probability entered as the covariate. Following the ANCOVA, the effect of pseudoword wordlikeness on total score was not significant ($F = .124, p = .725, df = 238, \eta^2 = .001$). Unlike length, the results were in the intended direction with the mean total score for wordlike pseudowords (9.45) being higher than the mean total score for items involving un-wordlike pseudowords (8.95). Thus, Hypothesis 2 was not supported.

As with length, the influence of wordlikeness was also examined at the item-level. A t-test was performed across all items excluding data collected from conditions with real words. Wordlikeness was found to significantly impact item performance for one item (item 6, $t = 2.188, p < .05, df = 235.4, d = .283$). A MANCOVA was performed at the item level including the covariates of participant vocabulary and SES to further investigate any patterns. Only item 6 featured a significant difference in performance based on pseudoword wordlikeness ($F = 5.013, p < .05, df = 1, \eta^2 = .021$). As with total score overall, the difference in performance for item 13 was in favor of items with wordlike rather than un-wordlike pseudowords. There were six items where performance was better when those items featured un-wordlike pseudowords. However, as with length these differences were very small and not significant.

Hypothesis 3 stated that pseudoword properties would interact such that items featuring both long and un-wordlike pseudowords will be more difficult than other pseudoword combinations. For this hypothesis, a t-test was performed comparing the mean score of condition 4 (i.e., long, un-wordlike pseudowords) to the mean score for conditions 1, 2, and 3. Following a

t-test, hypothesis 3 was not supported as the t-test was not significant ($t = .208, p = .835, df = 237, d = .031$). Additionally, though results were in the intended direction, performance was essentially equal between groups with the mean of condition 4 (9.11) being slightly lower than the combined mean of conditions 1, 2, and 3 (9.22). This finding makes sense in light of the results of hypothesis 1 where items involving longer pseudowords were easier than items involving shorter pseudowords.

Hypothesis 4a posited that linguistic status will interact with word type such that English-non-dominant participants will perform significantly worse on items featuring real words compared to pseudowords. A paired samples t-test was performed to compare the total score on items involving real words to the total score on items involving pseudowords for English-non-dominant participants. Following the t-test, hypothesis 4a was not supported ($t = -.714, p = .479, df = 43, d = .101$) as English-non-dominant participants achieved only a slightly lower score on items involving real words (3.52) compared to items involving pseudowords (3.70).

Hypothesis 4b stated that linguistic status will interact with word type such that English-dominant participants will perform significantly better on items featuring real words compared to English-non-dominant participants, while controlling for vocabulary and SES. An ANCOVA was performed with real score entered as the dependent variable, linguistic group as the independent variable, and vocabulary and SES entered as covariates. Following the ANCOVA, hypothesis 4b was not supported ($F = 1.632; p = .204, df = 123, \eta^2 = .013$). Results were in the intended direction with English-dominant participants achieving a higher mean score than English-non-dominant participants, 4.64 to 3.52, respectively. Also of note is that the covariate of vocabulary was significant ($F = 4.087, p = .045, \eta^2 = .033$); while the covariate of SES was not ($F = .916, p = .340, \eta^2 = .008$).

Hypothesis 5 stated linguistic status will interact with pseudoword length such that English-non-dominant participants will experience a greater decline in performance as pseudoword length moves from short to long compared English-dominant participants. Hypothesis 5 was tested via ANCOVA with participants' total score entered as the dependent variable, participant linguistic group and pseudoword length entered as independent variables, and orthographic probability, vocabulary score, and SES entered as covariates. Following the ANCOVA, hypothesis 5 was not supported as the interaction term between participant linguistic status and pseudoword length was not significant ($F = .660, p = .417, df = 238, \eta^2 = .003$). Of note is that the covariate for vocabulary was significant ($F = 17.870, p < .05, \eta^2 = .072$).

Hypothesis 6 stated that linguistic status will interact with pseudoword wordlikeness such that English-non-dominant participants will experience a greater decline in performance as pseudoword length moves from wordlike to un-wordlike compared English-dominant participants. Hypothesis 6 was tested via ANCOVA with participants' total score entered as the dependent variable, participant linguistic group and pseudoword wordlikeness entered as independent variables, and orthographic probability, vocabulary score, and SES entered as covariates. Following the ANCOVA, hypothesis 6 was not supported as the interaction term between linguistic status and pseudoword wordlikeness was not significant ($F = 3.303, p = .070, df = 238, \eta^2 = .014$). As with hypothesis 5, the covariate for vocabulary was significant ($F = 18.123, p < .05, \eta^2 = .072$). Interestingly, the mean performance for English-non-dominant participants was approximately the same for items involving wordlike and un-wordlike pseudowords at 8.66 and 8.67, respectively. However, the mean performance for English-dominant participants decreased considerably when moving from items involving wordlike to items involving un-wordlike pseudowords, at 10.24 to 9.12, respectively. Following a t-test, this

difference did not reach the traditional threshold for statistical significance ($t = 1.957, p = .052, df = 133, d = .343$).

Item-level analyses were conducted to further explore the influence of participant linguistic subgroup on performance. A t-test was performed across all items excluding data collected from conditions with real words. Participant linguistic status was found to be significantly related to performance for two items (item 3, $t = -3.458, p < .05, df = 226.6, d = .449$; item 9, $t = -2.206, p < .05, df = 220.7, d = .288$). A MANCOVA was performed at the item level including the covariates of participant vocabulary and SES to further investigate any patterns. After including the controls, only item 19 featured a significant difference in performance based on linguistic status ($F = 5.029, p < .05, df = 1, \eta^2 = .021$). Linguistic group differences on the two items with significant differences via t-test (items 3 and 9) were no longer significant. As with total score overall, the difference in performance for items 3 and 9 was in favor of English-dominant participants. Interestingly, performance was better for English-non-dominant participants on item 10. Though the difference in performance when not controlling for any variables were miniscule for item 10 (.260 vs. .259), the inclusion of SES and vocabulary leads to a significant difference in performance in favor of English-non-dominant participants. There were four items that English-non-dominant participants performed better on compared to English-dominant participants. However, these mean differences were very small and not significant.

Chapter 12: Discussion

Though pseudowords have featured an increasing presence in cognitive ability assessments, they have yet to be studied in detail. Specifically, can test-developers gain more control over verbal fluid intelligence item difficulty by varying the length and wordlikeness of pseudowords? Additionally, as pseudowords were introduced in cognitive ability assessments to reduce contamination due to prior familiarity, do manipulations to pseudoword properties impact linguistic sub-groups of test-takers differently?

Summary of Findings

Pseudoword Properties. Hypotheses 1 through 3 were concerned with understanding how item difficulty was impacted by manipulations to pseudoword length (operationalized as number of syllables) and wordlikeness (operationalized as the Levenshtein distance of a pseudoword's twenty closest neighbors). Put differently, hypotheses 1 through 3 were concerned with exploring manipulations to pseudoword properties in a cognitive ability testing context. Previous research in cognitive psychology had documented how changes in pseudoword properties impact tasks of pseudoword learning, recall, and recognition (Ellis & Beaton, 1993; Gathercole et al., 1999; Hulme et al., 1991; Papagno et al., 1991; Papagno & Vallar, 1992; Service & Craik, 1993; Vitevitch et al., 1997).

In the current research, while manipulations to pseudoword length did appear to influence item difficulty, these differences were not significant nor were they in the expected direction. Curiously, items featuring longer pseudowords appeared to be easier to solve than items featuring shorter pseudowords. One potential cause of this is that longer pseudowords are more amenable to being truncated than shorter ones. While the current study attempted to prevent against truncating pseudowords by featuring two that began with the same letter in every

problem stem, longer pseudowords can still be truncated to the first syllable. Thus, the effect of length is removed and in the vast majority of cases, the first syllable will be shorter in length and simpler than the one-syllable pseudowords used in the present research. While efforts were made to identify when participants were utilizing this strategy it is possible that participants were unaware this constituted a strategy or just did not report using this strategy. There is also some limited evidence of instances or contexts where longer words or pseudowords appear easier to accommodate cognitively (Romani et al., 2005), with cognitive ability contexts perhaps serving as another of those contexts. Lastly, the evidence surrounding (pseudo)word length effects is generally much more mixed than wordlikeness (Baddeley, 1975; Jalbert et al., 2011; Lovatt et al., 2000; Neath et al., 2003; Service, 1998).

Regarding wordlikeness, while results were not statistically significant, they were in the intended direction. Items featuring less wordlike pseudowords were more difficult. Unlike length, wordlikeness is a property less susceptible to being undermined by participant strategies such as first-letter or first-syllable strategies. Though keyword strategies (i.e., strategies where an unfamiliar word is replaced with a familiar word) are still possible to use, it's not expected that these will generally lead to improved success on items as the participant is adding another cognitive step (essentially, one of translation) to the item-solving process. The current results fit the extant literature when considering the stronger effects for wordlikeness compared to length (Jalbert et al., 2011; Lovatt et al., 2000; Neath et al., 2003) as well as the greater resistance to strategies that undermined wordlikeness.

Hypothesis 3 predicted an interaction effect between the properties of length and wordlikeness. It was anticipated that items featuring both long and un-wordlike pseudowords would be the most difficult of all four potential combinations. However, as manipulations to

length did not impact item difficulty as hypothesized, the interaction effect did not emerge as hypothesized.

Hypotheses 4 through 6 were concerned with exploring the interplay between linguistic status and word properties – both in terms of length and wordlikeness as well as real vs. pseudowords. Hypothesis 4a posited that English-non-dominant participants would perform worse on items involving real words compared to pseudowords. However, English-non-dominant participants ended up performing roughly equivalently across both types. In hindsight, this should have been anticipated. If English-non-dominant participants have less familiarity with the English language as well as smaller vocabularies (the latter supported by both the literature and the present research), then many of the real words used, if either completely or mostly unfamiliar to English-non-dominant participants, will effectively function cognitively as pseudowords. Additionally, if the purpose of pseudowords is to reduce contamination due to prior familiarity, then the effects will be more pronounced for English-dominant participants. This was in fact observed as Hypothesis 4b stated that English-dominant participants would perform significantly better than English-non-dominant participants on items involving real words. While the results were not statistically significant, they were in the intended direction. Also of note is that this was found after controlling for participant vocabulary.

Hypothesis 5 explored the interaction between linguistic status and pseudoword length and specifically posited that as pseudowords moved from short to long, English-non-dominant participants would experience a greater decline in performance compared to English-dominant participants. As with Hypothesis 1, the effect of pseudoword length was not in the intended direction and thus Hypothesis 5 was not supported. Given that the participants who generated data for Hypotheses 1 through 3 were a different sub-sample of participants, this is additional

evidence that pseudoword length does not behave as anticipated and that longer pseudowords, lead to better performance on cognitive ability items. As explained earlier, and in line with some research, it is believed that the longer words or pseudowords generally provide more possible memoranda (whether it be in terms of letters or syllables) and thus may be easier to accommodate cognitively in certain contexts (Romani et al., 2005).

Hypothesis 6 explored the interaction between linguistic status and pseudoword wordlikeness and stated that as pseudowords move from wordlike to un-wordlike, English-non-dominant participants will experience a greater decline in performance compared to English-dominant participants. Similar to the results for hypotheses 4a and 4b, English-non-dominant participant performance was fairly equal across both conditions of wordlikeness. It was in fact the English-dominant participants who experienced a decline in performance as pseudowords moved from wordlike to un-wordlike. While the interaction term between pseudoword wordlikeness and participant linguistic status for hypothesis 6 was not significant, it was near the traditional threshold for statistical significance ($p = .070$) and did provide a glimpse into two concepts regarding group score differences and the nature of pseudowords, to be discussed below.

