Summer 2018

Word Counts in Response to Cognitively Demanding Essay Prompts as Reflections of General Cognitive Ability and Broad Cognitive Abilities

Michael Beaumont Armstrong
Old Dominion University

WORD COUNTS IN RESPONSE TO COGNITIVELY DEMANDING ESSAY PROMPTS AS REFLECTIONS OF GENERAL COGNITIVE ABILITY AND BROAD COGNITIVE ABILITIES

by

Michael Beaumont Armstrong
B.A. May 2013, Western Kentucky University
M.S. December 2015, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

INDUSTRIAL/ORGANIZATIONAL PSYCHOLOGY

OLD DOMINION UNIVERSITY
August 2018

Approved by:

Richard N. Landers (Director)

Konstantin P. Cigularov (Member)

Ryan L. Klinger (Member)

ABSTRACT

WORD COUNTS IN RESPONSE TO COGNITIVELY
DEMANDING ESSAY PROMPTS AS REFLECTIONS OF GENERAL COGNITIVE
ABILITY AND BROAD COGNITIVE ABILITIES

Michael Beaumont Armstrong
Old Dominion University, 2018
Director: Dr. Richard N. Landers

Natural language processing techniques can be used to analyze text and speech data. These techniques have been applied within many domains to date but have only recently been examined in the domain of personnel assessment. By linking workplace-relevant constructs such as general cognitive ability (GCA) to natural language processing outcomes such as word counts, a foundation for language-based psychological assessment of those abilities can be laid. Over 400 participants were recruited through Amazon Mechanical Turk to write cognitively demanding essays and complete a battery of cognitive tests. Essays were analyzed using Linguistic Inquiry and Word Count (LIWC). Structural equation modeling was used to examine the relationship between GCA and word count categories as well as the relationship between broad cognitive abilities and word count categories. Latent GCA added incremental prediction of unique word use over latent verbal ability and incremental prediction of preposition use over latent short-term memory. Although not statistically significant, latent GCA and latent verbal ability related to various LIWC word count categories the strongest out of the abilities measured, yielding small to medium effect sizes in both positive and negative directions. Latent short-term memory and latent fluid reasoning were weakly related or unrelated to the LIWC word count categories observed. Word counting approaches to natural language processing may partially express GCA and latent verbal ability, but not latent short-term memory and latent fluid reasoning in cognitively demanding essay contexts.

This dissertation is dedicated to God, my wife Beth, and my family, all of whom supported me through this long and arduous process.

ACKNOWLEDGEMENTS

I would first like to thank my advisor and mentor Richard Landers. I have learned so much from you over the past 5 years. I don't know where I would have been at this point in my career if I had not ended up being your student, but I am glad that I ended up with you. Because of your instruction, I was equipped to complete this project, as well as many others along the way. Thank you for exposing me to new areas of research, applied experiences, and different perspectives of how psychology, science, and the world work.

Thank you to my committee members, Konstantin Cigularov and Ryan Klinger, for your excellent insights and being flexible in working with me in a crazy season of life. Thank you to all of the faculty that have instructed and guided me throughout the years, especially Konstantin, Debbie Major, Xiaoxiao Hu, and Don Davis. Thank you to Peggy Kinard, Linda Forchas, Mary Boswell, Jim Bliss, and Michelle Kelley for always going out of your way to make sure that things are running smoothly for me and for every other student.

Thank you to my TNTLAB mates, Andrew Collmus, Elena Auer, and Sebastian Marin for your help through tough problems, listening ears, and sympathetic hearts on this project and others. Y'all are the best! I wish you the best as you finish up your degrees in faraway places!

Thank you to my parents, Mike and Cuba Armstrong, for continually supporting me in so many different ways, and without whom this project would not have been possible.

Thank you to my wife, Beth. I would not have worked as hard on this if it weren't for you. I would not be done with this project if it weren't for you. I might not have even moved to Virginia to pursue a PhD if it weren't for you. Thank you for being awesome, for being an inspiration, for being fun when I had a tough day, for washing dishes when I had to work, and for keeping me company when I had to work late and deadlines loomed. I love you.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

INTRODUCTION

Natural language processing (NLP) techniques can be used to process and analyze text and audio data. NLP is a "range of computational techniques for the automatic analysis and representation of the human language" (Cambria & White, 2014, p. 48). These techniques may be used to analyze language at a variety of levels ranging from the syllables within a word or root word up to full sentences and discourse (Liddy, 1998; Feldman, 1999). Recently, NLP has been used in a variety of artificial intelligence technologies such as Google Search, IBM's Watson, and Apple's Siri (Cambria & White, 2014). Previously, NLP has been most frequently researched and practiced by computer scientists and computational linguists, but now, industrial-organizational (I-O) psychologists are beginning to take advantage of this technology to aid in workplace psychological assessment. The impetus for this sudden increased interest is likely the myriad text data available in the workplace that often remains underutilized: employee job application blanks, resumes, cover letters, employee emails, open-ended responses to employee engagement survey items, and employee writing samples, among others. With advances in audio transcription technology, I-Os may even have access to transcribed interview data, phone calls, and recorded employee conversations that were previously cost prohibitive. Through NLP, I-Os can quantify and analyze these historically qualitative data sources to derive new meaning in assessment contexts.

One possible application domain of I-O for NLP is personnel assessment, the use of NLP to measure desired knowledge, skills, abilities, and other characteristics (KSAOs) among job applicants and incumbents. Researchers have begun to explore the potential of using NLP to assess KSAOs typically examined in personnel assessment contexts such as personality, cognitive abilities, leadership skills, and communication skills (Park et al., 2015; Campion,

Campion, Campion, & Reider, 2016; Weaver, 2017). Text samples can easily be obtained via essays, resumes, and social media and analyzed using NLP. Analysis of existing text can reduce the amount of time applicants spend testing. Further, with less administrative and rating costs, organizations can save both time and funds.

One of the most accessible and most studied NLP frameworks of potential relevance to the measurement of KSAOs is Linguistic Inquiry and Word Count (LIWC; Pennebaker, Boyd, Jordan, & Blackburn, 2015). LIWC is a computer program and framework for categorizing different words that people use in speech or writing. Individual words are mapped onto different descriptive and psychological categories and counted each time they are used in a text sample. Although this technique may seem somewhat crude in comparison to best practices for psychometric test development, language is the basis for translating a person's inner thoughts and feelings into a form others can understand (Tausczik & Pennebaker, 2010). The words people use are behavioral in nature: they result from the interaction of personal characteristics (i.e., who they are, what they know, what they care about, how they feel, their relationships with other people) and the situations in which they exist. Analyzing these words can provide an opportunity to understand the minds and behaviors of the people speaking or writing them. LIWC consists of various word categories including, for example, the number of pronouns used in the writing sample, the number of common verbs used, articles, prepositions, affect-related words (e.g., "happy" or "cried"), social process words (e.g., "family" or "friend"), cognitive process words (e.g., "think" or "know"), and a variety of other topical categories covering work, leisure, religion, and death to name a few. LIWC and other similar word-counting approaches to NLP have a strong core of psychological research from which to base new developments in assessment (Short, McKenny, & Reid, 2018) compared with newer, less-studied NLP developments such as latent Dirichlet allocation (Blei, Ng, & Jordan, 2003) and deep learning (LeCun, Bengio, & Hinton, 2015). Decades of research support the correlations between the language categories in LIWC and psychological variables and processes (Tausczik & Pennebaker, 2010; Short et al., 2018), providing a theoretical foundation for text analysis without the need to develop and validate new word categories for each new experimental population. Additionally, LIWC is for the most part transparent, easy to use, and affordable.

Little research has investigated the relationship between general cognitive ability (GCA) and NLP variables (Weaver, 2017). GCA, an attribute of individuals that enables the correct or appropriate processing of mental information for successful performance on a given task (Carroll, 1993), is the strongest single predictor of job performance (Schmidt & Hunter, 1998). Essentially, all desirable tasks require some degree of cognition, so GCA is relevant to some extent in all work tasks (Carroll, 1993; Gottfredson, 1997b). Some research has theoretically tied NLP variables with cognition and cognitive *processes* in general (Tausczik & Pennebaker, 2010), which are distinct from GCA (Carroll, 1993), but cognitive *abilities* have only been linked to NLP variables directly in one available study (Weaver, 2017). Because both language interpretation and production require complex cognition, GCA is highly relevant in language contexts (Carroll, 1993, Schneider & McGrew, 2012). Being able to use NLP to assess GCA could remove the need for applicants to exert time and effort on cognitive test batteries, which are perceived only somewhat favorably by applicants despite being highly job-relevant and valid predictors of job performance (Hausknecht, Day, & Thomas, 2004).

A body of prior research suggests that broad cognitive abilities may be related to LIWC word categories. For example, vocabulary knowledge is moderately correlated with the proportion of unique words used within a speech sample (Kemper & Sumner, 2001). Vocabulary knowledge is a facet of the broad cognitive ability verbal ability, the "depth and breadth of knowledge relating to verbal and language skills in one's native language" (Stanek & Ones, 2018, p. 375). Higher vocabulary knowledge is a prerequisite for using a greater variety of words. A person cannot use a variety of words if he or she does not first know a variety of words. Other broad cognitive abilities that may be related to LIWC word categories include short-term memory ability and fluid reasoning ability (Kemper & Sumner, 2001; Graesser, McNamara, Louwerse, & Cai, 2004; Tausczik & Pennebaker, 2010).

In the present study, GCA is hypothesized to be positively related to various NLP variables (Weaver, 2017). Broad cognitive abilities such as verbal ability, short-term memory, and fluid reasoning are also hypothesized to be positively related to specific NLP variables. Both GCA and broad abilities should predict word use independently of one another, each providing unique incremental prediction of variance over the other. By exploring the influence of GCA and broad cognitive abilities on language, a firmer foundation for the assessment of GCA via NLP may be established.

Measuring Language Use via Natural Language Processing

Language, specifically the words that people use to express themselves when communicating with others, has been linked to a variety of individual differences, including mental health, personality, emotions, and writing skill (Tausczik & Pennebaker, 2010; Attali & Burnstein, 2006). Word choice and frequency have been studied primarily in the LIWC and automatic essay scoring research literatures. For example, positive emotion words (e.g., happy, pretty, good) are used more often when writing about amusing memories and negative emotion words (e.g., hate, nervous, cry) are used more often when writing about sad memories (Kahn, Tobin, Massey, & Anderson, 2007). Yarkoni (2010) found a variety of correlations among

LIWC word categories and the Big Five personality traits in a sample of several hundred bloggers. For example, measures of agreeableness correlated positively with the use of first person plural pronouns (e.g., we, ours, us) and negative emotion words were correlated positively with measures of neuroticism. Negative emotion words were negatively correlated with measures of conscientiousness, suggesting that conscientious bloggers are careful not to use sad, angry, or anxious words when blogging. Additionally, the Educational Testing Service utilizes word choice in student essays to automatically evaluate writing skill in the Graduate Record Examination, Test of English as a Foreign Language, and Graduate Management Admissions Test (Deane, 2013). Repeating the same words indicates lesser writing skill, whereas using more sophisticated words and topic-appropriate words indicates greater writing skill (Attali & Burnstein, 2006; Deane, 2013). In general, it appears that there is agreement within the literature that language often reflects individual differences. What remains unclear is precisely how.

NLP can be used to measure the complexity of both words and sentences within a given body of text, either descriptively or to infer characteristics of its writer. The complexity of such words is referred to as “lexical complexity” (Attali & Bernstein, 2006), whereas the complexity of sentences is called “syntactical complexity” (Kemper & Sumner, 2001; Graesser et al., 2004; Graesser, McNamara, & Kulikowich, 2011). The lexical complexity of a word increases as the number of morphemes and syllables needed to form the word increases (Carlisle & Stone, 2005). Accordingly, as the length of words increases, the words themselves become more complex, because adding morphemes and syllables necessarily adds more characters to a word. Reflecting an application of this theory, two common measurement created by NLP techniques are average word length and average number of syllables and morphemes per word. In contrast, the syntactical complexity of a sentence increases as the number of phrases embedded within the sentence increases (Kemper & Sumner, 2001). As the number of embedded phrases within a sentence increases, the length of the sentence also increases. Additionally, as more words are added per phrase or clause, the complexity of a sentence increases. Punctuation marks and connective words like conjunctions (e.g., and, but, or) and prepositions (e.g., of, to, in) often mark the boundaries of phrases or clauses within a sentence (Graesser et al., 2011). Accordingly, other common NLP techniques are counting the number of words per sentence and counting markers of clauses such as punctuation, conjunctions, and prepositions.

By examining the lexical and syntactical complexity of workplace language-based data sources, I-O psychologists can extract new meaning from qualitative data. Lexical and syntactical complexity are already accounted for in the automatic assessment of student writing skill (Deane, 2013), adding a quantitative aspect to historically qualitative evaluations of writing skill. Likewise, I-O psychologists can utilize automatic essay scoring as an assessment of job applicant writing skill, a valuable communication skill in the modern workplace. Lexical complexity and syntactical complexity may be indicators of other psychological constructs as well. These language features may possibly be tied to knowledge of grammar, sentence structure, and vocabulary. Additionally, it may be possible that lexical and syntactical complexity are aspects of linguistic style, reflecting individual differences like personality (Pennebaker & King, 1999). Some research also indicates that deceptive language is less complex, as honest speakers and writers do not need to focus on both maintaining a lie while also producing language (Newman, Pennebaker, Berry, & Richards, 2003).

Overall, NLP has much potential for use in personnel assessment. Language has already been established as a behavior caused by numerous individual differences. Using techniques to measure the constructs causing language enables researchers to utilize a variety of new text sources. Triangulating individual difference assessments on traditional tests with NLP-based assessments may also improve measurement reliability and validity. For example, the results of a personality survey may be compared with an NLP-based personality assessment as a type of reliability assessment. Text-based data that are traditionally evaluated qualitatively such as resumes or letters of recommendation may be evaluated quantitatively if measurement is accurate enough. Additionally, it is possible that data may not even need to be formally collected from the applicant. Pre-existing data, such as social media data, may be mined and analyzed without having to administer a test. NLP-based personnel assessment like this has already been demonstrated successfully; Campion and colleagues (2016) assessed a variety of constructs based on applicant accomplishment records (i.e., essays; Hough, 1984), reducing the need for human evaluators while maintaining valid assessment. The authors extracted words and phrases from the applicant text samples and combined them into similar categories which were then used to predict human ratings of each essay. Mean NLP-based ratings were nearly identical to human ratings of each construct, but with smaller standard deviations, demonstrating that NLP-based ratings could potentially supplement or replace human raters. Further, Campion and colleagues found no additional adverse impact in the NLP-based ratings beyond the pre-existing selection system, demonstrating that the benefits of NLP-based ratings come at no decrease in test fairness. Campion and colleagues assessed communication skills, critical thinking, people skills, leadership skills, managerial skills, and factual knowledge using their NLP-based approach. In another study, Weaver (2017) sought to assess a variety of psychological constructs using a LIWC-based approach. Weaver demonstrated ties among NLP variables and a variety of constructs including impression management, job performance, and general cognitive ability, though many effect sizes were small and sometimes contrary to hypothesized directions. Regardless, LIWC demonstrates some potential for indicating psychological constructs.

Linguistic Inquiry and Word Count (LIWC)

One promising approach to NLP-based assessment is a LIWC-based approach. As previously discussed, LIWC is both a text analysis program and theoretical framework for the psychological study of language. LIWC was initially developed to analyze changes in health and thinking via writing interventions (Pennebaker, 1993; Pennebaker & Francis, 1996). Text analysis in psychological research before that time often consisted of manually coding participant text samples (Pennebaker, 1993; Tausczik & Pennebaker, 2010). With small sample sizes, this may still have been acceptable, but researchers found that using human raters was unreliable, slow, expensive, and depending on the content of the texts, mentally harmful (Tausczik & Pennebaker, 2010). To address these problems, researchers sought to automate the analysis of text. Since the development of this automated text analysis program, additional studies have expanded upon and refined the theoretical underpinnings of LIWC. Originally beginning with 61 categories of words and a dictionary of 2000 words (Pennebaker & Francis, 1996), LIWC now contains a total of 125 word categories, 8 summary indices, and an internal dictionary of nearly 6,400 words, spanning topics from cognitive and perceptual topics to affective and biological processes (Pennebaker et al., 2015).

LIWC is theoretically supported by both its development process and decades of research providing evidence of its validity as measures of psychological traits. The initial word generation for LIWC began with subject matter experts identifying different dimensions of language that were of interest in research, such as negative or positive emotional expression (Pennebaker & Francis, 1996). To create the dictionary for each dimension or category of LIWC, the researchers utilized a thesaurus and dictionaries to generate words used in each category. Additionally, the researchers referenced psychological questionnaires to see what words were used to measure each dimension, as well as previously collected text data regarding each topic, to complete the initial LIWC dictionary. In both the initial LIWC framework and the most recent (Pennebaker et al., 2015), a content validation approach was used to determine the fit of each word to each category. In the most recent revision, four to eight judges rated whether a given word should be included in a given category. Although rater agreement generally exceeded 90%, if a majority of judges could not agree on a word even after consulting additional resources to determine a word's meaning or use, the word was removed from the dictionary (Pennebaker & Francis, 1996). After compiling the words and agreeing on their categorization, internal consistency statistics were computed for each word within a given category (Pennebaker et al., 2015). Words that detrimentally affected the internal consistency of the category were added to a list, which was reviewed by two to eight judges to determine if the words should be omitted from the category. This entire process of word generation, categorization, and reduction were repeated to catch any possible mistakes or oversights, after which two judges reviewed the final dictionary. This approach follows best practices for developing valid psychometric measures, treating words within a category similarly to items on a scale (Crocker & Algina, 1986). Pennebaker and King (1999) examined the reliability, factor structure, and validity of LIWC for reflecting personality. Across all LIWC category measures and across three participant samples, the authors found an average coefficient alpha of .59 with language composition categories (e.g., articles, prepositions) being more internally consistent than content categories (e.g., psychological process words, occupation-related words, leisure-related words), which can vary highly by the writing prompt given. Other researchers have provided evidence for the ability of LIWC to reflect emotional expression (Alpers et al., 2005), personal values (Boyd, Wilson, Pennebaker, Kosinkski, Stillwell, & Mihalecea, 2015), deception (Newman et al., 2003), and psychological health (Rude, Gortner, & Pennebaker, 2004), among other psychological phenomena (Tausczik & Pennebaker, 2010). The theoretical framework of LIWC has also been supported across multiple languages, including Chinese (Zhao, Jiao, Bai, & Zhu, 2016), Dutch (Boot, Zijlstra, & Geenen, 2017), and Spanish (Ramirez-Esparza, Pennebaker, Garcia, & Suria, 2007), among others (Pennebaker et al., 2015), further adding to its validity evidence.

Given this research, word count approaches to NLP such as LIWC appear to be useful methods for investigating psychological constructs. Word count approaches generally involve computing the frequency of words in a text sample that match a predefined list of words, often called a dictionary, which represents a psychological process or construct. For example, if the words “happy,” “joyful,” and “excited” are in a dictionary of positive emotion words, then a word count approach would tally the number of times each of those words appears in a text sample to compute a positive emotion score for that text sample. Word counting is a form of content analysis, a broad range of techniques used to organize and make sense of words, phrases, and language (Short et al., 2018), which until recently was conducted by manually coding themes throughout a passage of text (Duriau, Reger, & Pfarrer, 2007). Computer-aided text analysis approaches such as LIWC drastically improve the reliability, speed, and cost effectiveness of content analysis (Rosenberg, 1990, Dowling & Kabanoff, 1996; Neuendorf, 2002; Duriau et al., 2007; Short, Broberg, Cogliser, & Brigham, 2010). The LIWC framework provides a deductive rather than inductive approach to content analysis (Short et al., 2018), shifting the methodology from relying on subject matter expertise and subjective judgment to a more quantitative approach with less room for human biases (Short et al., 2018). Further, the ways people use words, provide markers of individuals' mental, social, and physical states. Word count approaches assume that word choice conveys additional psychological information (e.g., individual differences) beyond words' literal meaning or semantic context, the latter of which usually draw the attention of judges tasked with reading and analyzing the content of a body of text (Pennebaker, Mehl, & Niederhoffer, 2003). Automated approaches such as LIWC utilize a "bag-of-words" approach (Cambria & White, 2014), analyzing the words independent of context, which provides the basis for the exploration of word choice as a marker of individual differences such as GCA.

The Relationship between General Cognitive Ability and Language

GCA is the overall capacity of individuals which enables the correct or appropriate processing of mental information for successful performance on a given task, across contexts (Carroll, 1993). GCA is often used synonymously with the word "intelligence", described as "the general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly, and learn from experience," (Gottfredson, 1997a, p. 13). Spearman (1970) called this overall capability the "general factor" of intelligence, or *g*, describing it as a mathematical artifact of the positive correlations among all cognitive tests within a battery. Although Spearman (1970) did not consider *g* to be a "concrete thing but only a value or magnitude" (p. 75), *g* represents the similarities among test scores intended to represent various cognitive abilities, suggesting it to be a causal factor of these test scores. The existence of a statistical *g* is uncontroversial, though the explanation for this positive manifold of cognitive test scores remains debated to this day (Schneider & McGrew, 2012). To identify and estimate the GCA of individuals, psychologists typically administer a variety of cognitive tests and tasks which are positively intercorrelated (e.g., Ree, Earles, & Teachout, 1994). Stanek and Ones (2018) suggest that if any battery assesses three or more cognitive abilities, the resulting score is likely an index of GCA (p. 370). Likewise, if any battery assesses both fluid reasoning (i.e., fluid intelligence) and comprehension-knowledge (i.e., crystallized intelligence), the resulting score is also a likely index of GCA (Stanek & Ones, 2018), in accordance with Cattell's (1943) conceptualization that intelligence measurements are impacted by a combination of fluid and crystallized intelligence.

Much research has linked GCA with language development, understanding, and production. Carroll (1993) found that language abilities developed over time in conjunction with other cognitive abilities, concluding that any attempt at measuring language development within a person is confounded with measuring that person's cognitive development and abilities. Any written test attempting to assess GCA will by definition incorporate language to provide instructions and/or question prompts. Although some cognitive tests have been developed that do not utilize language (e.g., Raven's Progressive Matrices; Raven, 2000), these tests may only be measuring a single broad cognitive ability such as fluid reasoning (Stanek & Ones, 2018). According to Jensen (1980), vocabulary knowledge is one of the best indicators, if not the single best indicator, of GCA. All language is cognitively loaded to some extent, as it requires mental processing of information to comprehend and produce language (Carroll, 1993).