Additionally, item-level analyses were conducted. However, only a small number of items were significantly impacted for each variable of interest, whether that variable was length, wordlikeness, or the linguistic sub-group. Across those impacted items, there were no discernable patterns. Some impacted items required participants to provide the full sequence of items (e.g., ‘what would the correct order be?’) whereas other impacted items only required participants to provide either the product name or the position (e.g., ‘Which position would [pseudoword] be in?’). The smaller effects across items was anticipated as the intended effect

size overall was small to medium. The current phenomena appears difficult to capture in any individual item and rather, is expected to manifest either in a testing situation with a greater number of test-takers or in a test with more than sixteen items – as is currently experienced in employment or academic testing situations.

Theoretical and Practical Implications

The current research demonstrated a few concepts relevant for both future research and replication. The first is that individuals who are more familiar with English perform better on verbal reasoning questions containing real words rather than pseudowords. While the effects of real words/prior familiarity had been demonstrated in other research (Benjamin, 2009; Papagno et al., 1991) including research in cognitive ability contexts (Fagan & Holland, 2009; Sternberg, 2009), the current research adds to the extant literature – this time via a verbal reasoning format heretofore unexamined.

Regarding the properties of length and wordlikeness, a few patterns emerge that may have practical implications for test developers. Pseudoword length may be useful for test developers hoping to vary the difficulty of otherwise isomorphic items. It appears as though the instances where (pseudo)word length leads to a decrease in task performance may be limited to contexts where the stimuli disappear. In the current research, the stimuli remained on screen throughout the problem-solving process and something about the increased length led to increased item performance. There have been past instances where the increased length of a word was not viewed as placing additional strain on the cognitive processes of an individual but rather as providing more information for the individual to attend to/remember (Romani et al., 2005). This may be the case in a cognitive ability testing context where the stimuli remain available throughout the problem-solving process.

Wordlikeness may be relevant for test developers attempting to further minimize group score differences on verbal reasoning items. The smallest group differences were achieved on items involving un-wordlike pseudowords, with the differences in performance between linguistic subgroups going from $d = 0.40$ to $d = 0.16$ as pseudowords moved from wordlike to un-wordlike. This latter effect size of 0.16 is quite small for the group differences literature, though it should also be noted these effect sizes are generated between linguist sub-groups rather than between racial/ethnic groups studied in much of the group differences literature.

Ultimately, English-dominant participants experienced a decline in performance compared to English-non-dominant participants. The original hypotheses predicted a decrease in performance for English-non-dominant participants as pseudowords moved from wordlike to un-wordlike. This was based on the belief that the current batch of pseudowords were fairly representative of English words. Put differently, it was believed that the most wordlike pseudowords were similar enough to common words/sounds in English that even individuals with reduced English familiarity would be familiar enough with those words/sounds. Only by moving to less wordlike pseudowords would differences in linguistic background emerge to influence performance. What was in fact observed was that those with English as their dominant language experienced an advantage on items involving wordlike pseudowords and only when items contained un-wordlike pseudowords was the influence of linguistic background minimized. It was previously unclear where the benefits of prior familiarity would lie on the wordlikeness spectrum with the current research suggesting they lie on the wordlike rather than un-wordlike end of the spectrum. This finding, where there are larger group differences on items that are ostensibly ‘easier’ mirrors the work of Freedle and Kostin (1997) who reported that

racial group score differences on SAT and GRE verbal analogy items were larger on easier items than harder items.

Potential Limitations and Suggestions for Future Research

The current research attempted to reduce or remove certain potential limitations – carefully vetting both pseudowords and words across a number of properties (length, lexical neighborhood density, orthographic probability, lexical neighborhood frequency, part of speech, concreteness, and frequency) to ensure that differences in performance were not due to asymmetries in the sets of stimuli, as well as two pilot studies to identify administration issues (such as participants adopting unsuitable strategies when the questions are housed on a different screen from the relationships to be memorized), and to gather any feedback on the items (such as being too difficult or easy) and pseudowords themselves (such as a pseudoword being a word in a different language). Nevertheless, there are a handful of potential limitations to the current research that warrant discussion.

One of the more glaring potential limitations is that the entire study was English based. Specifically, all of the pseudowords constructed were constructed by a researcher who is English-speaking and monolingual and the pseudowords were designed to resemble words in English. It cannot be assumed that the phenomena observed in the current research would transfer to other languages or alphabets. However, while MCWord is English-only, it is derived from the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995) which contains information for English, Dutch, and German words. This may be a logical starting point for understanding how pseudowords operate in a non-English verbal Gf testing context. Additionally, another area of research – and perhaps a way to extend un-wordlikeness – may be to replicate the current research but using pseudowords that do not as closely conform to English but mirror another

language. Papagno et al., (1991) reported that in a study on memorizing and repeating lists of words, Italian participants struggled learning lists of Russian words when subvocalization was suppressed while English participants did not. However, once moving from Russian to Finnish words, the same phenomena was observed for the English participants. This echoes a central point made earlier – just because all pseudowords (and words in another language that operate as pseudowords) are unknown and novel does not mean they are equally unfamiliar (Papagno et al., 1991; Vitevitch et al., 2014).

Another limitation was the absence of two-syllable pseudowords. Based on the work of Jalbert et al., (2011), it was believed a more stringent test of pseudoword length was appropriate, thus the present comparison between pseudowords of one and three syllables. Given that the work of Jalbert and colleagues featured two syllable pseudowords that were generally around six letters in length, and that six letters was the upper bound of one-syllable words in the current research, there are numerous questions remaining. Chiefly among these – as it was found that difficulty on items decreased as pseudowords increased in length from one to three syllables, would two-syllable pseudowords fall somewhere in between these two difficulties? And would this be so even if in terms of letters there is little to no difference in length between pseudowords of one and two syllables?

Regarding wordlikeness, the current research favored lexical neighborhood density over other objective measures of wordlikeness (such as lexical neighborhood frequency or orthographic probability) as this was found to be the dominant measure of wordlikeness (Bailey & Hahn, 2001). The research attempted to take objectively measuring wordlikeness one step further by utilizing the Levenshtein distance, a heretofore under-utilized conceptualization of wordlikeness that affords much greater granularity amongst (pseudo)words with zero or one

neighbors. While this was a more comprehensive effort amongst the objective measures of wordlikeness, there are still subjective measures. Much research has determined wordlikeness simply by asking participants to rate how much a pseudoword resembles a typical word (Bailey & Hahn, 2001; Cheung, 1996; Gathercole, Willis, Emslie, & Baddeley, 1991). Often, these subjective ratings correspond fairly well to objective measures (Bailey & Hahn, 2001; Gathercole, Willis, Emslie, & Baddeley, 1991). However, it remains to be seen if a subjective approach to wordlikeness would have yielded different pseudoword groups and ultimately different results. Additionally, as the two linguistic groups under consideration differ on English ability, might those two groups have different understandings of how much a pseudoword subjectively resembles a word in English? Similar to the cultural-calibration work of Malda et al., 2010, future research may want to calibrate subjective ratings of wordlikeness within the populations to be studied.

The usage of pseudowords in cognitive ability testing contexts is still in relative infancy. Overall, future researchers will want to consider how pseudowords of two syllables behave on tests, how alternative conceptualizations of wordlikeness – primarily a subjective conceptualization – impact how pseudowords are classified and grouped together, and to see if phenomena observed in an English-testing context transfer to other languages.

Regarding the linguistic sub-groups, another potential limitation is the method for determining who is an English-non-dominant participant. While asking bilingual participants to identify their dominant language is pretty straight forward, it is unclear if there is a better way to determine this. There are measures that exist that understand an individual's linguistic profile by asking a series of questions regarding spoken, receptive, and written linguistic abilities. Additionally, the decision to use the age of ten as the cut point for age of learning English to

determine if someone who has English as their second language is English-dominant or not was both backed by the literature and somewhat arbitrary. At this point in time, it is unclear and somewhat doubtful that there will ever be a true answer to the question of ‘how late in life can an individual start learning a second language and still achieve dominance in that second language compared to their first language?’. Based on the available literature, it appears as though that point, were it to exist, would be somewhere between the age of five and thirteen. The current research chose age ten as a middle point. However, the results may look quite different if a different age was chosen or the research had simply relied solely on self-reports of language dominance. Future research may want to incorporate more objective measures of linguistic proficiency or use a different age as a cutpoint for determining dominance.

Lastly, a potential limitation was that the phenomena observed were only done so utilizing one item type and only one test format – a power test. While the item type used was chosen because it appeared as an ideal item type for the current research purposes, it is unclear how pseudowords would behave in other item types. The items utilized in the research by Fagan and Holland (2009) and Sternberg (1981, 2006) are a logical starting point to determine if the observed phenomena hold across alternative item types. Similarly, the current research utilized a power test. While there were time limits imposed on participants for the reasoning questions, they were generous time limits – too generous for the current research to be considered a speeded test. Speeded tests differ from power tests in that they feature a time limit that leaves little time per item and success requires fairly quick reasoning and responses (Mead & Drasgow, 1993). Generally, having to switch languages or reason in a less familiar language adds time to cognitive operations (Gollan & Ferreira, 2009), which can lead to better performance by monolingual or English-dominant relative to English-non-dominant test-takers on cognitive tasks

– particularly speeded cognitive tasks (Bialystok, Craik, Green, & Gollan, 2009). Additionally, test performance of English-non-dominant test-takers improves when given extra time to complete a test (Abedi, Hofstetter, Baker, & Lord, 2001). While it was believed that a speeded test design would not be an ideal cognitive ability context starting point, it nevertheless remains to be seen how changes to pseudoword properties impact performance in a speeded context across linguistic sub-groups.

Conclusions

Overall, the current research attempted to explore an old technique in a new context so that psychologists may assess verbal fluid reasoning in ways that could give more control to the item developer (i.e., attempting to use pseudoword properties to influence item difficulty) and could potentially minimize contamination due to prior familiarity. The findings suggest that while there is no silver bullet for the above issues, pseudowords are an effective way to reduce contamination due to prior familiarity compared to real words and that certain forms of pseudowords are better at this than others.

For test development purposes, pseudoword length emerged as an interesting way to potentially impact item difficulty, and the finding that longer pseudowords may lead to easier items was counterintuitive. Regarding the ability to measure verbal fluid reasoning in a way that minimizes group differences, the current research demonstrated that moving from real words to pseudowords helps reduce group score differences, and that moving from wordlike to un-wordlike short pseudowords may be the best technique of all to reduce contamination due to prior familiarity.

The effective understanding and measurement of cognitive ability is one of the cornerstones of psychology. The current research examined an under-explored technique that is

useful for measuring a facet of fluid intelligence that research has increasingly moved away from – verbal fluid intelligence. It is hoped that this research is part of a renaissance for creating improved items and measuring the construct of verbal fluid intelligence with ever-increasing precision.

Appendix A: Verbal Fluid Intelligence Items with Pseudowords

Item Stem 1

You are learning about the differences in value between products.

Trawns is more valuable than Murnt.

Knills is either most valuable or least valuable.

Boust is more valuable than Trawns.

If Knills is the most valuable, Brench is less valuable than Murnt.

If Knills is the least valuable, Brench is more valuable than Murnt.