Recent research has linked GCA to specific LIWC word categories. Weaver (2017) examined seven LIWC categories in relation to cognitive ability: words with seven or more characters, conjunctions, prepositions, cognitive process words, and three of the subcategories of cognitive process words (i.e., differentiation words, causal words, and insight words). Using the spot-the-word test, a measure of verbal intelligence and GCA (Baddeley, Emslie, & Nimmo-Smith, 1993; Yuseph & Vanderploeg, 2000), Weaver found significant small to moderate correlations between GCA and each of the categories (*r* = .09-.29) except for insight words (*r* = .06). Though these effects are intriguing, they are far from any attempt at replacing traditional measures of GCA with a text-based sample as Weaver proposes. Further, Weaver used a combined domestic and international sample with varying levels of expertise in English, analyzing resume texts. LIWC word category base rates indicate differences across genres (Pennebaker et al., 2015), suggesting that resume text may also differ from other workplace text samples. The present study seeks to examine the predictive validity of GCA in relation to LIWC word categories in an essay context using an entirely domestic sample. Through the present study, further evidence may be gathered regarding the relationship between GCA and language through word frequency. The full theoretical hypothesized model is presented in Figure 1.

Hypothesis 1: Latent GCA will provide incremental prediction of (a) seven or more character word use, (b) unique word use, (c) conjunction word use, (d) preposition word use, and (e) cognitive process word use beyond broad abilities.

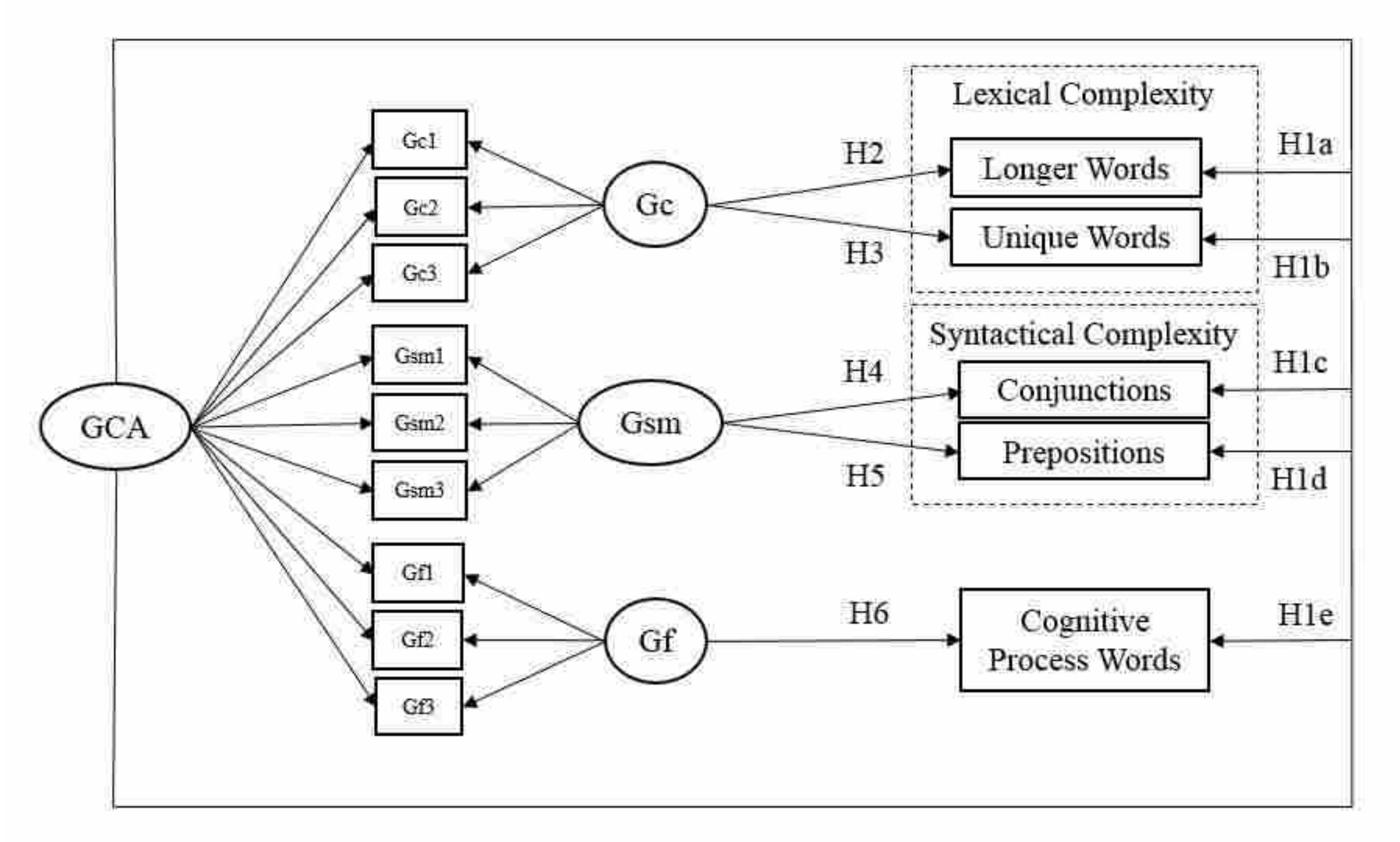


Figure 1. Theoretical hypothesized model.
Note: GCA = general cognitive ability. Gc = verbal ability. Gsm = short-term memory. Gf = fluid reasoning.

The Relationship between Broad Cognitive Abilities and Language

Broad cognitive abilities are mental capacities more specific than GCA, enabling the correct or appropriate processing of mental information for successful performance in a specific domain (e.g., language, memory, reasoning; Carroll, 1993, Schmidt & Hunter, 2004). In I-O psychology, broad and narrow cognitive abilities are often called "specific abilities" (Ree et al., 1994) or "*s*" for short (Spearman, 1970). Broad abilities are in specific domains such as language, mathematics, memory, and reasoning, whereas narrow abilities are specific aspects of each domain (e.g., reading comprehension, vocabulary knowledge, and grammatical knowledge in the language domain; induction and deduction in the reasoning domain). Although few would disagree about the existence of statistical *g*, debates remain over the exact number of broad abilities that exist, what they should be called, and what narrow abilities are subsumed within each broad ability (Schneider & McGrew, 2012). Recent research has proposed a total of 16 broad cognitive abilities (Schneider & McGrew, 2012). These include fluid reasoning, short-term memory, long-term storage and retrieval, processing speed, comprehension-knowledge, quantitative knowledge, reading and writing ability, visual processing, and auditory processing, among others.

In the intelligence literature, there are two major frameworks for modeling the relationships among GCA, broad abilities, and narrow abilities (Murray & Johnson, 2013). First, there is the Cattell-Horn-Carroll (CHC) framework of cognitive ability (McGrew, 1997; McGrew, 2005; Schneider & McGrew, 2012), or the hierarchical model, where GCA is a higher-order latent construct causing each broad ability. Each broad ability fully mediates the effects of GCA on each observed cognitive test score, also called narrow abilities (Carroll, 1993; Schneider & McGrew, 2012; Murray & Johnson, 2013; Stanek & Ones, 2018). The second framework for modeling cognitive abilities is the bi-factor model of cognitive abilities, where broad abilities are assumed to be independent from GCA, representing specific domains beyond general intelligence. Using this approach, GCA is not modeled as a higher-order latent construct causing each broad ability. Instead, each cognitive test score loads onto both GCA and a single broad cognitive ability directly.

The present study utilized a bi-factor model of cognitive abilities because it currently provides the best test of incremental prediction of outcomes by broad abilities over general abilities (Chen, West, & Sousa, 2006; Chen, Hayes, Carver, Laurenceau, & Zhang, 2012). It is currently difficult to discern which structural model of cognitive abilities is “correct,” but both models have distinct advantages over the other. For example, the bi-factor model is less constrained and tends to display better model fit in statistical tests (Murray & Johnson, 2013). A bifactor model of a multi-faceted construct like intelligence is more useful for examining the incremental variance of broad abilities compared to the general ability (Chen et al., 2006; Chen et al., 2012) than a traditional hierarchical model. This is in part because attempting to examine the incremental variance of broad abilities over general abilities within a hierarchical approach requires nonstandard structural equation modeling that is both difficult to execute and to interpret (Chen et al. 2006; Chen et al., 2012). In such an approach, instead of using latent broad abilities to predict outcomes, the residual variance of each broad ability is modeled separately from the latent constructs used to predict outcome variables (Gustafsson & Balke, 1993). A bifactor model is more easily interpretable and allows for the examination of broad abilities as distinct causal individual differences that can be contrasted directly with GCA.

Broad cognitive abilities can provide incremental explanation over the relationship between GCA and language. Specific aptitude theory suggests that broad cognitive abilities provide incremental prediction over GCA in personnel selection contexts in certain jobs (Schmidt & Hunter, 2004). Although evidence mounts against the notion that broad abilities predict job or training performance better than GCA (Ree et al., 1994; Schmidt & Hunter, 2004; Brown, Le, & Schmidt, 2006), it is possible that broad abilities can provide incremental prediction of other outcomes beyond GCA. The reason that specific aptitude theory is often rejected by researchers is that the jobs, training, and other outcomes under observation are highly *g*-loaded (Mount, Oh, & Burns, 2008). Language, on the other hand, may be *g*-loaded, but a variety of other influences also impact language, such as context, writing style, and individual differences (Tausczik & Pennebaker, 2010). Broad abilities are still cognitive abilities, but represent unique variance in cognitive ability tests unexplained by *g* which may prove relevant in the prediction of language outcomes like LIWC word categories (Spearman, 1970).

The present study explicitly tested specific aptitude theory in the context of written communication, an aspect of job performance (Campbell, McCloy, Oppler, & Sager, 1993). Specific aptitude theory would be supported by finding that broad cognitive abilities predicted writing behavior incrementally over GCA. Specific aptitude theory has been tested in the context of task proficiency (e.g., Ree et al., 1994) and training performance (e.g., Brown et al., 2006), but it has not been tested in the context of communication. Although research suggests GCA generally predicts workplace outcomes better than broad abilities, certain broad abilities may be more predictive in specific cases (e.g., job knowledge, perceptual speed; Brown et al., 2006; Mount et al., 2008). If the predictive power of broad abilities and GCA varies depending on context, then it is necessary to test the theory within a broad range of contexts, including the context of cognitively demanding writing, such as in the workplace.

Of all broad cognitive abilities, verbal ability (Gc), short-term memory (Gsm), and fluid reasoning (Gf) emerge as the most promising broad abilities for predicting NLP outcomes. Verbal ability encompasses both comprehension-knowledge and reading and writing ability (Schneider & McGrew, 2012), effectively tapping into acquired knowledge and skills in reading, writing, listening, and speaking (Carroll, 1993). Verbal ability is poised as the most likely broad ability to predict language-related outcomes even beyond the LIWC framework (Floyd, McGrew, & Evans, 2008; Cormier, Bulut, McGrew, & Frison, 2016). Short-term memory, "the ability to encode, maintain, and manipulate information in one's immediate awareness" (Schneider & McGrew, 2012, p. 114), is relevant to speech production (Kemper & Sumner, 2001) as well as in reading (Graesser et al., 2011). Forming or reading more complex sentences demands more cognitive resources from a person's short-term memory. Fluid reasoning, is involved in language at a higher-level: Gf involves following and applying rules to solve ambiguous problems such as writing. This form of reasoning and planning aids in structuring and developing a piece of writing, an important component of writing skill (Flower & Hayes, 1981; Attali & Burnstein, 2006; Deane, 2013). All three of these broad abilities were found to be strong predictors of writing skill across childhood development and several broad abilities (Floyd et al., 2008; Cormier et al., 2016).

The relationship between verbal ability (Gc) and language. Verbal ability (Gc) is a broad cognitive ability representing all acquired knowledge in the domain of language. It is thought by some researchers to be superordinate to other broad abilities such as comprehension-knowledge and reading and writing ability (Schneider & McGrew, 2012, Stanek & Ones 2018), but there is scholarly disagreement on the exact factor structure of cognitive abilities. Cattell (1963) considered acquired knowledge like verbal abilities to be a part of crystallized intelligence, and in factor analyses, Carroll (1993) found verbal abilities to align with a crystallized intelligence factor. Some researchers have proposed that abilities relating to language should be distinguished along the lines of speech and listening versus reading and writing (McGrew, 1997; McGrew, 2005; Schneider & McGrew, 2012). Others maintain that though this distinction is justified, grouping these two related abilities together under the heading of “verbal ability” may be necessary for balancing the level of specificity and abstraction present in current theories of cognitive abilities (Stanek & Ones, 2018). Comprehension-knowledge represents “the depth and breadth of knowledge and skills that are valued by one’s culture,” (Schneider & McGrew, 2012, p. 122), covering the listening and speaking end of the verbal ability spectrum, as speech is generally valued by one’s culture. In complement, reading and writing ability represents the “depth and breadth of knowledge and skills related to written language” (Schneider & McGrew, 2012, p. 125).

The relationship between Gc and word choice is best explained through the development of lexical knowledge (i.e., vocabulary, Schneider & McGrew, 2012). Verbal ability affects the development and use of speech and writing. It impacts knowledge of grammar, spelling, reading comprehension, listening comprehension, vocabulary, and overall language development (Schneider & McGrew, 2012). Each of these narrow abilities may affect word choice to some extent, but vocabulary has the largest impact. For example, reading or listening comprehension impacts word choice in that one must understand what is written or said by someone else before articulating a response. Although these narrow abilities impact the language one will choose to produce, it is in a less predictable fashion and largely dependent on situational context. Thus, vocabulary knowledge emerges as the best predictor of word choice. For example, a person cannot say or write a word that he or she does has not heard of or does not know. As a person’s lexical knowledge increases, the greater the variety of words he or she can use in speech or writing. Thus, vocabulary size should be predictive of the variance in word choice such that people with smaller vocabularies are only able to use a limited variety of words while people with larger vocabularies are able to produce a larger variety of words in their speech and writing. As vocabularies increase in size, the average length of the words in one's vocabulary also increase. For example, in childhood language development, a child begins with single syllable sounds before proceeding on to learning simple words then progressively more complex words as they age and learn (McCarthy, 1933). In this regard, it might be predicted that people with larger vocabularies are more capable of using longer words than people with smaller vocabularies.

Higher levels of Gc should be more positively related to use of larger words (i.e., words consisting of seven or more characters). A person must know a word before he or she can use that word. As vocabularies grow, the length of the average known word should increase (McCarthy, 1933). Thus, a person with a smaller vocabulary knowledge base should produce slightly shorter words on average compared to a person with a larger vocabulary knowledge base. However, this potential connection between word length and vocabulary knowledge is likely a weak connection. Knowing more words or longer words does not guarantee that a person will use longer words. However, the use of any word does depend on knowledge of that word first, which provides a sort of lower bound to the average word length for a person in each sample of text or speech. Little research has examined the psychological correlates of word length, but Pennebaker and King (1999) did examine this LIWC category when investigating the factor structure of LIWC. The authors found that longer words (i.e., words consisting of seven or more characters) loaded negatively on a factor they labeled "immediacy." Other categories loading on the “immediacy” factor included use of first-person singular pronouns (e.g., I, me, my), discrepancy words (e.g., need, could, lack), present tense verbs, and a negative loading of articles (e.g., a, an, the) on the factor. The immediacy factor was negatively correlated with SAT scores and school exam grades, as well as a need for cognition, openness to experience, and parent education, all of which are positively correlated with cognitive abilities (Sewell & Shah, 1968; Ackerman & Heggestad, 1997; Frey & Detterman, 2004; Gignac, Stough, & Loukomitis, 2004; Furnham & Monsen, 2009; Hill, Foster, Elliott, Shelton, McCain, & Gouvier, 2013). This evidence suggests that lower scores on the immediacy factor are positively related to many of correlates of cognitive abilities, including the increased use of longer words in speech or writing samples. It is likely that cognitive abilities, such as verbal ability, are positively related to the use of longer words.

Hypothesis 2: Latent Gc will provide incremental prediction of seven or more character word use over GCA.

Higher levels of Gc should be more positively related to unique word use in each writing sample. In analyzing the factor structure of verbal abilities in adults, Kemper and Sumner (2001) found that measures of vocabulary knowledge correlated moderately with type-token ratio (r = .21-.44), an index of unique word use. As scores on a variety of vocabulary tests increased, the ratio of unique words used to total words increased, suggesting that verbal ability and unique word use are related. If a person has more lexical knowledge, there are more words that he or she can potentially use in speech or writing. Again, vocabulary knowledge does not cause a person to use more unique words, but it does provide a minimum capability for producing a high ratio of unique words compared to all words used. The automatic essay scoring literature supports the notion that verbal ability is connected to unique word use. Algorithms assessing writing skill factor the "sophistication" of vocabulary into essay scores by assessing typical word length and word uniqueness (Attali & Burnstein, 2006; Deane, 2013). On a more fundamental level, Jensen (1980) explains that word uniqueness plays a key role in the creation of vocabulary tests. Discriminating vocabulary assessments should include words across a range of difficulty. Difficult words are those words that are less frequently seen or used, while easier words are more commonly known. Word frequency is inherently tied to word uniqueness; as words begin to repeat within a text sample, the ratio of unique words to total words decreases. Thus, it is hypothesized that verbal ability impacts unique word use such that those with larger vocabularies will have larger proportions of unique words to total words used when compared to those with smaller vocabularies.

Hypothesis 3: Latent Gc will provide incremental prediction of unique word use over GCA.

The relationship between short-term memory (Gsm) and language. Short-term memory ability (Gsm) is a domain-free capacity not associated with a specific sensory system (Schneider & McGrew, 2012), which is under the broader category of all memory (Stanek & Ones, 2018). Gsm refers to "individual differences in both the capacity (size) of primary memory and to the efficiency of attentional control mechanisms that manipulate information within primary memory," (Schneider & McGrew, 2012, p. 114-115). Gsm is typically measured with tests of memory span (i.e., reproducing a sequence of visual or audio information in the same order that it was presented) and working memory capacity (i.e., performing simple operations, manipulations, transformations, or combinations of information in primary memory; Schneider & McGrew, 2012; Stanek & Ones, 2018). The working memory aspect of Gsm is most relevant to language production. As a person speaks or writes, he or she is encoding and maintaining information into primary memory. As the statement a person is trying to make becomes more complex, the information and language is manipulated, increasingly taxing the working memory. The person must give increased attention to the meaning of the statement itself as well as the words being used to convey that meaning, ignoring distractions and irrelevant information.

The relationship between Gsm and word choice is best explained through sentence structure (i.e., syntactical complexity). Sentences with more complex syntactical structure tax Gsm more than simpler sentences, as they include more words, descriptions, ideas, and phrases than a simpler sentence (Graesser et al., 2011). In their factor analysis of verbal abilities, Kemper and Sumner (2001) identified an indicator of syntactical complexity (i.e., the "development level" of participant speech) that loaded strongly and positively onto a working memory factor. Development level is an index of syntactical complexity ranging from "simple one-clause sentences to complex sentences with multiple forms of embedding and subordination" (Rosenberg & Abbeduto, 1987; Kemper & Sumner, 2001, p. 315). People with better short-term memory abilities tended to produce sentences with more embeddings and subordinate clauses. These embeddings and combinations of clauses are generally marked by specific grammatical syntax such as conjunctions and prepositions (Rosenberg & Abbeduto, 1987), both of which are word categories found within the LIWC framework (Pennebaker et al., 2015). Kemper and Sumner (2001) also found moderately strong correlations of short-term memory measures with the mean length of speech utterances (i.e., mean sentence length). These data suggest that people with stronger short-term memory abilities can produce longer sentences in speech and writing, as they are able to hold more ideas, descriptions, and phrases within their primary memory before and during language production. Increases in sentence length are also generally marked by an increased usage of conjunctions and prepositions as these parts of speech combine simple phrases with other phrases to provide additional information and meaning.

Higher levels of Gsm should also be associated with increased use of both conjunctions and prepositions. Conjunctions and prepositions indicate greater syntactical complexity, which is tied to short-term memory ability (Kemper & Sumner, 2001). Conjunctions (e.g., and, but, or) are used to combine multiple statements and phrases together, increasing both sentence length and syntactical complexity. Many conjunctions are also logical operators (e.g., or, and, if-then), which in larger numbers in a language sample can create a larger need for cognitive processing, taxing the working memory (Graesser et al., 2004). Prepositions (e.g., of, under, to) indicate that a speaker or writer is providing more complex information about a topic, adding additional description beyond the clauses of the root sentence (Tausczik & Pennebaker, 2010). These descriptions require more attentional resources from the short-term memory when forming sentences. Thus, people with high levels of Gsm are hypothesized to use more complex sentences than people with lower levels of Gsm, which are accordingly indicated by a greater use of conjunctions and prepositions than people with lower levels of Gsm.

Hypothesis 4: Latent Gsm will provide incremental prediction of conjunction word use over GCA.

Hypothesis 5: Latent Gsm will provide incremental prediction of preposition word use over GCA.

The relationship between fluid reasoning (Gf) and language. Fluid reasoning (Gf), traditionally labeled fluid intelligence (Cattell, 1943; 1963), is the “deliberate but flexible control of attention to solve novel, ‘on-the-spot’ problems that cannot be performed by relying exclusively on previously learned habits, schemas, or scripts,” (Schneider & McGrew, 2012, p.

111). Broadly, Gf includes inductive and general sequential (i.e., deductive) reasoning abilities, which are also the primary means for measuring Gf (Schneider & McGrew, 2012; Stanek & Ones, 2018). Induction involves discovering underlying rules or patterns whereas deduction involves applying known rules or premises to reason logically (Stanek & Ones, 2018). Within language production, Gf plays a role in the organization and structure of written or spoken discourse. Drafting a speech or manuscript is an inherently cognitive task requiring analysis of the prompt, context, and audience while synthesizing information and ordering it in a logical fashion. In writing, the author must both comprehend the task as well as compose a written response, which are influenced by the author's ability to interpret the task (Deane, 2013). In the automatic essay scoring literature, the organization of writing is an important factor in scoring writing skill (Attali & Burnstein, 2006; Deane, 2013). Scoring algorithms are trained to identify thesis statements, supporting points, and conclusions, a logical order for making a specific point in written discourse (Attali & Burnstein, 2006). The ability of a writer to supply appropriate information in a logical order is based on his or her fluid reasoning ability. This relationship between Gf and language is through problem solving, as all written or spoken tasks, formal or informal, rely on problem solving ability to some extent in generating a response.

The relationship between Gf and word choice is best explained through problem solving, a cognitive process. Cognitive processes refer to any actions used to operate upon mental contents to produce some result or response (Carroll, 1993). Cognitive processes are the mental actions taken by an individual to solve a cognitively-oriented task, such as solving a math problem, mentally rotating a figure, or interpreting meaning from written text. Problem solving is a specific cognitive process enabled by fluid reasoning (Stanek & Ones, 2018). Thus, problem solving can take many different forms across a variety of domains, but the consistent thread is mentally processing, analyzing, and evaluating information. This processing, analysis, and evaluation is often indicated in speech or writing. It is likely that words that indicate cognitive processing and problem solving behavior also indicate some degree of Gf.