1. Which order is possible?
 - a. Boust, Trawns, Brench, Knills, Murnt
 - b. Knills, Boust, Murnt, Brench, Trawns
 - c. Boust, Brench, Murnt, Trawns, Knills
 - d. Brench, Boust, Trawns, Murnt, Knills
2. If Knills is the most valuable, which is the third-most valuable?
 - a. Boust
 - b. Brench
 - c. Trawns
 - d. Murnt
3. If Murnt and Boust switched value, which order could be possible?
 - a. Knills, Murnt, Boust, Trawns, Brench
 - b. Brench, Murnt, Trawns, Boust, Knills
 - c. Murnt, Trawns, Boust, Knills, Brench
 - d. Knills, Brench, Murnt, Trawns, Boust
4. If Brench is more valuable than Trawns, which must be true?
 - a. Knills is most valuable
 - b. Boust is most valuable
 - c. Murnt is least valuable
 - d. Knills is least valuable

Item Stem 2

You are setting up a display for five different products but the company has rules for which products are allowed where.

Praugh can only be on either end of the display.

Plilge cannot be next to Choulf.

Jevsh cannot be first or third in the display.

Smoelb and Choulf must be exactly one product away from each other.

5. If Praugh is on the left end of the display, which product is third?
 - a. Smoelb
 - b. Choulf
 - c. Jevsh
 - d. Plilge
6. Which order is possible?
 - a. Choulf, Jevsh, Smoelb, Plilge, Praugh
 - b. Plilge, Smoelb, Jevsh, Choulf, Praugh
 - c. Praugh, Plilge, Smoelb, Choulf, Jevsh
 - d. Praugh, Jevsh, Choulf, Plilge, Smoelb
7. When Choulf is the product that is on the right-end of the display, which product is second from the left?
 - a. Smoelb
 - b. Jevsh
 - c. Praugh
 - d. Plilge
8. If all the rules remained the same except the rule for Jevsh was changed to state that Jevsh must be third in the display, which order is possible?
 - a. Smoelb, Plilge, Jevsh, Choulf, Praugh
 - b. Plilge, Smoelb, Jevsh, Choulf, Praugh
 - c. Praugh, Plilge, Jevsh, Smoelb, Choulf
 - d. Praugh, Choulf, Jevsh, Plilge, Smoelb

Item Stem 3

You are ordering stock for different products. Due to space limits, the products must be ordered according to the following rules:

When Lathera is ordered first-most, Becated is ordered fourth-most

Verises can be ordered in any amount except first

Lathera can only be ordered first- or second-most

Forrier can be ordered in any amount except fifth-most

Bantiness is always ordered one position more than Becated

9. When Forrier is ordered the most, which quantities can Verises be ordered in?
 - a. Second
 - b. Third
 - c. Fourth
 - d. Fifth
10. Across all correct ways to place an order, which quantities can Bantiness be ordered in? (Check all that apply).
 - a. Second
 - b. Third
 - c. Fourth
 - d. Fifth
11. Which of the following orders is not possible?
 - a. Forrier, Lathera, Verises, Bantiness, Becated
 - b. Lathera, Forrier, Bantiness, Becated, Verises
 - c. Forrier, Lathera, Bantiness, Becated, Verises
 - d. Lathera, Forrier, Verises, Bantiness, Becated
12. If all the rules remained the same except the last rule was changed so that Bantiness is always ordered two positions more than Becated, which order is not possible?
 - a. Lathera, Bantiness, Forrier, Becated, Verises
 - b. Forrier, Bantiness, Lathera, Becated, Verises
 - c. Forrier, Lathera, Bantiness, Verises, Becated
 - d. Bantiness, Lathera, Becated, Forrier, Verises

Item Stem 4

Below are sales trends for the summer. You also know that the sales order reverses during the winter.

Womerer sells less than Isolents

Waffias sells more than Isolents

Tecression sells less than Waffias

Tecression sells more than Socidence

Socidence sells less than Tecression

13. Which product sells the most in the summer?

- a. Waffias
- b. Tecression
- c. Womerer
- d. Isolents

14. If Tecression also sells less than Womerer during the summer, which order could be true?

- a. Waffias, Isolents, Socidence, Womerer, Tecression
- b. Isolents, Waffias, Womerer, Tecression, Socidence
- c. Waffias, Isolents, Womerer, Tecression, Socidence
- d. Waffias, Isolents, Womerer, Socidence, Tecression

15. Which order is possible in the winter?

- a. Socidence, Womerer, Tecression, Isolents, Waffias
- b. Womerer, Isolents, Tecression, Socidence, Waffias
- c. Womerer, Tecression, Socidence, Isolents, Waffias
- d. Socidence, Womerer, Isolents, Tecression, Waffias

16. If Socidence went on sale and started selling more than Isolents in the summer, which product is the third-worst selling product in the winter?

- a. Womerer
- b. Socidence
- c. Tecression
- d. Waffias

Appendix B: Linguistic Background Measure

1. Do you speak a language or languages other than English?
 - a. If yes, please, list all of the other languages you speak.

(Questions 2 through 6 of the linguistic background measure are skipped if the participant indicates that they only speak English)
2. Based on your response to the previous question, what do you consider to be your primary language (that is the language you feel most comfortable speaking)?
3. At what age did you start learning to speak English?
4. How many years have you been speaking English?
5. How often is English spoken in your home?
 - a. Never
 - b. Rarely
 - c. Sometimes
 - d. Quite often
 - e. Always
6. What language do you generally speak when you are with friends?

Appendix C: Socioeconomic Status Measure

1. What is your family's annual gross (pre-tax) income?¹²
 - a. \$1,000 to \$10,000
 - b. \$10,001 to \$20,000
 - c. \$20,001 to \$30,000
 - d. \$30,001 to \$40,000
 - e. \$40,001 to \$50,000
 - f. \$50,001 to \$75,000
 - g. \$75,001 to \$100,000
 - h. \$100,001 to \$150,000
 - i. More than \$150,000

2. a. How many separate rooms are in your house or apartment? (Do not count bathrooms, porches, balconies, foyers, halls, or unfinished basements).¹³

b. How many of these rooms are bedrooms? (If you live in a studio apartment, enter '0').

3. In your house or apartment, do you (or any person who lives in your house or apartment) have the following:
 - a. Desktop or laptop
 - b. Smartphone
 - c. Tablet or other portable wireless computer
 - d. Some other type of computer

¹² Item adapted from the U.S. Census.

¹³ Items 2a, 2b, and 3 are modified versions of those appearing in the American Community Survey (i.e., the long-form version of the U.S. Census). Available online at: <http://www2.census.gov/programs-surveys/acs/methodology/questionnaires/2016/quest16.pdf>

Appendix D: Demographic Measure

1. What year are you in college?
 - a. Freshman
 - b. Sophomore
 - c. Junior
 - d. Senior
2. What is your current GPA?
 - a. Do not yet have a GPA/do not remember
 - b. 0.00-0.49
 - c. 0.50-0.99
 - d. 1.00-1.49
 - e. 1.50-1.99
 - f. 2.00-2.49
 - g. 2.50-2.99
 - h. 3.00-3.49
 - i. 3.50-3.99
 - j. 4.00
3. What is your age?
4. What is your gender?
 - a. Male
 - b. Female
 - c. Other
5. What is your ethnicity? (Check all that apply)
 - a. White, non-Latino or Hispanic
 - b. Black/African, non-Latino or Hispanic
 - c. Latino or Hispanic
 - d. East-Asian
 - e. South-Asian
 - f. Middle Eastern
 - g. Native Hawaiian or Pacific Islander
 - h. Native American or Alaska Native
 - i. Other (please specify)

Table 1. Pilot Study One Condition Properties

Condition	Number of statements	Number of pseudowords	Number of pseudoword relationships	Number of response options	n	Mean difficulty	Standard deviation of difficulty
1	4	4 or 5	1	2	21	.800	.190
2	3	4	2	2	25	.908	.122
3	4	4 or 5	2	2	31	.855	.136
4	3	4	2	3	21	.800	.200
5	4	4 or 5	2	3	19	.775	.133
6	4	4 or 5	1	2	22	.732	.178
7	4	4 or 5	1	2	21	.819	.218
8	4	4 or 5	1	2	25	.860	.191

Table 2. Pilot Study One Pseudoword Properties

Pseudoword	Number of Letters	Number of Syllables	Lexical Neighborhood Density	Lexical Neighbor Frequency	Levenshtein Distance	Orthographic Probability
ance*	4	1	2	289.82	1.75	660.87
blalt	5	1	1	11.78	1.95	1,028.15
fause	5	1	3	66.89	1.85	1,485.82
jould*	5	1	4	1,207.75	1.80	4,927.49
jures*	5	1	3	1.51	1.80	1,296.23
kearce*	6	1	0	0.00	2.70	923.92
knace*	5	1	2	1.81	1.90	1,092.05
orke	4	1	0	0.00	1.95	1,275.27
reand	5	1	0	0.00	1.80	1,732.93
yopte	5	1	0	0.00	2.00	645.50
bathic	6	2	0	0.00	2.15	1,065.02
bumen*	5	2	0	0.00	1.95	1,089.19
dampur	6	2	1	0.77	2.15	427.18
egspro	6	2	0	0.00	3.20	4.85
etstak*	6	2	0	0.00	3.00	286.41
gerev	5	2	0	0.00	2.00	380.37
gractions	9	2	1	1.25	2.15	597.15
lightling	9	2	1	14.75	2.15	1,077.27
luxua	5	2	0	0.00	2.95	43.64
meattered	9	2	0	0.00	1.95	790.50
migo	4	2	0	0.00	2.00	490.32
orkno*	5	2	0	0.00	3.00	80.90
oxcagh	6	2	0	0.00	3.40	402.43
qulof	5	2	0	0.00	2.80	327.49
screading	9	2	2	20.08	2.05	1,122.78

Pseudoword	Number of Letters	Number of Syllables	Lexical Neighborhood Density	Lexical Neighbor Frequency	Levenshtein Distance	Orthographic Probability
udsti	5	2	0	0.00	3.00	109.50
communted*	9	3	1	13.98	2.05	1,471.46
compoution	10	3	0	0.00	2.85	1,172.21
contration*	10	3	1	1.01	2.15	1,336.76
delection	9	3	4	9.73	1.65	1,166.99
demactint	9	3	0	0.00	1.90	1,298.65
destering*	9	3	2	0.80	3.00	1,593.39
gractimer	9	3	0	0.00	3.20	472.79
interpries	10	3	0	0.00	2.80	730.70
physiders	9	3	0	0.00	3.50	320.90
propospere*	10	3	0	0.00	3.50	311.36
vanderies*	9	3	0	0.00	2.80	819.88
vollolate*	9	3	0	0.00	3.50	351.30

*only used in condition 8.