Higher levels of Gf should be more positively related to the use of cognitive process words. The LIWC framework includes a higher-order category of “cognitive process” words that imply the writer or speaker is actively thinking, making comparisons, evaluations, and analyses (Tausczik & Pennebaker, 2010). The cognitive processes word category can be broken down into insight words, causation words, discrepancy words, tentative words, certainty words, and differentiation words (Pennebaker et al., 2015). Insight words (e.g., think, know, explain) indicate that a person has made or is in the process of making some sort of realization. Making a realization implies that the person was previously thinking about, for example, a problem to be solved. Causation words (e.g., because, effect, change) indicate that a person is analyzing the relationship between two or more entities or how something may have been changed over time. Discrepancy words (e.g., should, would, lack) indicate that a person is making some sort of evaluation. These words mark a discrepancy between a current state and an ideal or predicted state, or a contrast between two or more entities. Tentative words (e.g., maybe, perhaps, depend) indicate that a person’s evaluation of something is subjective or not yet finalized, while certainty words (e.g., always, never, absolute) indicate a person’s evaluation of something is conclusive and final. Differentiation words (e.g., exclude, but, else) indicate that a person is contrasting two or more entities, analyzing the similarities and differences between them. These word categories all appear to indicate cognitive processing and problem solving behaviors. Thus, people who use more cognitive processing words likely engage in more problem solving, which may be indicative of higher levels of Gf.

Hypothesis 6: Latent Gf will provide incremental prediction of cognitive process word use over GCA.

Model Specification and Previously Unexplored Relationships

Although previous research suggests that the hypothesized paths between broad cognitive abilities and word count categories exist, there is a dearth of research regarding many of the other paths implied by the theoretical model. This is problematic because SEM requires a strong a priori statement of each potential path's existence; yet if the research literature does not support the inclusion or exclusion of a relationship between two constructs in the theoretical model, it is impossible to hypothesize one way or another about such a path's existence. Because constraining relationships to zero when they are in fact non-zero decreases the fit between the data and hypothesized model, paths not previously explored in the literature will be freed in order to explore the strength of these relationships and establish preliminary estimates of their magnitude. Regarding verbal ability, the relationships between Gc and conjunctions, prepositions, and cognitive process words will be explored. Regarding short-term memory, the relationships between Gsm and longer words, unique word use, and cognitive process words will be explored. Regarding fluid reasoning, the relationship between Gf and longer words, unique word use, conjunctions, and prepositions will be explored.

Research Question 1: Of what strength are the relationships between Gc and other word count outcomes, Gsm and other word count outcomes, and Gf and other word count outcomes while controlling for GCA and the other broad cognitive abilities?

METHOD

Participants

To determine the number of participants needed for this study, a Monte Carlo simulation was conducted using Mplus. A Monte Carlo study can be used to decide on sample size and determine the statistical power when conducting structural equation modeling (Muthén & Muthén, 2002). In Monte Carlo studies, population data are simulated and a large number of samples are drawn from these simulated data. A model is estimated for each sample and parameter values and standard errors are averaged across all of the samples (Muthén & Muthén, 2002). A theorized structural equation model was specified for this study with population estimates derived from the literature where available (i.e., primarily from Landers, Armstrong, & Collmus, 2017; Weaver, 2017; and Pennebaker & King, 1999). When estimates were not available in the literature (i.e., for the relationships between LIWC variables and broad cognitive abilities), medium effect sizes (r = .30; Cohen, 1992) were used to be as realistic as possible (i.e., given the medium effect size correlations of Weaver), yet stringent enough to avoid underpowered effects. In a 10,000-replication Monte Carlo simulation, it was found that a sample size of 350 would be required for model stability and for all anticipated path estimates to reach 80% power. For path estimates to reach 90% power, a sample size of 475 would be required. However, with a sample size of 400, all path estimates would have 90% power except for one, which would have 80% power. Thus, a sample size of 400 participants was set as a target recruitment goal in order to balance statistical requirements and practical considerations.

Participants were recruited for this study using Amazon Mechanical Turk (MTurk). MTurk sampling provides a wide variety of individuals across a variety of educational and professional backgrounds, which aids in the generalization of these results across all working adults (Landers & Behrend, 2015). Participants were compensated at a rate of about US$4.50-5.15 per hour (i.e., $9.00 total for 105 to 120 minutes of participation), which was based on previous research examining the expected wage of MTurk workers (Horton & Chilton, 2010; Armstrong & Landers, 2017). Broad criteria were used to increase the variance in MTurk work experience: a 95% or higher task acceptance rate, completion of at least 50 previous MTurk tasks, and a location in the United States.

First, the study was posted to MTurk's website with 20 participation slots as a test of the technology delivering the survey and the payment structure. Participants were paid a total of $7.00, which generated complaints from several participants. The payment structure was revised to total $9.00 for the remaining participants, and the previous 20 were given a $2.00 bonus within a day of completing their surveys to maintain fairness. MTurk allows Requesters (i.e., the researcher) to approve or reject MTurk workers' task submissions. All 20 of the first batch of participants' work were approved. Next, 383 participants were recruited with the revised payment structure. Of these 383 participants, 44 participants' work was rejected by the researcher. In determining whether to approve or reject survey responses, bogus item responses were examined, as well as essay responses. Several rules were established for determining if work should be approved or rejected, excluding these participants from the sample. These rules are presented in Table 1. Another 41 participants were recruited, with 1 participant's work being rejected. In total, 444 participants were recruited and paid. However, due to the nature of MTurk, three additional participants were recruited and completed part of the survey, dropping out before completion. One of these participants declined the consent form, immediately terminating the study. Another began the survey, but never finished it. It is unclear how the third additional participant ended up in the data set, as all other cases completed the survey protocol, entering their Worker ID on the last page.

A total of 445 participants completed the study protocol. Of those 445 cases, the surveys submitted by 45 participants were rejected for the reasons outlined in Table 1. These participants were not paid. The data were examined to determine if additional cases needed to be excluded before analysis. First, geographic location was examined via latitude and longitude coordinates. Only MTurk workers registered with American accounts were allowed to participate in the study, but some MTurk workers from other countries may try to register an American account to access American work tasks and surveys. Participants were retained if their latitude and longitude coordinate data when completing the survey were roughly within the contiguous United States (i.e., latitude between 15 and 50 degrees North and longitude between -60 and -130 degrees West). Of the 400 participants with approved submissions, 16 participants did not meet these criteria and their data were discarded before conducting analyses because of the higher likelihood of being non-American. Additionally, some participants experienced technical errors, prompting their data to be discarded before analyses. On the second test of Gsm, some participants encountered errors playing the audio files due to their web browsers not supporting Flash-based videos. These participants noted the error in the response space provided (e.g., “Video did not play.”) in addition to most emailing the researcher. Of the 384 remaining cases, 8 participants experienced technical errors and were excluded from analyses.

Table 1
Rules for Approving/Rejecting MTurk Survey Submissions

Rule	Reject If…
1	The participant failed 1 or 2 (of 4 total) bogus items AND did not follow all essay instructions.
2	The participant failed 1 or 2 bogus items AND only spent approximately 1-2 seconds on pages with audio/video stimuli lasting 5+ seconds.
3	The participant failed 1 or 2 bogus items AND typed numerical responses on a task asking for alphabetical responses.
4	At least 1 essay (of 3 total) was left blank.
5	The participant copy-pasted the essay writing prompt into the response box and added no original essay content.
6	Each of the 3 essays contained less than 5 sentences each.
7	At least 2 of the participant's essays were searchable online OR are word-for-word identical to another participant's essay.
8	There is some combination of essays with less than 5 sentences each AND plagiarism.

Next, the data were examined for possible cases to exclude. A variety of exclusionary criteria were investigated, creating a new variable for each criterion. These criteria are outlined and summarized, including frequency counts and pass rates, in Table 2. Criteria 1 through 4 were correctly answering bogus items (Meade & Craig, 2012) mixed into the cognitive test battery (see Appendix A). Criteria 5 and 6 asked participants if they gave an honest effort at the task if their data should be used for research purposes. Although most participants passed criteria 5 and 6, a small number did concede that they did not try their hardest or that their data should not be included in analyses. Criterion 7 was that participants' had to identify English as their native language. Criteria 8 through 10 were based on following instructions and paying attention. For the tests of short-term memory ability, participants were instructed to listen to and view a series of audio and video files. There were 24 files to play for each of the three tests. Timestamp data from the last stimulus and item of each test were examined to determine if participants played the entire file before moving on to the next page. Without spending enough time on each stimulus, it would be impossible to correctly remember the number and letter sequences presented without cheating. Finally, some participants did not follow instructions when formatting their responses to these tasks. When merging data files in SPSS, the format for the entire variable was converted to whatever 95% of the variables already are. For these tests, 95% of participant data were numerical responses (e.g., 1234), but some formatted their responses as strings (e.g., "1, 2, 3, 4"). These strings were converted into missing data by SPSS and thus counted as incorrect when recoding variables, resulting in scale scores of zero for Gsm1 (criterion 11) and Gsm2 (criterion 12). This issue with string data did not apply to the third test of Gsm, which involved letters instead of numbers.

Table 2
Exclusionary Criteria Pass Rates

Criterion	Description	Passing	Failing	Percent Passing
1	Bogus item 1. Participants had to select the response most like the word "happy" with response options: "sad," "angry," "afraid," "joyful," and "disgusted." If participants selected "joyful," they passed the item.	372	4	98.94%
2	Bogus item 2. Participants had to select the response most like the word "mother" with response options: "aunt," "uncle," "mom," and "dad." If participants selected "mom" for this item, they passed the item.	371	5	98.67%
3	Bogus item 3. Participants were presented with the visual stimulus "3, 4, 5" to remember and recall. If participants correctly recalled the number sequence, they passed the item.	353	23	93.88%
4	Bogus item 4. Participants had to select 1 of 5 letter sequences that did not match the others: "NNNN," "NNNN," "NNNN," "MMMM," and "NNNN." Participants who selected "MMMM," passed the item.	361	15	96.01%
5	A single question at the end of the study protocol: "I gave an honest effort at all of these assessments. True or False?" Participants answering with "true" passed the item.	366	10	97.34%
6	A single question at the end of the study protocol: "In all honesty, you should not use my data for research purposes because I did not respond completely honestly or to the best of my ability. Yes or No?" Participants responding "yes," passed the item.	370	6	98.40%
7	If participants identified English as their native language, they passed the item.	374	2	99.47%
8	Participants failed if they spent less than 10 seconds on a page requiring they listen to a 12-second audio file.	366	10	97.34%
9	Participants failed if they spent less than 17 seconds viewing a 19-second video file.	352	24	93.62%
10	Participants failed if they spent less than 10 seconds on another 12-second audio file.	356	20	94.68%
11	Participants failed if they used improper formatting on Gsm1, resulting in a Gsm1 score of zero.	369	7	98.14%
12	Participants failed if they used improper formatting on Gsm2, resulting in a Gsm2 score of zero.	367	9	97.60%

N = 376.

Of the 376 remaining cases, 295 participants correctly passed 12 of 12 criteria. A total of 344 participants passed at least 11 of 12 criteria, 362 participants passed at least 10 of 12 criteria, and 373 participants passed at least 9 of 12 criteria. The three participants passing the fewest criteria (i.e., 5 of 12, 6 of 12, and 8 of 12, respectively), were excluded from analyses. Thus, all participants passing at least 9 of 12 criteria were retained, resulting in a final sample size of 373. Although more liberally excluding cases may have better preserved data integrity, a larger sample was needed in accordance with the power analysis.

After data cleaning and exclusions, 373 participants were retained for analysis whose demographics are presented in Tables 3 and 4. In summary, participant ages ranged from 18 to 65+ years, averaging 36 years. Participant gender was evenly split between male and female and the sample was mostly non-Hispanic and Caucasian. Almost all participants spoke English as their native language. Most participants were employed either full-time or part-time beyond MTurk, although 83 participants were either unemployed or only worked on MTurk. Participants worked across a variety of industries, with average job tenure at their current job being just over 6 years. Participants reported working an average of about 38 hours per week.

Table 3
Descriptive Statistics for Participant Age, Tenure, and Hours Worked

	N	Mean	*SD*	Min	Max
Age (Years)	372	35.86	10.28	18.00	65.00
Tenure (Months)	290	72.48	65.21	1.00	497.00
Tenure (Years)	290	6.04	5.43	0.08	41.42
Hours/week	290	37.58	8.55	6.00	70.00

Table 4
Participant Demographic Responses Frequencies and Percentages

	Response Option	Total	Percent
Gender	Male	186	50.00%
N = 372	Female	185	49.73%
	Other (Transgender)	1	0.27%
Ethnicity	Hispanic	37	9.97%
N = 371	Non-Hispanic	334	90.03%
Race	African American or Black	39	10.48%
N = 372	Asian American	18	4.84%
	Caucasian or White	297	79.84%
	Native American or Native Alaskan	3	0.81%
	Other single race	4	1.08%
	Two or more races	10	2.69%
	Not American	1	0.27%
Native	English	371	99.46%
Language	Mandarin	1	0.27%
N = 373	Other (Norwegian)	1	0.27%
Employment	Full time	234	62.73%
N = 373	Part time	56	15.82%
	Unemployed	83	22.25%
Industry	Business Services	61	21.03%
N = 290	Education	27	9.31%
	Finance	19	6.55%
	Health Care	34	11.72%
	Insurance	8	2.76%
	Manufacturing	29	10.00%
	Retail	56	19.31%
	Wholesale	5	1.72%
	Other	51	17.59%
Master Worker	No	306	82.26%
N = 372	Yes	66	17.74%

Other single races: Latino (2), Mestiza (1), Puerto Rican(1). Two or more races: Black-White (1), Black-Native American (1), Asian American-White (4), White-Native American (4).

Measures

Writing samples. Three writing prompts from the Graduate Record Examination (GRE) Analytical Writing Measure were administered to participants in order to collect a writing sample. Specifically, participants answered three “Analyze an Issue” tasks with a 5-minute time limit on task. This task states an opinion on a general issue and asks test-takers to address the issue form any perspective, providing relevant reasons and examples to support their claims (Powers, Burstein, Chodorow, Fowles, & Kukich, 2000). The GRE is a cognitively-demanding high-stakes test often determining entrance into graduate programs of study. In this way, the GRE analytic writing task is similar to high-stakes employment testing, helping enable the generalization of this writing sample to workplace pre-employment testing contexts. The following writing prompts were used: 1) “As people rely more on technology to solve problems, the ability of humans to think for themselves will surely deteriorate.” 2) “To understand the most important characteristics of a society, one must study its major cities.” 3) “Scandals are useful because they focus our attention on problems in ways that no speaker or reformer ever could.” Participants were instructed to write responses in which they discussed the extent to which they agreed or disagreed with the claims provided. In developing and supporting their positions, participants were encouraged to address the most compelling reasons and/or examples that could be used to challenge their positions. The writing samples were not assessed for participants’ ability to articulate complex ideas or build arguments; instead, they were used as a method for obtaining cognitively-loaded writing. Participants were required to write a minimum of 5 sentences and spend no less than 1 minute writing before proceeding with the next essay and the remainder of the study. There are no guidelines for how many words are recommended per text sample to provide reliable and valid measures in LIWC, although the manual stated that in acquiring base rates for each category, a minimum of 25 words per text corpus were required for inclusion in analyses (Pennebaker et al., 2015). To improve the external validity of this task in relation to a high-stakes testing context where participants would be writing an essay in order to apply for a job or promotion, the top five best written essays each received a $50 bonus payment.

General cognitive ability and broad cognitive abilities. Verbal ability (Gc), short-term memory ability (Gsm), and fluid reasoning ability (Gf) were all assessed using tests from the Educational Testing Services' Kit of Factor-Referenced Cognitive Tests (Ekstrom, French, Harman, & Dermen, 1976), which were designed for research purposes (Carroll, 1993). Factor-analytic and correlational validity evidence has supported the use of this cognitive battery for measuring Gc, Gsm, and Gf (Bunderson, 1967; Lemke, Klausmeier, & Harris, 1967; Dunham & Bunderson, 1969; Traub, 1970). Carroll's (1993) review positioned these tests within the greater intelligence literature, each loading onto its intended broader cognitive ability. In the present study, each broad cognitive ability was measured with three tests, which is generally considered a lower-limit of indicator variables for model identification (Marsh, Hau, & Balla, 1998). Test descriptions, length, and time limits for each test in the cognitive battery are presented in Table 5. Correct responses were coded as "1" and incorrect or missing responses were coded as "0". Mean scores were calculated for each cognitive ability test, resulting in three scale scores per broad cognitive ability test. Scores were converted to percentages to match the scaling of the LIWC outcome variables (Muthén & Muthén, 2012). To do so, each scale score was multiplied by 100.00 (e.g., changing a score of .50 to 50.00%).

Table 5
Cognitive Ability Test Battery Details

Broad Cognitive Ability	Test	Description	Number of Items	Time Limit
Verbal Ability (Gc)	Extended Range Vocabulary Test	A 5-choice synonym test having items ranging from very easy to very difficult.	24	6 mins
	Advanced Vocabulary Test I	A 5-choice synonym test consisting mainly of difficult items.	18	4 mins
	Advanced Vocabulary Test II	A 4-choice synonym test consisting mainly of difficult items.	18	4 mins
Short-Term Memory Ability (Gsm)	Auditory Number Span	A conventional digit-span test with digits in series of varying length. Each digit is read aloud to the participant at a speed of one digit per second. Once the series is completed, participants recall the order of the digits and write them down.	24	N/A
	Visual Number Span Test	A conventional digit-span test with digits in series of varying length. Each digit is visually displayed for one second for the participant to see. Once the series is completed, participants recall the order of the digits and write them down.	24	N/A
	Auditory Letter Span	A test like the Auditory Number Span Test, but using letters instead of numerical digits.	24	N/A
Fluid Reasoning (Gf)	Letter Sets Test	Five sets of four letters are presented. The task is to find the rule which relates four of the sets to each other and identify the one which does not fit the rule.	15	7 mins
	Locations Test	For each item, five rows of dashes and gaps are given. In each of the first four rows one dash in each row is marked with an "X" according to a rule. The task is to discover the rule and to select one of 5 numbered places in the fifth row accordingly.	14	6 mins
	Figure Classification Test	Each item presents 2 or 3 groups each containing 3 geometrical figures that are alike in accordance with some rule. The second row of each item contains 8 test figures. The task is to discover the rules and assign each test figure to one of the groups.	14	8 mins

Verbal ability (Gc). Gc was measured using three tests from the verbal comprehension factor: the Extended Range Vocabulary Test (Gc1), Advanced Vocabulary Test I (Gc2), and Advanced Vocabulary Test II (Gc3). These tests represent lexical knowledge (i.e., vocabulary knowledge, Stanek & Ones, 2018), which is a prerequisite narrow ability for other verbal abilities such as reading comprehension (Schneider & McGrew, 2012). An example item is for participants to select a synonym for the word "orthodox" from a list of possible responses: 1) conventional, 2) straight, 3) surgical, 4) right-angled, or 5) religious. Internal consistency reliability estimates for were acceptable for basic research (i.e., α = .70; Nunnally, 1978) for all three tests (α = .783, .770, and .743 for Gc1, Gc2, and Gc3, respectively).

Short-term memory (Gsm). Gsm was measured using three tests from the memory span factor: the Auditory Number Span Test (Gsm1), Visual Number Span Test (Gsm2), and Auditory Letter Span Test (Gsm3). These tests represent the memory span factor of Gsm (Stanek & Ones, 2018), which Schneider and McGrew (2012) recommended as the most important factor to measure when assessing Gsm. An example item involved participants listening to a pre-recorded sequence of numbers such as "8, 1, 9, 5, 7, 2" then recalling the order of the numbers after the recording is finished. Two items from the Auditory Number Span Test and three items from the Visual Number Span Test were dropped from analysis due to having zero variance. These items were so difficult that no participant answered them correctly. Internal consistency reliability estimates for were acceptable for basic research for all three tests (α = .898, .893, and .875 for Gsm1, Gsm2, and Gsm3, respectively).

Fluid reasoning (Gf). Gf was assessed using three tests from the induction factor: The Letter Sets Test, Locations Test, and Figure Classification Test. These tests represent the induction factor of Gf (Stanek & Ones, 2018), which is considered the core underlying factor of

Gf (Schneider & McGrew, 2012). An example item presented participants with five sets of letters (e.g., QPPQ, HGHH, TTTU, DDDE, MLMM). Four of the letter sets were associated with one another through an underlying rule (e.g., a letter that repeats three times in the set). The participant had to identify which letter set did not fit with the others. Internal consistency reliability estimates were acceptable for basic research for Gf1 and Gf3 (α = .792 and .939, respectively. Gf2, the Locations Test, was less internally consistent (α = .624), unlike historical reliability estimates for this test (α = .75; Ekstrom et al., 1976).

Linguistic Inquiry and Word Count (LIWC). Text responses were downloaded in a CSV file in separate cells and accessed by LIWC. For each cell, LIWC read one target word at a time, searching its internal dictionary for a match with the target word (Pennebaker et al., 2015). For each match, that category was incremented. After each file was analyzed, LIWC produced a table of output variables, which was merged with the remaining dataset using identifier variables. Composite LIWC scores were calculated by averaging the proportions of each category across all three essays, resulting in one score per category across all participant writing samples. Each category under observation is described in the following sections.

In general, the psychometrics of natural language processing are less well understood than questionnaires. In natural language, when a person says something, they generally tend to not repeat the same information within the same paragraph or essay. It is generally considered good discourse to move on to the next topic. However, in self-report questionnaires, the same item content is usually repeated with slight variations several times in order to obtain a stable estimate to minimize systematic error influences. Thus, in natural language processing, internal consistency estimates of reliability tend to be much lower than traditional psychometric standards (Pennebaker et al., 2015). Using the Spearman-Brown prediction formula to correct coefficient alphas generally provides a more accurate approximation of the psychometric internal consistency for a LIWC word category than raw uncorrected alphas (Pennebaker et al., 2015). Both are presented for the conjunctions, prepositions, and cognitive process word categories below. Reliability estimates for word with seven or more characters or unique words are not given in the most recent LIWC manual (Pennebaker et al., 2015). Pennebaker and King (1999) reported test-retest reliability of .59 across all LIWC categories, which gives some indication of the reliability for these two categories. However, in structural equation modeling, low reliability is not an issue of concern due to the way that common factors are modeled.

Words with seven or more characters (i.e., long words). The word length metric of LIWC is calculated in a similar fashion to the other word categories in the program. The number of words with seven or more characters is divided by the total number of words in the text sample, yielding a proportion for long words used out of all words used. Although it is odd to convert a numerical quantity like character count into a categorical variable (i.e., long word vs. short word), this conversion is consistent with other metrics produced by LIWC. Further, measuring word categories as proportions provides meaningful results independent of total word count or writing sample size. In the first manual for LIWC, Pennebaker, Francis, and Booth (2001) stated that natural language generally has a lower percentage of long words compared to short words. This is evident in the current base rates of word frequency in LIWC where words with seven or more characters make up 15.6% of all language across a variety of text genres (Pennebaker et al., 2015). This was further evidenced by Miller, Newman, and Friedman (1958), who analyzed word length and word frequency in a large text sample. Miller and colleagues found that among all unique words in their text sample, the most frequent length of words was seven characters. However, regarding the most frequently used word lengths, Miller and colleagues found a large positively skewed distribution, where 2-letter, 3-letter, and 4-letter words are used most frequently, then a sharp decline in use of 5-letter words and exponentially less use of words longer that. This was due to the tendency in English to use function words (e.g., articles, prepositions, conjunctions) at a greater rate than content words, which are generally shorter in length (Miller et al., 1958). The proportion of long words in each essay was averaged together for each person to create a composite long word use score. Treating each essay score as one item in the composite, internal consistency reliability was moderate (α = .726).