Table 3. Short and Wordlike Pseudoword Properties

Pseudoword	Number of Letters	Lexical Neighborhood Density	Lexical Neighbor Frequency	Levenshtein Distance	Orthographic Probability
boust	5	3	4.24	1.70	2,987.59
brench	6	9	24.69	1.50	692.09
clent	5	3	1.57	1.70	959.40
coids	5	7	3.08	1.60	1,573.20
drants	6	2	8.66	1.80	737.01
druld	5	1	0.18	1.95	3,148.43
fimps	5	2	0.83	1.85	735.09
foungs	6	2	0.03	1.90	947.77
grangs	6	3	5.53	1.70	803.57
grouge	6	1	1.67	1.85	1,477.32
knills	6	1	0.18	1.90	410.63
krough	6	1	3.27	1.90	1,407.37
meange	6	1	0.24	1.85	1,079.51
murnt	5	1	19.34	1.95	544.22
neaks	5	6	3.31	1.65	1,552.49
neaves	6	4	27.04	1.70	1,350.66
ounte	5	1	5.47	1.95	612.39
ourch	5	1	2.32	1.90	1,276.57
plarce	6	1	1.31	1.90	796.76
plench	6	3	0.40	1.85	611.94
strint	6	3	15.23	1.80	1,911.64
swelds	6	1	1.90	1.90	264.05
tinch	5	5	1.98	1.70	1,728.73
trawns	6	3	5.22	1.85	353.39
Means	5.58	2.71	5.74	1.81	1,165.08

Table 4. Short and Un-wordlike Pseudoword Properties

Pseudoword	Number of Letters	Lexical Neighborhood Density	Lexical Neighbor Frequency	Levenshtein Distance	Orthographic Probability
blaarg	6	0	0.00	2.90	271.76
blorzk	6	0	0.00	2.90	233.96
choulf	6	0	0.00	2.80	1,670.89
curngh	6	0	0.00	2.90	844.50
dournz	6	0	0.00	2.65	632.02
dreeck	6	0	0.00	2.60	446.47
fersch	6	0	0.00	2.80	453.95
frufth	6	0	0.00	2.65	277.21
glyth	5	0	0.00	2.60	489.85
gremph	6	0	0.00	2.90	401.65
holnge	6	0	0.00	2.65	634.75
hydst	5	0	0.00	2.50	541.18
jarpht	6	0	0.00	2.95	487.44
jevsh	5	0	0.00	2.55	417.74
knulth	6	0	0.00	2.95	196.02
kroulk	6	0	0.00	3.00	1,174.29
plilge	6	0	0.00	2.70	435.01
praugh	6	0	0.00	2.65	750.20
quenth	6	1	0.24	2.90	401.80
quolsh	6	0	0.00	2.95	82.58
skosch	6	0	0.00	2.70	203.54
smoelb	6	0	0.00	2.90	381.60
zoyce	5	1	11.90	2.60	590.26
zurpt	5	0	0.00	2.50	134.91
Means	5.79	0.08	0.51	2.76	506.40

Table 5. Long and Wordlike Pseudoword Properties

Pseudoword	Number of Letters	Lexical Neighborhood Density	Lexical Neighbor Frequency	Levenshtein Distance	Orthographic Probability
alashing	8	5	2.80	1.70	1,782.38
attering	8	3	2.04	1.65	2,053.29
austiness	9	4	0.06	1.80	434.31
bantiness	9	1	0.06	2.05	426.35
becated	7	2	1.43	1.90	2,803.25
concested	9	4	1.98	2.00	1,408.84
conderting	10	2	2.94	1.95	1,181.18
delating	8	6	3.00	1.60	2,213.33
delection	9	4	9.73	1.65	1,166.99
eleating	8	3	1.01	1.75	2,183.05
eration	7	2	1.34	1.80	996.90
faminess	8	1	0.00	1.90	648.80
forrier	7	3	0.14	1.85	1,519.84
landiness	9	4	0.03	1.95	351.39
lathera	7	2	0.03	1.85	849.06
manuted	7	2	0.15	1.90	2,050.41
multiness	9	1	0.24	1.95	315.99
murdiest	8	2	0.00	1.90	416.36
polations	9	1	0.06	1.95	699.25
promiting	9	4	6.07	2.30	1,678.25
recorted	8	4	22.83	1.80	1,897.05
remented	8	5	3.45	1.75	1,923.43
verises	7	1	0.00	1.90	1,173.04
vetation	8	1	0.77	1.95	960.22
Means	8.17	2.79	2.51	1.87	1,297.21

Table 6. Long and Un-wordlike Pseudoword Properties

Pseudoword	Number of Letters	Lexical Neighborhood Density	Lexical Neighbor Frequency	Levenshtein Distance	Orthographic Probability
bleconers	9	0	0.00	3.00	343.57
brarions	8	0	0.00	2.70	507.22
cavoral	7	0	0.00	2.65	715.73
communate	9	1	0.00	2.45	781.68
decepter	8	0	0.00	2.70	902.14
distinere	9	0	0.00	3.00	709.00
folitless	9	0	0.00	2.95	416.58
fondiction	10	0	0.00	2.80	1,097.69
incorted	9	0	0.00	3.00	1,135.98
isolents	8	0	0.00	2.75	503.67
leriper	7	0	0.00	2.70	1,345.21
levaron	7	0	0.00	2.90	564.52
polighted	9	0	0.00	2.50	1,069.69
porserins	9	0	0.00	3.00	1,084.92
recrement	9	0	0.00	2.75	943.86
resilencs	9	0	0.00	2.70	525.98
socidence	9	0	0.00	2.95	492.61
spriously	9	1	64.85	2.45	425.22
tecreSSION	10	0	0.00	2.75	947.41
teleption	9	0	0.00	2.75	992.62
viblited	8	0	0.00	2.85	1,382.27
vutraction	10	0	0.00	2.70	928.75
waffias	7	0	0.00	2.70	213.16
womerer	7	0	0.00	2.80	1,379.81
Means	8.54	0.08	2.70	2.77	808.72

Table 7. Short and Wordlike Word Properties

Word	Number of Letters	Word Frequency	Lexical Neighborhood Density	Lexical Neighbor Frequency	Levenshtein Distance	Orthographic Probability	Concreteness Mean*
bleach	6	1.90	3	3.35	1.75	396.86	4.74
breeze	6	11.24	2	4.88	1.80	365.92	3.62
chive	5	0.12	5	1.78	1.65	2,272.44	4.81
clique	6	2.14	2	0.06	1.90	207.15	2.68
dearth	6	0.95	1	4.64	1.85	965.65	--
drape	5	1.01	4	1.01	1.60	552.17	4.50
finch	5	0.65	4	1.98	1.70	2,071.40	4.30
frieze	6	1.96	1	8.75	1.90	422.55	--
gorge	5	8.09	3	2.10	1.65	1,140.51	4.15
grouse	6	1.67	1	6.19	1.80	1,352.69	--
haunt	5	4.64	5	1.33	1.60	1,079.62	2.03
hearse	6	2.20	1	4.40	1.85	1,023.49	4.85
maize	5	6.43	1	1.73	1.90	386.03	4.58
mulch	5	0.65	4	0.37	1.75	1,194.51	4.59
plague	6	7.56	2	1.46	1.95	384.91	3.41
prawn	5	1.19	2	33.41	1.80	623.90	4.62
quaff	5	0.00	1	0.12	1.90	303.38	--
quark	5	0.18	3	1.65	1.80	872.74	2.79
shrink	6	6.19	4	3.93	1.75	1,562.85	3.55
splint	6	0.71	1	2.32	1.85	1,368.48	4.69
throne	6	9.76	2	1.90	1.85	924.72	4.64
troop	5	4.64	1	1.49	1.90	958.93	4.34
whisk	5	2.56	1	0.77	1.85	3,311.53	4.33
wrench	6	3.93	4	37.79	1.75	584.45	4.93
Means	5.50	3.35	2.42	5.31	1.80	1013.62	4.11

*Concreteness statistics were unavailable for some words.

Table 8. Long and Un-wordlike Word Properties

Word	Number of Letters	Word Frequency	Lexical Neighborhood Density	Lexical Neighbor Frequency	Levenshtein Distance	Orthographic Probability	Concreteness Mean**
ambience	8	1.13	0	0.00	2.85	600.53	1.72
asterisk	8	0.06	0	0.00	2.80	542.74	4.00
bassinet	8	0.12	0	0.00	2.85	611.53	4.71
billiards	9	2.74	1	0.24	2.70	187.62	4.61
condiment	9	0.48	0	0.00	2.80	989.81	4.72
constable	9	10.23	0	0.00	2.75	680.95	4.54
decanter*	8	0.95	1	0.30	2.55	945.21	3.43
defroster	9	0.00	1	0.30	2.90	510.60	4.23
ganglion	8	0.00	0	0.00	2.90	671.84	--
grenadine	9	0.12	0	0.00	2.80	733.38	4.19
hexagon	7	0.83	0	0.00	2.80	322.26	4.52
holograph	9	0.06	1	0.00	2.50	201.83	4.27
megalith	8	0.06	0	0.00	2.90	415.47	--
molasses	8	1.61	1	0.06	2.55	672.64	4.84
nautilus	8	0.59	0	0.00	3.00	236.59	--
novella	7	0.06	1	0.00	2.65	383.37	--
pancreas	8	0.54	0	0.00	2.90	325.75	4.79
prospectus	10	2.68	0	0.00	2.85	275.93	2.65
racketeer	9	0.06	0	0.00	2.95	181.53	3.27
regatta	7	0.36	0	0.00	2.85	536.35	--
scorpion	8	0.95	1	0.00	2.70	558.43	4.84
syndicate	9	1.96	0	0.00	2.65	288.28	2.33
telegraph	9	2.97	0	0.00	2.85	224.47	4.59
trimester	9	0.06	0	0.00	2.85	527.63	2.78
Means	8.38	1.19	0.29	0.04	2.79	484.36	3.95

*Concreteness statistic provided for the word lemma 'decant'.

**Concreteness statistics were unavailable for some words.

Table 9. Pseudoword and Word Property Means by Stimuli Set

Word Type	Stimuli Set	Number of Letters	Lexical Neighborhood Density	Lexical Neighbor Frequency	Levenshtein Distance	Orthographic Probability	Word Frequency	Concreteness
Pseudoword	Short, Wordlike	5.58	2.71	5.74	1.81	1,165.08	--	--
	Short, Un-wordlike	5.79	0.08	0.51	2.76	506.40	--	--
	Long, Wordlike	8.17	2.79	2.51	1.87	1,297.21	--	--
	Long, Un-wordlike	8.54	0.08	2.70	2.77	808.72	--	--
Real Words	Short, Wordlike	5.50	2.42	5.31	1.80	1,013.62	3.35	4.11
	Long, Un-wordlike	8.38	0.29	0.04	2.79	484.36	1.19	3.95

Table 10. Pilot Study Two Item Stem and Item Difficulties

Stem	Item	n	Mean difficulty	Standard deviation
1		60	0.588	0.328
	1	60	0.700	0.462
	2	60	0.550	0.502
	3	60	0.583	0.497
	4	60	0.517	0.504
2		60	0.483	0.252
	1	60	0.633	0.486
	2	60	0.717	0.454
	3	60	0.500	0.504
	4	60	0.083	0.279
3		60	0.342	0.225
	1	60	0.467	0.503
	2	60	0.233	0.427
	3	60	0.383	0.490
	4	60	0.283	0.454
4		60	0.617	0.232
	1	60	0.917	0.279
	2	60	0.817	0.390
	3	60	0.517	0.504
	4	60	0.217	0.415

Table 11. Mean Total Score, Vocabulary Score, and Socioeconomic Status (SES) by Stimuli Set

Word Types	Stimuli Set	n	Total Score		Vocabulary		SES	
			mean	sd	mean	sd	mean	sd
Pseudowords	Short, Wordlike	55	9.05	3.08	9.27	2.41	-0.03	1.01
	Short, Un-wordlike	62	8.79	3.56	9.23	2.49	-0.03	0.99
	Long, Wordlike	61	9.80	3.34	8.54	2.60	0.01	1.04
	Long, Un-wordlike	61	9.11	3.46	8.59	2.97	0.02	1.01
Pseudowords and Real Words	Short, Wordlike	73	8.16	3.65	9.05	2.28	-0.02	0.99
	Long, Un-wordlike	56	8.55	3.82	8.80	2.81	0.00	0.98
Total		368	8.89	3.52	8.92	2.59	-0.01	1.00

Table 12. Mean Item and Total Scores Overall and by Linguistic Sub-group

Item	Overall		English-dominant		English-non-dominant	
	Mean difficulty*	sd	Mean difficulty	sd	Mean difficulty	sd
1	0.647	0.479	0.677	0.469	0.601	0.491
2	0.543	0.499	0.591	0.493	0.473	0.501
3	0.451	0.498	0.523	0.501	0.345	0.477
4	0.511	0.501	0.532	0.500	0.480	0.501
5	0.571	0.496	0.600	0.491	0.527	0.501
6	0.685	0.465	0.727	0.446	0.622	0.487
7	0.484	0.500	0.491	0.501	0.473	0.501
8	0.677	0.468	0.727	0.446	0.601	0.491
9	0.465	0.499	0.518	0.501	0.385	0.488
10	0.247	0.432	0.268	0.444	0.216	0.413
11	0.609	0.489	0.632	0.483	0.574	0.496
12	0.563	0.497	0.573	0.496	0.547	0.499
13	0.897	0.305	0.900	0.301	0.892	0.312
14	0.793	0.405	0.800	0.401	0.784	0.413
15	0.489	0.501	0.514	0.501	0.453	0.499
16	0.261	0.440	0.259	0.439	0.264	0.442
Total Score	8.891	3.515	9.331	3.558	8.236	3.357

*Higher difficulty values indicate easier items.

References

- Abedi, J. (2010). Performance assessments for English language learners. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.
- Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). NAEP math performance and test accommodations: Interactions with student language background. CSE Technical Report 536. Retrieved from: <https://files.eric.ed.gov/fulltext/ED466961.pdf>.
- Abedi, J., Lord, C., & Plummer, J. R. (1997). Final report of language background as a variable in NAEP mathematics performance. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.
- Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, 97(1), 129-133.
- Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, 30(4), 481-509.
- Abutalebi, J., Della Rosa, P. A., Green, D. W., Hernandez, M., Scifo, P., Keim, R., Cappa, S. F., & Costa, A. (2011). Bilingualism tunes the anterior cingulate cortex for conflict monitoring. *Cerebral Cortex*, 22(9), 2076-2086.
- Abutalebi, J., & Green, D. W. (2007). Bilingual language production: The neurocognition of language representation and control. *Journal of Neurolinguistics*, 20 (3), 242-275.
- Abutalebi, J., & Green, D. W. (2008). Control mechanisms in bilingual language production: Neural evidence from language switching studies. *Language and Cognitive Processes*, 23(4), 557-582.

- Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, 131(1), 30-60.
- Agnello, P., Ryan, R., & Yusko, K. (2015). Implications of modern intelligence research for assessing intelligence in the workplace. *Human Resource Management Review*, 25(1), 47-55.
- Allen, R., & Hulme, C. (2006). Speech and language processing mechanisms in verbal serial recall. *Journal of Memory and Language*, 55(1), 64-88.
- Anderson, J. R. (1981). Effects of prior knowledge on memory for new information. *Memory & Cognition*, 9(3), 237-246.
- Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, 128(2), 186-197.
- Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2), 234.
- Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). CELEX2 LDC96L14. Web Download. Philadelphia: Linguistic Data Consortium, 1995. Retrieved from: <https://catalog.ldc.upenn.edu/ldc96l14>, last retrieved on 11/16/2017.
- Baddeley, A. D. (1998). Recent developments in working memory. *Current Opinion in Neurobiology*, 8(2), 234-238.
- Baddeley, A. D. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, 36(3), 189-208.
- Baddeley, A. D. (2007). *Working Memory, Thought, and Action*. Oxford University Press.

- Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, 63, 1-29.
- Baddeley, A. D., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, 105(1), 158-173.
- Baddeley, A.D. & Logie, R.H. (1999). Working memory: The Multiple-Component Model. In A. Miyake, & P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (pp. 28-61). Cambridge University Press.
- Baddeley, A. D., Papagno, C., & Vallar, G. (1988). When long-term learning depends on short-term storage. *Journal of Memory and Language*, 27(5), 586-595.
- 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.
- Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44(4), 568-591.
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133(2), 283.
- Barbey, A. K., Colom, R., Paul, E. J., & Grafman, J. (2014). Architecture of fluid intelligence and working memory revealed by lesion mapping. *Brain Structure and Function*, 219(2), 485-494.
- Bard, E., & Shillcock, R. (1993). Competitor effects during lexical access: Chasing Zipf's tail. In G. T.M. Altmann & R. Shillcock (Eds.) *Cognitive Models of Speech Processing* (pp. 235-275). Psychology Press.

- Bartolotti, J., & Marian, V. (2014). Wordlikeness and Novel Word Learning. In annual meeting of the Cognitive Science Society, Quebec City, Canada. Paper retrieved from <https://mindmodeling.org/cogsci2014/papers/036/paper036.pdf>.
- Barton, J. J., Hanif, H. M., Eklinder Björnström, L., & Hills, C. (2014). The word-length effect in reading: A review. *Cognitive Neuropsychology*, 31(5-6), 378-412.
- Basnight-Brown, D.M. (2013). Models of lexical access and bilingualism. In J. Altarriba, & R. Heredia (Eds.). *Understanding Bilingual Memory: Theory and Applications*. New York: Springer.
- Benjamin, L. T. (2006). *A Brief History of Modern Psychology*. Wiley-Blackwell.
- Berko, J. (1958). The child's learning of English morphology. *Word*, 14(2-3), 150-177.
- Bialystok, E. (2005). Consequences of bilingualism for cognitive development. In J.F. Kroll and A. M.B. DeGroot (Eds.) *Handbook of Bilingualism: Psycholinguistic Approaches*, (pp. 417-432). Oxford University Press, London.
- Bialystok, E. (2006). Effect of bilingualism and computer video game experience on the Simon task. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 60(1), 68-79.
- Bialystok, E., Craik, F. I., Grady, C., Chau, W., Ishii, R., Gunji, A., & Pantev, C. (2005). Effect of bilingualism on cognitive control in the Simon task: Evidence from MEG. *NeuroImage*, 24(1), 40-49.
- Bialystok, E., Craik, F. I., Green, D. W., & Gollan, T. H. (2009). Bilingual minds. *Psychological Science in the Public Interest*, 10 (3), 89-129.
- Bialystok, E., Craik, F. I., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: evidence from the Simon task. *Psychology and Aging*, 19(2), 290-303.

- Bialystok, E., Craik, F. I., & Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends in Cognitive Sciences*, 16(4), 240-250.
- Bialystok, E., & Feng, X. (2009). Language proficiency and executive control in proactive interference: Evidence from monolingual and bilingual children and adults. *Brain and Language*, 109(2), 93-100.
- Bialystok, E., & Luk, G. (2012). Receptive vocabulary differences in monolingual and bilingual adults. *Bilingualism: Language and Cognition*, 15(02), 397-401.
- Bialystok, E., Luk, G., Peets, K. F., & Yang, S. (2010). Receptive vocabulary differences in monolingual and bilingual children. *Bilingualism: Language and Cognition*, 13(04), 525-531.
- Bijeljac-Babic, R., Millogo, V., Farioli, F., & Grainger, J. (2004). A developmental investigation of word length effects in reading using a new on-line word identification paradigm. *Reading and Writing*, 17(4), 411-431.
- Binning, J., & Barrett, G. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, 74(3), 478-494.
- Birdsong, D. (2005). Interpreting age effects in second language acquisition. In J.F. Kroll and A. M.B. DeGroot (Eds.) *Handbook of Bilingualism: Psycholinguistic Approaches*, (pp. 109-127). Oxford University Press, London.
- Bireta, T. J., Neath, I., & Surprenant, A. M. (2006). The syllable-based word length effect and stimulus set specificity. *Psychonomic Bulletin & Review*, 13(3), 434-438.
- Bobko, P., & Roth, P. L. (2013). Reviewing, categorizing, and analyzing the literature on black-white mean differences for predictors of job performance: Verifying some perceptions and updating/correcting others. *Personnel Psychology*, 66, 91-126.

- Boring, E. G. (1923). Intelligence as the tests test it. *New Republic* 36, 35-37.
- Brand, Christopher. 1987. "The Importance of General Intelligence." In S. Magil & C. Magil, (Eds.), *Arthur Jensen: Consensus and Controversy*. New York: Falmer.
- Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. In A. Conway, C. Jarrold, M. Kane, A. Miyake, J. Towse (Eds.), *Variation in Working Memory*, (pp. 76-106). Oxford University Press.
- Brenders, P., Van Hell, J. G., & Dijkstra, T. (2011). Word recognition in child second language learners: Evidence from cognates and false friends. *Journal of Experimental Child Psychology*, 109, 383–396.
- Brouwers, S. A., & van de Vijver, F. J. (2012). Intelligence 2.0 in I–O Psychology: Revival or contextualization? *Industrial and Organizational Psychology*, 5(2), 158-160.
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904-911.
- Buehner, M., Krumm, S., Ziegler, M., & Pluecken, T. (2006). Cognitive abilities and their interplay: Reasoning, crystallized intelligence, working memory components, and sustained attention. *Journal of Individual Differences*, 27(2), 57-72.
- Burgess, G. C., Gray, J. R., Conway, A. R., & Braver, T. S. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *Journal of Experimental Psychology: General*, 140(4), 674-692.
- Carrell, P.L. (1984). Evidence of a formal schema in second language comprehension. *Language Learning*, 34, 87-111.

- Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge, England: University of Cambridge Press.
- Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading*, 36(2), 84-95.
- Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin*, 40(3), 153-193.
- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1), 1-22.
- Chen, J. & Gardner, H. (2012). Assessment of intellectual profile: A perspective from Multiple-Intelligences Theory. In D. P. Flanagan and P. L. Harrison (Eds), *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, (3rd ed., pp.145-155). New York: Guilford Press.
- Cheung, H. (1996). Nonword span as a unique predictor of second-language vocabulary language. *Developmental Psychology*, 32(5), 867-873.
- Chuderski, A. (2013). When are fluid intelligence and working memory isomorphic and when are they not? *Intelligence*, 41(4), 244-262.
- Chuderski, A., Taraday, M., Nęcka, E., & Smoleń, T. (2012). Storage capacity explains fluid intelligence but executive control does not. *Intelligence*, 40(3), 278-295.
- Cohen, A. D., & Aphek, E. (1980). Retention of second-language vocabulary overtime: Investigating the role of mnemonic associations. *System*, 8(3), 221-235.
- Colom, R., Abad, F. J., Quiroga, M. Á., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why? *Intelligence*, 36(6), 584-606.

- Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence*, 32(3), 277-296.
- Conway, A. R., Cowan, N., Bunting, M. F., Theriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, 30(2), 163-183.
- Conway, A. R., Kane, M., & Engle, R. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, 7, 547-552.
- Costa, A., Hernández, M., & Sebastián-Gallés, N. (2008). Bilingualism aids conflict resolution: Evidence from the ANT task. *Cognition*, 106(1), 59-86.
- Cowan, N. (2005). Working-memory capacity limits in a theoretical context. In C. Izawa & N. Ohta (eds.), *Human Learning and Memory: Advances in Theory and Applications* (pp.155-175). Erlbaum.
- Cowan, N., Baddeley, A. D., Elliott, E. M., & Norris, J. (2003). List composition and the word length effect in immediate recall: A comparison of localist and globalist assumptions. *Psychonomic Bulletin & Review*, 10(1), 74-79.
- Cowan, N., Day, L., Saults, J. S., Keller, T. A., Johnson, T., & Flores, L. (1992). The role of verbal output time in the effects of word length on immediate memory. *Journal of Memory and Language*, 31(1), 1-17.
- Cowan, N., Rouders, J. N., Blume, C. L., & Saults, J. S. (2012). Models of verbal working memory capacity: what does it take to make them work? *Psychological Review*, 119(3), 480-499.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4), 450-466.