Unique words. Unique words were originally tabulated by LIWC (Pennebaker et al., 2001), but the metric was removed from later revisions to the program. Unique words were removed because they tended to correlate highly negatively with total word count (r = -0.80; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007). However, this metric is commonly used in linguistic research under the moniker "type-token ratio" (e.g., Miller et al., 1958; Kemper & Sumner, 2001), and the research literature supports the link between Gc and unique words whereas there is no support for an inverse connection with total word count. Regardless, both variables were examined in relation to Gc and GCA to investigate Pennebaker and colleagues' (2007) proposition that unique word proportions are no different than total word count. Unique words score were calculated by counting the total number of words that appeared at least one time in a given text sample then dividing that number by the total number of words present in the same text sample, resulting in a proportion of unique words used to total words used. A larger proportion means that an individual's text sample contained many unique words rather than repeated the same words multiples times within the sample. The proportion of unique words in each essay was averaged together for each person to create a composite unique word use score.

Treating each essay score as one item in the composite, internal consistency reliability was moderate (α = .697).

Conjunctions. The conjunctions score were calculated by LIWC as the proportion of words used falling under LIWC's conjunction word category out of all words used in that text sample. The current list of LIWC conjunctions includes 43 words, including contractions, misspellings, and slang used as conjunctions. Examples of words in this category include "also," "and," "because," "but," "or," and "while." Internal consistency measures of conjunction word use are generally low for psychometric standards, but comparable to other LIWC word categories ($\alpha_{uncorrected}$ = .14, $\alpha_{corrected}$ = .50; Pennebaker et al., 2015). The proportion of conjunctions in each essay was averaged together for each person to create a composite conjunction use score. Treating each essay score as one item in the composite, internal consistency reliability was low (α = .384).

Prepositions. The prepositions score were calculated by LIWC as the proportion of words used falling under LIWC's preposition word category out of all words used in that text sample. The current list of LIWC prepositions includes 74 words, 3 of which are stems with multiple possible word endings (i.e., among, through, toward). Examples of words in this category include "about," "above," "behind," "during," "into," "of," and "within." Internal consistency measures of preposition word use are very low for psychometric standards and are generally low even compared to other LIWC word categories ($\alpha_{uncorrected}$ = .04, $\alpha_{corrected}$ = .18; Pennebaker et al., 2015). The proportion of prepositions in each essay was averaged together for each person to create a composite preposition use score. Treating each essay score as one item in the composite, internal consistency reliability was low (α = .389).

Cognitive process words. The cognitive process words score were calculated by LIWC as the proportion of words used falling under LIWC's cognitive process word category out of all words used in that text sample. The current list of LIWC prepositions includes 797 words across six subcategories: insight words, causation words, discrepancy words, tentative words, certainty words, and differentiation words. Examples of words in this category include "cause," "know," "ought," "think," "because," "would," "perhaps," "always," and "else." Internal consistency measures of cognitive process words approach psychometric standards and are generally high compared to other LIWC word categories ($\alpha_{uncorrected}$ = .65, $\alpha_{corrected}$ = .92; Pennebaker et al., 2015). The proportion of cognitive process words in each essay was averaged together for each person to create a composite cognitive process word use score. Treating each essay score as one item in the composite, internal consistency reliability was very low (α = .178).

Demographics. Basic demographic information were collected, including gender, age, race, ethnicity, employment status, employment industry, job tenure, average hours per week of work, and Amazon MTurk Master Worker status.

Procedure

Participants were paid $9.00, a rate approximately equivalent to US$4.50-5.15 per hour for 105 to 120 minutes of participation. Participants signed up for the HIT, then followed a link to a Qualtrics survey. They read a description of the study and were given consent information. If they accepted, they responded to three GRE analytical writing task prompts lasting 5 minutes each (1 minute each at a minimum). After the writing task, they completed the battery of 12 cognitive tests. Finally, they completed a demographic survey to complete the HIT. Participants typed their Worker ID into the last page of the survey to ensure they were compensated later on the MTurk website, where they typed the same worker ID into a separate form.

RESULTS

Data Management

The raw data were downloaded from Qualtrics as a CSV file. First, the data were processed through LIWC. The three GRE essay responses for each participant were selected and analyzed in LIWC2015 with default settings, creating proportions of word counts for every LIWC category available to total words produced. Because each essay needed to be analyzed separately, LIWC analyses were conducted three times, once per essay prompt. This resulted in three copies of the original CSV data set, each with a different set of LIWC variables appended to the last column of each file. As discussed previously, LIWC no longer calculates scores for unique word use. Thus, unique word use proportions were next calculated using R. The CSV was imported into R and the text data were cleaned to prepare for analyses. The class of the essay variables was changed to characters from the default, factors. All punctuation marks were removed except for apostrophes, intra-word dashes, ampersands, dollar signs, and percent signs. Double white spaces between sentences were removed. All text was converted to lower case. A single space character was often remaining at the end of each essay, which was also removed. Cleaning the data in this way allowed the separation of essays into lists of individual words separated by single blank spaces. The total number of words in each essay were then counted. A function was written for identifying and counting unique words in each list, which was then applied to the word lists. The number of unique words was divided by the total words for each participant's essay, then multiplied by 100 to become a percentage, effectively recreating the unique words score created in earlier versions of LIWC. These new variables were written to a new CSV file. There were five separate data files in total after text analyses. Variables were renamed and merged into one master data file on participants' Amazon MTurk Worker ID codes.

Data Cleaning

First, the data set were checked for missing data. Participants who did not complete all three of the essays were excluded from analysis (i.e., in accordance with Rule 4 of Table 1). Because of the nature of the cognitive ability battery, participants may have run out of time on each individual test before they were able to answer every question on the test. Instead of excluding participants for incomplete tests, scores were calculated based on the number of correct responses and missing data were counted as incorrect. Composite essay scores were calculated by averaging the three observed scores for each variable.

Assumption Checking

After the data were cleaned, descriptive statistics were examined. Descriptive statistics for the LIWC word category composites are presented in Table 6, descriptive statistics for the cognitive test battery are presented in Table 7, a correlation matrix of all observed variables is presented in Table 8, and a correlation matrix of all latent variables is presented in Table 9. First, the LIWC word category composite scores were compared to the base rates in the LIWC manual (Pennebaker et al., 2015) to ensure that the data were representative of other text data. These estimates are presented in Table 6, along with descriptive statistics for these language outcomes. The means and standard deviations of the LIWC word category scores varied similarly to the estimates from the LIWC manual, suggesting that these data are representative of typical text data. Mean composite scores in all categories were slightly higher than the base rates, which may be due to the cognitively-demanding nature of the writing task.

The individual LIWC scores from each essay were compared to one another to investigate meaningful differences among the essay prompts other than essay content. In general, essay length did not vary differently across essay prompts, yielding similar means, standard deviations, minimum word counts, and maximum word counts. Essay differences are presented in Table 10. Means of all LIWC variables differed by 0% to 4%, yielding very similar means across each variable type (e.g., long word use). Standard deviations differed by less than 1% across all LIWC variables. Of all essay variables calculated, the total word count for essay 2 was the most skewed distribution, while the other variables were not greatly skewed positively or negatively. The kurtosis of the distributions of each variable across essays did not vary in any consistent way. Given the data in Table 10, the essay prompts did not appear to differ in a meaningful way beyond essay content.

Next, the statistical assumptions necessary for regression and structural equation modeling were checked. First, the linearity of the relationships examined was assessed by plotting each relationship with a scatterplot, regression line, and loess line. All the examined relationships exhibited small linear effects. None of the loess lines greatly departed from the linear regression line, indicating that the variables under observation were linearly related to one another. Second, multivariate normality and multivariate outliers were assessed by examining the standardized residuals of the covariances and correlation matrices among all of the observed variables. Multivariate normality was assumed if the number of standardized residuals exceeding 1.96 in magnitude (i.e., the z-score value for 2 standard deviations) was at a rate equal to or less than .05 (i.e., 1 in 20) across tests. Using Mplus, 76 standardized residuals were estimated and 6 residuals exceeded 1.96 in magnitude (i.e., a rate of .078). Thus, the assumption of multivariate normality was questionable. To combat the violation of this assumption, bootstrapped confidence intervals were estimated for each parameter value (i.e., with 1000 iterations) and referenced for hypothesis testing instead of estimated symmetric standard errors. Bootstrapping draws randomly from the sampled data to create a data-derived sampling distribution of each estimated parameter. Thus, the exclusion of 0 within each 95% bootstrapped confidence interval indicated statistical significance.

Table 6

Descriptive Statistics of the Present Study LIWC Composites and LIWC2015 Estimates

Variable	Mean	*SD*	Min	Max	Skewness	Kurtosis	Alpha
Word Count							
Present Study	113.84	36.09	23.67	295.67	0.84	1.79	.901
LIWC2015 Estimate	N/A	N/A					
Long Words							
Present Study	21.62	4.52	8.07	40.75	0.44	1.19	.726
LIWC2015 Estimate	15.60	3.76					
Unique Words							
Present Study	66.72	5.57	48.20	84.04	0.02	0.33	.697
LIWC2015 Estimate	N/A	N/A					
Conjunctions							
Present Study	6.73	1.63	2.15	11.06			.384
LIWC2015 Estimate	5.90	1.57					
Prepositions							
Present Study	14.87	2.02	4.77	20.09	-0.93	2.91	.389
LIWC2015 Estimate	12.93	2.11					
Cognitive Process Words							
Present Study	16.42	2.88	9.87	25.78	0.18	-0.10	.178
LIWC2015 Estimate	10.61	3.02					

N = 373.

Note: Mean, SD, Min, and Max for all variables except word count are percentages out of 100%. LIWC2015 Mean and SD are base rates sampled from a variety of writing outlets in a variety of contexts provided by Pennebaker and colleagues (2015). Coefficient alpha was calculated by treating each essay as 1 item of a 3-item test.

Table 7
Descriptive Statistics of the Cognitive Ability Battery

Test	Mean	*SD*	Min	Max	Skewness	Kurtosis	Items	Alpha
Gc1	58.45	17.65	8.33	100.00	-0.25	-0.40	24	.783
Gc2	60.32	18.74	5.56	100.00	-0.08	-0.39	18	.770
Gc3	61.02	18.77	5.56	100.00	-0.31	-0.25	18	.743
Gsm1	47.37	23.20	0.00	100.00	0.44	-0.30	22	.898
Gsm2	52.50	23.47	0.00	100.00	0.08	-0.53	21	.893
Gsm3	31.09	19.18	0.00	95.83	0.89	0.75	24	.875
Gf1	60.88	22.61	6.67	100.00	-0.40	-0.83	15	.792
Gf2	40.12	19.68	0.00	85.71	0.24	-0.49	14	.624
Gf3	51.63	15.58	6.25	94.64	-0.01	-0.16	112	.939

N = 352.

Note: Gsm1 items 13 and 17 were dropped because they had no variance (i.e., everyone got them wrong). These were the two longest digit span items, 12 digits each in length. Gsm2 items 9, 13, and 17 were dropped for zero variance as well. These were the longest items and thus the hardest spanning 12 to 13 digits each.