- Dijkstra, T., & Rekké, S. (2010). Towards a localist-connectionist model of word translation. *The Mental Lexicon*, 5(3), 401-420.
- Drasgow, F. (2003). Intelligence and the Workplace. *Handbook of Psychology, Part One*. 107–130. American Psychological Association.
- Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly*, 16(4), 486-514.
- Duncan, J., Seitz, R. J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A., ... & Emslie, H. (2000). A neural basis for general intelligence. *Science*, 289, 457-460.
- Duyck, W., & Brysbaert, M. (2004). Forward and backward number translation requires conceptual mediation in both balanced and unbalanced bilinguals. *Journal of Experimental Psychology: Human Perception and Performance*, 30(5), 889-908.
- Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning*, 43(4), 559-617.
- Engel de Abreu, P. M. (2011). Working memory in multilingual children: Is there a bilingual effect? *Memory*, 19(5), 529-537.
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology: General*, 128(3), 309-331.
- Fagan, J. F. (2000). A theory of intelligence as processing: Implications for society. *Psychology, Public Policy, and Law*, 6(1), 168-179.
- Fagan, J. F., & Holland, C. (2002). Equal opportunity and racial differences in IQ. *Intelligence*, 30, 361–387.

- Fagan, J. F., & Holland, C. (2007). Racial equality in intelligence: Predictions from a theory of intelligence as processing. *Intelligence*, 35, 319-334.
- Fagan, J. F., & Holland, C. (2009). Culture-fair prediction of academic achievement. *Intelligence*, 37, 62-67.
- Fellbaum, C. (1998) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.
Available at <http://www.easydefine.com>, last retrieved: November, 12th, 2016.
- Fernandes, M. A., Craik, F., Bialystok, E., & Kreuger, S. (2007). Effects of bilingualism, aging, and semantic relatedness on memory under divided attention. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 61(2), 128-141.
- Flanagan, D. P., Alonso, V. C., Ortiz, S. O. (2012). The Cross-Battery Assessment approach. In D., Flanagan, & P., Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (3rd ed., pp. 459-483). New York: Guilford Press.
- Flanagan, D. P., & Harrison, P. (2012) (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (3rd ed.). New York: Guilford Press.
- Francis, W. S., & Baca, Y. (2014). Effects of language dominance on item and order memory in free recall, serial recall and order reconstruction. *Memory*, 22(8), 1060-1069.
- Freedle, R. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review*, 73, 1-42.
- Freedle, R., & Kostin, I. (1997). Predicting black and white differential item functioning in verbal analogy performance. *Intelligence*, 24, 417-444.
- Fischer, C. T. (1969). Intelligence defined as effectiveness of approaches. *Journal of Consulting and Clinical Psychology*, 33(6), 668.

- Floyd, R. G., Evans, J. J., & McGrew, K. S. (2003). Relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and mathematics achievement across the school-age years. *Psychology in the Schools, 40*(2), 155-171.
- Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language, 42*(4), 481-496.
- Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory & Cognition, 23*(1), 83-94.
- Gathercole, S. E., & Adams, A. M. (1994). Children's phonological working memory: Contributions of long-term knowledge and rehearsal. *Journal of Memory and Language, 33*(5), 672-688.
- Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language, 28*(2), 200-213.
- Gathercole, S. E., & Baddeley, A. D. (1990). The role of phonological memory in vocabulary acquisition: A study of young children learning new names. *British Journal of Psychology, 81*(4), 439-454.
- Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(1), 84-95.
- Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental Psychology, 28*(5), 887-898.

- Geake, J. G., & Hansen, P. C. (2005). Neural correlates of intelligence as revealed by fMRI of fluid analogies. *Neuroimage*, 26(2), 555-564.
- Geake, J. G., & Hansen, P. C. (2010). Functional neural correlates of fluid and crystallized analogizing. *Neuroimage*, 49(4), 3489-3497.
- Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In Kuczaj, S. A., (Ed.) *Language Development: Language, Thought, and Culture*, Vol. 2 (pp. 301-334). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Glanzer, M., & Duarte, A. (1971). Repetition between and within languages in free recall. *Journal of Verbal Learning and Verbal Behavior*, 10(6), 625-630.
- Glaze, J. A. (1928). The association value of non-sense syllables. *The Pedagogical Seminary and Journal of Genetic Psychology*, 35(2), 255-269.
- Gold, B. T., Kim, C., Johnson, N. F., Kryscio, R. J., & Smith, C. D. (2013). Lifelong bilingualism maintains neural efficiency for cognitive control in aging. *The Journal of Neuroscience*, 33(2), 387-396.
- Goldstein, H., Scherbaum, C., & Yusko, K. (2009). Adverse impact and measuring cognitive ability. In J. Outtz's (Ed.) *Adverse Impact: Implications for Organizational Staffing and High Stakes Testing* (pp. 95-134). New York: Psychology Press.
- Gollan, T.H., & Ferreira, V.S. (2009). Should I stay or should I switch? A cost-benefit analysis of voluntary language switching in young and aging bilinguals. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 35 (3), 640-665.
- Goodenough, F. L. (1949). *Mental Testing, its History, Principles, and Applications*. Oxford, England.

- Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, 6(3), 316-322.
- Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, 1(02), 67-81.
- Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, 36(1), 3-15.
- Gupta, P. (2005). Primacy and recency in nonword repetition. *Memory*, 13(3-4), 318-324.
- Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, 8(3), 179-203.
- Gustafsson, J. E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, 28, 407-434.
- Haier, R. J., Siegel, B. V., Nuechterlein, K. H., Hazlett, E., Wu, J. C., Paek, J., Browning, H. L., & Buchsbaum, M. S. (1988). Cortical glucose metabolic rate correlates of abstract reasoning and attention studied with positron emission tomography. *Intelligence*, 12(2), 199-217.
- Hakuta, K., & Diaz, R. M. (1985). The relationship between degree of bilingualism and cognitive ability: A critical discussion and some new longitudinal data. *Children's Language*, 5, 319-344.
- Hakuta, K., Ferdman, B. M., & Diaz, R. M. (1987). Bilingualism and cognitive development: Three perspectives. In S. E. Rosenberg (Ed.) *Advances in Applied Psycholinguistics*, Vol. 2 (pp. 284-319). New York: Cambridge University Press.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17-27.

- Hebb, D. O. (1949). *The Organization of Behavior*. New York: Wiley & Sons.
- Helms-Lorenz, M., Van de Vijver, F., & Poortinga, Y. (2003). Cross-cultural differences in cognitive performance and Spearman's hypothesis: *g* or *c*? *Intelligence*, 31, 9-29.
- Henry, L. A., & Millar, S. (1991). Memory span increase with age: A test of two hypotheses. *Journal of Experimental Child Psychology*, 51(3), 459-484.
- Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, 26(1), 55-88.
- Hoffman, B. (1962). *The Tyranny of Testing*. Dover Publications.
- Horn, J. & Blankson, N. (2012). Foundations for better understanding cognitive abilities. In D. P. Flanagan and P. L. Harrison (Eds), *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, (3rd ed., pp. 73-98). New York: Guilford Press.
- Horn, J., & Cattell, R. (1966). Refinement of the theory of fluid and crystalized general intelligences. *Journal of Educational Psychology*, 57, 253-270.
- Hossiep, R., Turck, D., & Hasella, M. (1999). Bochumer Matrizentest. BOMAT-advanced-short version. Göttingen: Hogrefe.
- Hull, C. L. (1933). The meaningfulness of 320 selected nonsense syllables. *The American Journal of Psychology*, 45(4), 730-734.
- Hulme, C., Maughan, S., & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language*, 30, 685-701.
- Hulme, C., Neath, I., Stuart, G., Shostak, L., Surprenant, A. M., & Brown, G. D. (2006). The distinctiveness of the word-length effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(3), 586-594.

- Hulme, C., Suprenant, A. M., Bireta, T. J., Stuart, G., & Neath, I. (2004). Abolishing the word-length effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(1), 98-106.
- Hulme, C., Thomson, N., Muir, C., & Lawrence, A. (1984). Speech rate and the development of short-term memory span. *Journal of Experimental Child Psychology*, 38(2), 241-253.
- Ittenbach, R. F., Spiegel, A. N., McGrew, K. S., & Bruininks, R. H. (1992). Confirmatory factor analysis of early childhood ability measures within a model of personal competence. *Journal of School Psychology*, 30(3), 307-323.
- Izawa, H. (1993). The English vocabulary of 21 Japanese adults on a high proficiency level. *JALT Journal*, 15(1), 63-75.
- Jaeggi, S. M., Buschkuhl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, 105(19), 6829-6833.
- Jaeggi, S. M., Buschkuhl, M., Jonides, J., & Shah, P. (2011). Short-and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences*, 108(25), 10081-10086.
- Jaeggi, S. M., Studer-Luethi, B., Buschkuhl, M., Su, Y. F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning-implications for training and transfer. *Intelligence*, 38(6), 625-635.
- Jalbert, A., Neath, I., Bireta, T. J., & Surprenant, A. M. (2011). When does length cause the word length effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(2), 338-353.
- Jalbert, A., Neath, I., & Surprenant, A. M. (2011). Does length or neighborhood size cause the

- word length effect? *Memory & Cognition*, 39(7), 1198-1210.
- Janse, E., & Newman, R. S. (2013). Identifying nonwords: Effects of lexical neighborhoods, phonotactic probability, and listener characteristics. *Language and Speech*, 56(4), 421-441.
- Jensen, A. R. (1998). *The g Factor: The Science of Mental Ability*. Westport, CT: Praeger.
- Jensen, A. R. (2002). Galton's legacy to research on intelligence. *Journal of Biosocial Science*, 34(02), 145-172.
- Johnston, P. (1984). Prior knowledge and reading comprehension test bias. *Reading Research Quarterly*, 19(2), 219-239.
- Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology*, 59, 193-224.
- Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development*, 64(3), 675-687.
- Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33(5), 630-645.
- Jung, R. E., & Haier, R. J. (2007). The Parieto-Frontal Integration Theory (P-FIT) of intelligence: converging neuroimaging evidence. *Behavioral and Brain Sciences*, 30 (02), 135-154.
- Kail, R., & Hall, L. K. (2001). Distinguishing short-term memory from working memory. *Memory & Cognition*, 29(1), 1-9.
- Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin and Review*, 9, 637-671.

- Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working Memory Capacity and Fluid Intelligence Are Strongly Related Constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, 131(1), 66-71.
- Kaushanskaya, M., & Marian, V. (2009a). The bilingual advantage in novel word learning. *Psychonomic Bulletin & Review*, 16(4), 705-710.
- Kaushanskaya, M., & Marian, V. (2009b). Bilingualism reduces native-language interference during novel-word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3), 829-835.
- Kaushanskaya, M., & Yoo, J. (2013). Phonological short-term and working memory in bilinguals' native and second language. *Applied Psycholinguistics*, 34(5), 1005-1037.
- Kaushanskaya, M., Yoo, J., & Van Hecke, S. (2013). Word learning in adults with second-language experience: Effects of phonological and referent familiarity. *Journal of Speech, Language, and Hearing Research*, 56(2), 667-678.
- Keith, T. Z., & Reynolds, M. R. (2010). Cattell–Horn–Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools*, 47(7), 635-650.
- Kena, G., Hussar, W., McFarland, J., de Brey, C., Musu, Gillette, L., Wang, X., Zhang, J., Rathbun, A., et al. (2016). *Conditions of Education 2016: Racial/ethnic enrollment in public schools* (NCES 2016144). U.S. Department of Education.
- King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30(5), 580-602.
- Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, 10(1), 62-102.
- Kobrin, J. L., Sathy, V., & Shaw, E. J. (2007). *A Historical View of Subgroup Performance Differences on the SAT Reasoning Test*. Research Report No. 2006-5. College Board.

- Koch, A. J., McCloy, R. A., Trippe, D. M., & Paullin, C. L. (2012). "Hello, Dolly!": Parameter variation in cloned ability items. In C. J. Paullin (chair), *Practical IRT: Applications in real-word situations*. Paper presented at the Annual Conference of the Society for Industrial & Organizational Psychology, San Diego, CA.
- Kroll, J. F., Dussias, P. E., Bice, K., & Perrotti, L. (2015). Bilingualism, mind, and brain. *Annual Review of Linguistics*, 1(1), 377-394.
- Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, 33(2), 149-174.
- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology*, 86, 148–161.
- Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?!. *Intelligence*, 14(4), 389-433.
- La Pointe, L. B., & Engle, R. W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(6), 1118-1133.
- Lang, J., & Bliese, P. (2012). I–O psychology and progressive research programs on intelligence. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 5, 161–168.
- Lang, J. W., Kersting, M., Hülshager, U. R., & Lang, J. (2010). General mental ability, narrower cognitive abilities, and job performance: The perspective of the nested-factors model of cognitive abilities. *Personnel Psychology*, 63(3), 595-640.

- Langdon, D., & Warrington, E. K. (2000). The role of the left hemisphere in verbal and spatial reasoning tasks. *Cortex*, 36(5), 691-702.
- Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3), 375-419.
- Lewis, M., & Sullivan, M. W. (1985). Infant intelligence and its assessment. In B. Wolman (Ed.), *Handbook of Intelligence: Theories, Measurements, and Applications* (pp.505-600). New York: Wiley
- Li, P. & Zhao, X. (2013). Connectionist bilingual representation. In J. Altarriba, & R. Heredia (Eds.). *Understanding Bilingual Memory: Theory and Applications* (pp. 63-84). New York: Springer.
- Loewen, J. W, Rosser, P., & Katzman, J. (1988). Gender Bias in SAT Items. New Orleans: paper presented at the annual meeting of the American Educational Research Association.
- Lopez, M., & Young, R. K. (1974). The linguistic interdependence of bilinguals. *Journal of Experimental Psychology*, 102(6), 981-983.
- Lovatt, P., Avons, S. E., & Masterson, J. (2000). The word-length effect and disyllabic words. *The Quarterly Journal of Experimental Psychology: Section A*, 53(1), 1-22.
- Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "'General Intelligence,' Objectively Determined and Measured". *Journal of Personality and Social Psychology*, 86 (1), 96-111.
- Luce, P. A. (1986). Neighborhoods of Words in the Mental Lexicon. Research on Speech Perception. Technical Report No. 6.
- Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, 16(5-6), 565-581.

- Luk, G., Green, D. W., Abutalebi, J., & Grady, C. (2011). Cognitive control for language switching in bilinguals: A quantitative meta-analysis of functional neuroimaging studies. *Language & Cognitive Processes*, 27(10), 1479-1488.
- Majerus, S., Van der Linden, M., Mulder, L., Meulemans, T., & Peters, F. (2004). Verbal short-term memory reflects the sublexical organization of the phonological language network: Evidence from an incidental phonotactic learning paradigm. *Journal of Memory and Language*, 51(2), 297-306.
- Malda, M., van de Vijver, F. J., & Temane, Q. M. (2010). Rugby versus soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence*, 38(6), 582-595.
- Mandler, G. (1967). Organization and memory. *Psychology of Learning and Motivation*, 1, 327-372.
- Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing: Within-and between-language competition. *Bilingualism: Language and Cognition*, 6(02), 97-115.
- Martin-Rhee, M. M., & Bialystok, E. (2008). The development of two types of inhibitory control in monolingual and bilingual children. *Bilingualism: Language and Cognition*, 11(01), 81-93.
- Martínez, K., & Colom, R. (2009). Working memory capacity and processing efficiency predict fluid but not crystallized and spatial intelligence: Evidence supporting the neural noise hypothesis. *Personality and Individual Differences*, 46(3), 281-286.
- Masoura, E. V., & Gathercole, S. E. (1999). Phonological short-term memory and foreign language learning. *International Journal of Psychology*, 34(5-6), 383-388.

- Masoura, E. V., & Gathercole, S. E. (2005). Contrasting contributions of phonological short-term memory and long-term knowledge to vocabulary learning in a foreign language. *Memory*, 13(3-4), 422-429.
- Mather, N., & Wendling, B. J. (2012). Linking cognitive abilities to academic interventions for students with specific learning disabilities. In D., Flanagan, & P., Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (3rd ed., pp. 553-581). New York: Guilford Press.
- Matthews, C. R., Riccio, C. A., & Davis, J. L. (2012). The NEPSY-III. In D., Flanagan, & P., Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (3rd ed., pp. 422-435). New York: Guilford Press.
- Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114(3), 449-458.
- McCallum, R. S. & Bracken, B. A. (2012). The Universal Nonverbal Intelligence Test. In D., Flanagan, & P., Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (3rd ed., pp. 357-375). New York: Guilford Press.
- McClelland, J. L. (2000). Connectionist models of memory. In E. Tulving & F.I.M.. Craik (Eds.) *The Oxford Handbook of Memory*, (pp. 583-596). New York: Oxford University Press.
- McClelland, J. L. & Cleeremans, A. (2009). Connectionist models. In T. Byrne, A. Cleeremans, & P. Wilken (Eds.) *Oxford Companion to Consciousness*. New York: Oxford University Press.
- McDonough, C., Song, L., Hirsh-Pasek, K., Golinkoff, R. M., & Lannon, R. (2011). An image is worth a thousand words: Why nouns tend to dominate verbs in early word learning. *Developmental Science*, 14(2), 181-189.

- McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37(1), 1-10.
- McGrew, K. S., & Wendling, B. J. (2010). Cattell–Horn–Carroll cognitive-achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools*, 47(7), 651-675.
- McNulty, J. A. (1965). An analysis of recall and recognition processes in verbal learning. *Journal of Verbal Learning and Verbal Behavior*, 4(5), 430-436.
- McNulty, J. A. (1966). The measurement of “adopted chunks” in free recall learning. *Psychonomic Science*, 4(1), 71-72.
- Medler, D. A., & Binder, J.R. (2005). MCWord: An On-Line Orthographic Database of the English Language. <http://www.neuro.mcw.edu/mcword/>.
- Meeker, M. (1985). Toward a psychology of giftedness: A concept in search of measurement. In B. Wolman (Ed.), *Handbook of Intelligence: Theories, Measurements, and Applications*, (pp. 787-799). New York: John Wiley.
- Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50(9), 741-749.
- Morales, J., Calvo, A., & Bialystok, E. (2013). Working memory development in monolingual and bilingual children. *Journal of Experimental Child Psychology*, 114(2), 187-202.
- Morford, J. P., Wilkinson, E., Villwock, A., Piñar, P., & Kroll, J. F. (2011). When deaf signers read English: Do written words activate their sign translations? *Cognition*, 118(2), 286-292.
- Morton, J. B., & Harper, S. N. (2007). What did Simon say? Revisiting the bilingual advantage. *Developmental Science*, 10(6), 719-726.

- Mueller, S. T., Seymour, T. L., Kieras, D. E., & Meyer, D. E. (2003). Theoretical implications of articulatory duration, phonological similarity, and phonological complexity in verbal working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6), 1353-1380.
- Naglieri, J. A., (2005). The cognitive assessment system. In D. P. Flanagan and P. L. Harrison (Eds.) *Contemporary Intellectual Assessment* (2nd Edition) (pp. 441-460). New York: Guilford.
- Naglieri, J. A., & Das, J. P. (1990). Planning, attention, simultaneous, and successive (PASS) cognitive processes as a model for intelligence. *Journal of Psychoeducational Assessment*, 8(3), 303-337.
- Naglieri, J. A., & Das, J. P. (1997). Intelligence revised. In R. Dillon (Ed.), *Handbook on Testing* (pp. 136-163). Westport, CT: Greenwood Press.
- Naglieri, J. A., Das, J. P., & Goldstein, S. (2012). Planning, attention, simultaneous, successive: A cognitive-processing-based theory of intelligence. In D., Flanagan, & P., Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (3rd ed., pp. 178-195). New York: Guilford Press.
- Naglieri, J. A., & Otero, T. M. (2012). The Cognitive Assessment System: From theory to practice, in Eds. Flanagan, D. P. & Harrison, P. L. *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (3rd ed., pp. 376-399). The Guilford Press.
- 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.
- Neisser, U., Boodoo, G., Bouchard Jr, T. J., Boykin, A. W., Brody, N., Ceci, S. J., & Urbina, S. (1996). Intelligence: knowns and unknowns. *American Psychologist*, 51(2), 77.

- New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, 13(1), 45-52.
- Nott, C. R., & Lambert, W. E. (1968). Free recall of bilinguals. *Journal of Verbal Learning and Verbal Behavior*, 7(6), 1065-1071.
- Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3), 411-421.
- Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H. M. (2005). Working memory and intelligence--their correlation and their relation: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, 131(1), 61-65.
- Oberauer, K., Süß, H. M., Wilhelm, O., & Wittmann, W. W. (2008). Which working memory functions predict intelligence? *Intelligence*, 36(6), 641-652.
- Oller, D. K. & Eilers, R. E. (2002). *Language and Literacy in Bilingual Children, Multilingual Matters*. Clevedon, England.
- Olmedo, E. L. (1981). Testing linguistic minorities. *American Psychologist*, 36(10), 1078-1085.
- Ortiz, S. O., Ochoa, S. H., & Dynda, A. M. (2012). Testing with culturally and linguistically diverse populations: Moving beyond the verbal-performance dichotomy into evidence-based practice. *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (3rd ed, pp. 526-552). Guilford Press.
- Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, 25(1), 46-59.

- Paap, K. R., Johnson, H. A., & Sawi, O. (2014). Are bilingual advantages dependent upon specific tasks or specific bilingual experiences? *Journal of Cognitive Psychology*, 26(6), 615-639.
- Paap, K. R., Johnson, H. A., & Sawi, O. (2015). Bilingual advantages in executive functioning either do not exist or are restricted to very specific and undetermined circumstances. *Cortex*, 69, 265-278.
- Papagno, C., Valentine, T., & Baddeley, A. (1991). Phonological short-term memory and foreign-language vocabulary learning. *Journal of Memory and Language*, 30, 331-347.
- Papagno, C., & Vallar, G. (1992). Phonological short-term memory and the learning of novel words. The effect of phonological similarity and item length. *The Quarterly Journal of Experimental Psychology*, 44A(1), 47-67.
- Papagno, C., & Vallar, G. (1995). Verbal short-term memory and vocabulary learning in polyglots. *The Quarterly Journal of Experimental Psychology*, 48(1), 98-107.
- Peal, E., & Lambert, W. E. (1962). The relation of bilingualism to intelligence. *Psychological Monographs: General and Applied*, 76(27), 1-23.
- Perani, D., Abutalebi, J., Paulesu, E., Brambati, S., Scifo, P., Cappa, S. F., & Fazio, F. (2003). The role of age of acquisition and language usage in early, high-proficient bilinguals: An fMRI study during verbal fluency. *Human Brain Mapping*, 19(3), 170-182.
- Phillips, L. H., & Hamilton, C. (2001). The working memory model in adult aging research. In J. Andrade (Ed.), *Working Memory in Perspective* (pp. 101-125). Psychology Press.
- Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication*, 4(1), 75-95.

- Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*(1), 153-172.
- Poortinga, Y. H., & Van de Vijver, F. J. (2004). Cultures and cognition: Performance differences and invariant structures. In R. J. Sternberg & E. L. Grigorenko (Eds), *Culture and Competence: Contexts of Life Success* (pp. 139-162). Washington, DC, US: American Psychological Association.
- Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*(1), 160.
- Pyburn, K. M., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity–validity dilemma: Overview and legal context. *Personnel Psychology, 61*(1), 143-151.
- Ratiu, I., & Azuma, T. (2015). Working memory capacity: is there a bilingual advantage? *Journal of Cognitive Psychology, 27*(1), 1-11.
- Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*(1), 1-48.
- Raven, J., Raven, J.C., & Court, J.H. (2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. San Antonio, TX: Harcourt Assessment.
- Redick, T. S., Unsworth, N., Kelly, A. J., & Engle, R. W. (2012). Faster, smarter? Working memory capacity and perceptual speed in relation to fluid intelligence. *Journal of Cognitive Psychology, 24*(7), 844-854.
- Resing, W. C., & Elliott, J. G. (2011). Dynamic testing with tangible electronics: Measuring children's change in strategy use with a series completion task. *British Journal of Educational Psychology, 81*(4), 579-605.

- Resing, W. C., Tunteler, E., de Jong, F. M., & Bosma, T. (2009). Dynamic testing in indigenous and ethnic minority children. *Learning and Individual Differences*, 19 (4), 445-450.
- Rodgers, T. S. (1969). On measuring vocabulary difficulty. An analysis of item variables in learning Russian-English vocabulary pairs. *International Review of Applied Linguistics in Language Teaching*, 7(4), 327-343.
- Roediger III, H. L. (1980). Memory metaphors in cognitive psychology. *Memory & Cognition*, 8(3), 231-246.
- Rohde, D. L. T. & Plaut, D. C. (2003). Connectionist models of language processing. *Cognitive Studies*, 10, 10-28.
- Romani, C., McAlpine, S., Olson, A., Tsouknida, E., & Martin, R. (2005). Length, lexicality, and articulatory suppression in immediate recall: Evidence against the articulatory loop. *Journal of Memory and Language*, 52(3), 398-415.
- Roodenrys, S., & Hinton, M. (2002). Sublexical or lexical effects on serial recall of nonwords?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(1), 29-33.
- Roodenrys, S., Hulme, C., & Brown, G. (1993). The development of short-term memory span: Separable effects of speech rate and long-term memory. *Journal of Experimental Child Psychology*, 56(3), 431-442.
- Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. M. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(6), 1019-1034.
- Rosser, P. (1989) *The SAT Gender Gap: Identifying the Causes*. Washington, D.C.: Center for Women Policy Studies.
- Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences

- in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, 54(2), 297-330.
- Sabet, J., Scherbaum, C., & Goldstein, H. (2013). Examining the potential of neuropsychological intelligence tests for predicting academic performance and reducing racial/ethnic test scores differences. In F. Metzger's (Ed.) *Neuropsychology: New Research* (pp. 1-24). New York: Nova Publishers.
- Sanchez, C. A., Wiley, J., Miura, T. K., Colflesh, G. J., Ricks, T. R., Jensen, M. S., & Conway, A. R. (2010). Assessing working memory capacity in a non-native language. *Learning and Individual Differences*, 20(5), 488-493.
- Scherbaum C., Goldstein, H., Ryan, R., Agnello, P., Yusko, K. & Hanges, P. (2015). New developments in intelligence theory and assessment: Implications for personnel selection In I. Nikolaou & J. K. Oostrom (Eds.) *Employee Recruitment, Selection, and Assessment. Contemporary Issues for Theory and Practice* (1st ed., pp. 99-116). Psychology Press.
- Scherbaum, C.A., Goldstein, H.W., Yusko, K.P., Ryan, R., and Hanges, P.J. (2012). Intelligence 2.0: Reestablishing a research program on g in I-O Psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 5(2), 128-148.
- Schmidt, F. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance*, 15, 187–210.
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262-274.

- Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. Flanagan & P. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (3rd ed., pp. 99–144). New York: Guilford.
- Schroeder, S. R., & Marian, V. (2012). A bilingual advantage for episodic memory in older adults. *Journal of Cognitive Psychology*, 24(5), 591-601.
- 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: Section A*, 51(2), 283-304.
- Service, E., & Craik, F. I. M. (1993). Differences between young and older adults in learning a foreign vocabulary. *Journal of Memory and Language*, 32(5), 608-623.
- Service, E., Simola, M., Metsänheimo, O., & Maury, S. (2002). Bilingual working memory span is affected by language skill. *European Journal of Cognitive Psychology*, 14(3), 383-408.
- Shook, A., & Marian, V. (2012). Bimodal bilinguals co-activate both languages during spoken comprehension. *Cognition*, 124(3), 314-324.
- Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*, 138(4), 628-654.
- Singer, J. K., & Lichtenberger, E. O. (2012). The Kaufman Assessment Battery for Children-Second Edition and the Kaufman Test of Educational Achievement. In D. P. Flanagan and P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, (3rd ed., pp. 269-296). New York: Guilford Press.
- Singleton, D. (2005). The Critical Period Hypothesis: A coat of many colours. *International Review of Applied Linguistics in Language Teaching*, 43(4), 269-285.

- Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment, 13*(2-3), 108-131.
- Spearman, C. (1927). *The Abilities of Man*. New York: Macmillan.
- Stauffer, J. M., Ree, M. J., & Carretta, T. R. (1996). Cognitive-components tests are not much more than g: An extension of Kyllonen's analyses. *The Journal of General Psychology, 123*(3), 193-205.
- Sternberg, R. (1981). Intelligence and non-entrenchment. *Journal of Educational Psychology, 73*, 1-16.
- Sternberg, R. J. (2004). Culture and intelligence. *American Psychologist, 59*(5), 325-338.
- Sternberg, R. J. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence, 34*, 321-350.
- Sternberg, R. J., & Detterman, D. K. (1986). *What is Intelligence?: Contemporary Viewpoints on its Nature and Definition*. Praeger Pub Text.
- Sternberg, R. J., Ferrari, M., Clinkenbeard, P., & Grigorenko, E. L. (1996). Identification, instruction, and assessment of gifted children: A construct validation of a triarchic model. *Gifted Child Quarterly, 40*(3), 129-137.
- Storkel, H. L. (2004). Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density. *Journal of Speech, Language, and Hearing Research, 47*(6), 1454-1468.
- Storkel, H. L., Armbrüster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research, 49*, 1175-1192.

- Storkel, H. L., Bontempo, D. E., Aschenbrenner, A. J., Maekawa, J., & Lee, S. Y. (2013). The effect of incremental changes in phonotactic probability and neighborhood density on word learning by preschool children. *Journal of Speech, Language, and Hearing Research, 56*(5), 1689-1700.
- Storkel, H. L., Bontempo, D. E., & Pak, N. S. (2014). Online learning from input versus offline memory evolution in adult word learning: Effects of neighborhood density and phonologically related practice. *Journal of Speech, Language, and Hearing Research, 57*(5), 1708-1721.
- Storkel, H. L., & Lee, S. Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes, 26*(2), 191-211.
- Suárez, L., Tan, S. H., Yap, M. J., & Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin & Review, 18*(3), 605-611.
- Süß, H. M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability - and a little bit more. *Intelligence, 30*(3), 261-288.
- Surprenant, A. M., Brown, M. A., Jalbert, A., Neath, I., Bireta, T. J., & Tehan, G. (2011). Backward recall and the word length effect. *The American Journal of Psychology, 124*(1), 75-86.
- Takano, Y., & Noda, A. (1993). A temporary decline of thinking ability during foreign language processing. *Journal of Cross-Cultural Psychology, 24*(4), 445-462.
- Takano, Y., & Noda, A. (1995). Interlanguage dissimilarity enhances the decline of thinking ability during foreign language processing. *Language Learning, 45*(4), 657-681.

- Tehan, G., Hendry, L., & Kocinski, D. (2001). Word length and phonological similarity effects in simple, complex, and delayed serial recall tasks: Implications for working memory. *Memory*, 9(4-6), 333-348.
- Tehan, G., & Tolan, G. A. (2007). Word length effects in long-term memory. *Journal of Memory and Language*, 56(1), 35-48.
- Thorn, A. S., & Frankish, C. R. (2005). Long-term knowledge effects on serial recall of nonwords are not exclusively lexical. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(4), 729-735.
- Thorn, A. S., & Gathercole, S. E. (2001). Language differences in verbal short-term memory do not exclusively originate in the process of subvocal rehearsal. *Psychonomic Bulletin & Review*, 8(2), 357-364.
- Thorn, A. S., Gathercole, S. E., & Frankish, C. R. (2002). Language familiarity effects in short-term memory: The role of output delay and long-term knowledge. *The Quarterly Journal of Experimental Psychology: Section A*, 55(4), 1363-1383.
- van de Vijver, F. J. (1997). Meta-analysis of cross-cultural comparisons of cognitive test performance. *Journal of Cross-Cultural Psychology*, 28, 678-709.
- van de Vijver, F. J., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, 13(1), 29-37.
- van de Vijver, F. J., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, 54(2), 119-135.
- Van den Noort, M. W., Bosch, P., & Hugdahl, K. (2006). Foreign language proficiency and working memory capacity. *European Psychologist*, 11(4), 289-296.

- van der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review*, 113(4), 842-861.
- van der Maas, H. L., Kan, K. J., & Borsboom, D. (2014). Intelligence is what the intelligence test measures. Seriously. *Journal of Intelligence*, 2(1), 12-15.
- Van Hell, J. G., & Mahn, A. C. (1997). Keyword mnemonics versus rote rehearsal: Learning concrete and abstract foreign words by experienced and inexperienced learners. *Language Learning*, 47(3), 507-546.
- Van Heuven, W. J., Dijkstra, T., & Grainger, J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language*, 39(3), 458-483.
- Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology*, 61(4), 871-925.
- Vejnović, D., Milin, P., & Zdravković, S. (2010). Effects of proficiency and age of language acquisition on working memory performance in bilinguals. *Psihologija*, 43(3), 219-232.
- Vernon, P. E. (1950). *The Structure of Human Abilities*. London: Methuen.
- Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, 9(4), 325-329.
- Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, 40(1), 47-62.
- Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, 68(1), 306-311.

- Vitevitch, M. S., Storkel, H. L., Francisco, A. C., Evans, K. J., & Goldstein, R. (2014). The influence of known-word frequency on the acquisition of new neighbours in adults: Evidence for exemplar representations in word learning. *Language, Cognition and Neuroscience*, 29(10), 1311-1316.
- Wager, T. D., & Smith, E. E. (2003). Neuroimaging studies of working memory. *Cognitive, Affective, & Behavioral Neuroscience*, 3(4), 255-274.
- Wasserman, J. D., & Tulskey, D. S. (2005). A history of intelligence assessment. In D. P. Flanagan and P. L. Harrison (Eds), *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, (2nd ed., pp3-22). New York: Guilford Press.
- Welsh Jr., J. R., Kucinkas, S. K., & Curran, L. T. (1990). Armed services vocational battery (ASVAB): Integrative review of validity studies. Operational Technologies Corp San Antonio TX.
- Wolf, M. K., Herman, J. L., & Dietel, R. (2010, Spring). Improving the validity of English language learner assessment systems (CRESST Policy Brief No. 10 - Full Report). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Wolman, B. (1985). *Handbook of Intelligence: Theories, Measurements, and Applications*. New York: Wiley.
- Wu, Y. J., & Thierry, G. (2010). Chinese–English bilinguals reading English hear Chinese. *The Journal of Neuroscience*, 30(22), 7646-7651.
- Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, 60(4), 502-529.

- Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5), 971-979.
- Zhao, X., & Li, P. (2010). Bilingual lexical interactions in an unsupervised neural network model. *International Journal of Bilingual Education and Bilingualism*, 13(5), 505-524.
- Zipf, G. K. (1935). *The Psycho-Biology of Language*. Oxford, England: Houghton Mifflin.