Table 8
Correlation Matrix of All Observed Variables

	WC	Long	Unique	Conj	Prep	Cogproc	Gc1	Gc2	Gc3	Gsm1	Gsm2	Gsm3	Gf1	Gf2	Gf3
WC	1														
Long	-.01	1													
Unique	-.74	.16	1												
Conj	.24	-.04	-.14	1											
Prep	.14	.10	-.10	-.13	1										
Cogproc	-.01	-.05	-.10	.09	-.12	1									
Gc1	.27	.10	-.09	-.03	.07	-.01	1								
Gc2	.28	.06	-.09	-.01	.08	-.05	.73	1							
Gc3	.24	.09	-.10	-.05	.02	-.10	.72	.73	1						
Gsm1	.06	.05	-.07	.05	.05	-.01	.00	.04	.03	1					
Gsm2	.07	.14	-.04	-.01	.10	.01	.06	.05	.05	.66	1				
Gsm3	.20	-.08	-.15	.03	.11	.02	.22	.22	.22	.48	.47	1			
Gf1	.21	.05	-.10	-.10	.16	.04	.41	.38	.37	.06	.21	.22	1		
Gf2	.07	.03	-.02	-.01	.06	.01	.24	.19	.20	.03	.07	.12	.54	1	
Gf3	.14	-.01	-.08	.04	.10	.04	.17	.06	.14	.03	.13	.17	.33	.32	1

Note. N = 373. All correlations greater than or equal to .10 in magnitude are significant at the α = .05 level. All correlations greater than or equal to .14 in magnitude are significant at the α = .01 level. WC = word count composite. Long = words with seven or more characters composite. Unique = unique words composite. Conj = conjunctions composite. Prep = prepositions composite. Cogproc = cognitive process words composite. Gc = verbal ability. Gsm = short-term memory. Gf = fluid reasoning.

Table 9

Correlation Matrix of All Latent Variables

	GCA	Gc	Gsm	Gf	Long	Unique	Conj	Prep	Cogproc
GCA	1								
Gc	.00	1							
Gsm	.00	.00	1						
Gf	.00	.00	.00	1					
Long	-.04	.17	.09	.09	1				
Unique	-.23	.06	-.05	.09	.16	1			
Conj	-.14	.09	.07	.05	-.04	-.14	1		
Prep	.30	-.19	.03	-.07	.10	-.10	-.13	1	
Cogproc	.12	-.20	-.03	-.06	-.05	-.10	.09	-.12	1

Note. N = 373. GCA = general cognitive ability. Gc = verbal ability. Gsm = short-term memory. Gf = fluid reasoning. Long = words with seven or more characters composite. Unique = unique words composite. Conj = conjunctions composite. Prep = prepositions composite. Cogproc = cognitive process words composite.

Table 10
Differences in LIWC Variables across Essay Prompts

	M	*SD*	Min	Max	Skewness	Kurtosis
WC1	118.75	40.77	19.00	293.00	0.88	1.48
WC2	111.03	38.25	30.00	286.00	1.07	1.98
WC3	111.67	39.41	18.00	308.00	0.85	2.31
Long1	21.65	5.80	6.33	42.22	0.47	0.43
Long2	20.55	5.23	4.58	38.38	0.12	0.29
Long3	22.65	5.83	7.29	48.35	0.45	1.43
Unique1	67.39	7.00	46.79	91.84	0.16	0.67
Unique2	64.11	7.19	42.71	91.43	0.21	0.17
Unique3	68.66	7.00	43.50	92.11	0.25	0.55
Conj1	6.57	2.49	0.00	14.53	0.16	0.31
Conj2	6.55	2.38	1.33	13.43	0.36	-0.05
Conj3	7.07	2.41	0.00	13.85	0.05	-0.09
Prep1	15.75	3.02	3.85	24.00	-0.24	0.58
Prep2	14.67	3.01	6.38	23.53	-0.06	0.00
Prep3	14.18	3.02	0.00	22.22	-0.35	1.50
Cogproc1	17.49	4.74	4.35	33.33	0.19	0.32
Cogproc2	14.15	4.42	3.12	29.23	0.46	0.59
Cogproc3	17.63	4.90	3.45	31.53	-0.03	-0.09

N = 373. WC = total word count. Long = use of words with seven or more characters. Unique = unique word use. Conj = conjunction use. Prep = preposition use. Cogproc = cognitive process word use.

A bifactor analysis was used to check the dimensionality of GCA and the broad cognitive abilities. Model fit indices were calculated and standards for good model fit were set a priori. A non-significant chi square statistic would indicate good model fit. Additionally, an SRMR index less than .05, a CFI index greater than .95, a TLI index greater than .90, and an RMSEA index less than .05 would indicate good model fit. Each of the cognitive tests were loaded onto latent factors representing the broad cognitive abilities underlying test performance. Each of the cognitive tests was also loaded onto a latent GCA factor simultaneously. All the correlations among the broad cognitive abilities and GCA were set to equal zero. A chi-square goodness of fit test indicated that the data did not fit the model well, $\chi^2(18, N = 373) = 30.41, p = .034$. However, chi-square tests have two limitations which are relevant to the present study. First, the chi-square test assumes multivariate normality, which may cause a model to be rejected even when it is properly specified (McIntosh, 2007; Hooper, Coughlan, & Mullen, 2008). Second, the chi-square test is sensitive to sample size, meaning it will almost always reject the model with a large enough sample size (Bentler & Bonett, 1980; Hooper et al., 2008). Due to the sample size and multivariate non-normality of the sample, other fit indices were investigated to triangulate the model fit of the confirmatory factor analysis. By the standards set for multiple fit indices, the model fit the data well, RMSEA = .043, CFI = .990, TLI = .980, SRMR = .028. The measurement model with factor standardized factor loadings for the bi-factor GCA model is presented in Figure 2. The tests of Gc consistently loaded onto the latent GCA and Gc factors across all three tests, loading more strongly on Gc. The tests of Gsm loaded highly on latent Gsm, but not very highly onto latent GCA. Specifically, Gsm1, the audio number span test, had the weakest loading onto GCA of all 9 indicators. The tests of Gf were moderately loaded onto both latent Gf and GCA, but Gf1, the letter sets test, loaded much higher than the other tests onto

GCA while Gf2, the locations test, loaded much higher than the others onto Gf. The test for Gf1 was the largest loading across all tests onto GCA.

Hypothesis Testing

The composite LIWC outcomes were added to the CFA measurement model for GCA and the broad cognitive abilities to form the full structural equation model. Both hypothesized and exploratory paths were added connecting the latent GCA and broad abilities to the composite LIWC outcomes. Each LIWC outcome was freely correlated with each other LIWC outcome. This full model is presented in Figure 3. A chi-square goodness of fit test indicated that the data did not fit the model well, $\chi^2(43, N = 373) = 67.13$, $p = .011$. Relative model fit indices were examined as chi-square tests are sensitive to sample size and multivariate non-normality (Bentler & Bonett, 1980; McIntosh, 2007; Hooper et al., 2008), which was present in this sample. By the standards of these fit indices, the model fit the data well, RMSEA = .039, CFI = .982, TLI = .961, SRMR = .027. Overall, the model fit was adequate for testing hypotheses.

Once the exploratory model was fitted, exploratory path estimates were examined as an investigation of Research Question 1. Statistical significance at the $p < .05$ level as well as practically meaningful effect sizes in the hypothesized direction were set a priori as criteria indicating support for each hypothesis and exploratory path. Bootstrapped 95% confidence intervals were calculated around the unstandardized estimates. Confidence intervals that did not contain zero were interpreted as statistically significant. Because each latent cognitive ability was modeled while controlling for the others, each path estimate indicated the incremental predictive variance over the other latent cognitive abilities. The standardized parameter estimates, unstandardized parameter estimates, and bootstrapped confidence intervals around the unstandardized estimates are presented in Table 12.

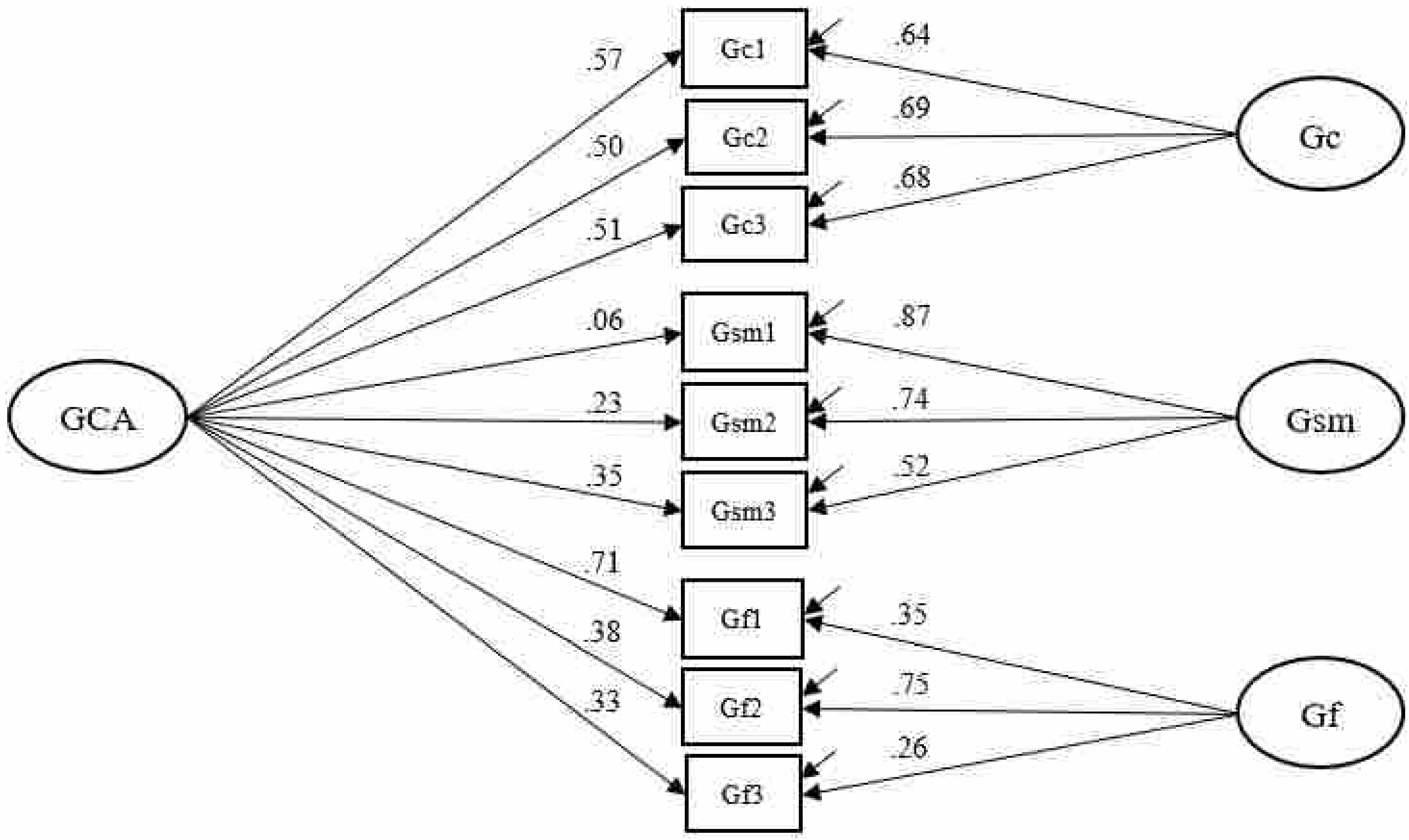


Figure 2. Confirmatory Factor Analysis of the Bi-factor GCA Model.
Note. GCA = general cognitive ability. Gc = verbal ability. Gsm = short-term memory. Gf = fluid reasoning. $\chi^2(18, N = 373) = 30.41$, $p = .034$; RMSEA = .043; CFI = .990; TLI = .980; SRMR = .028.

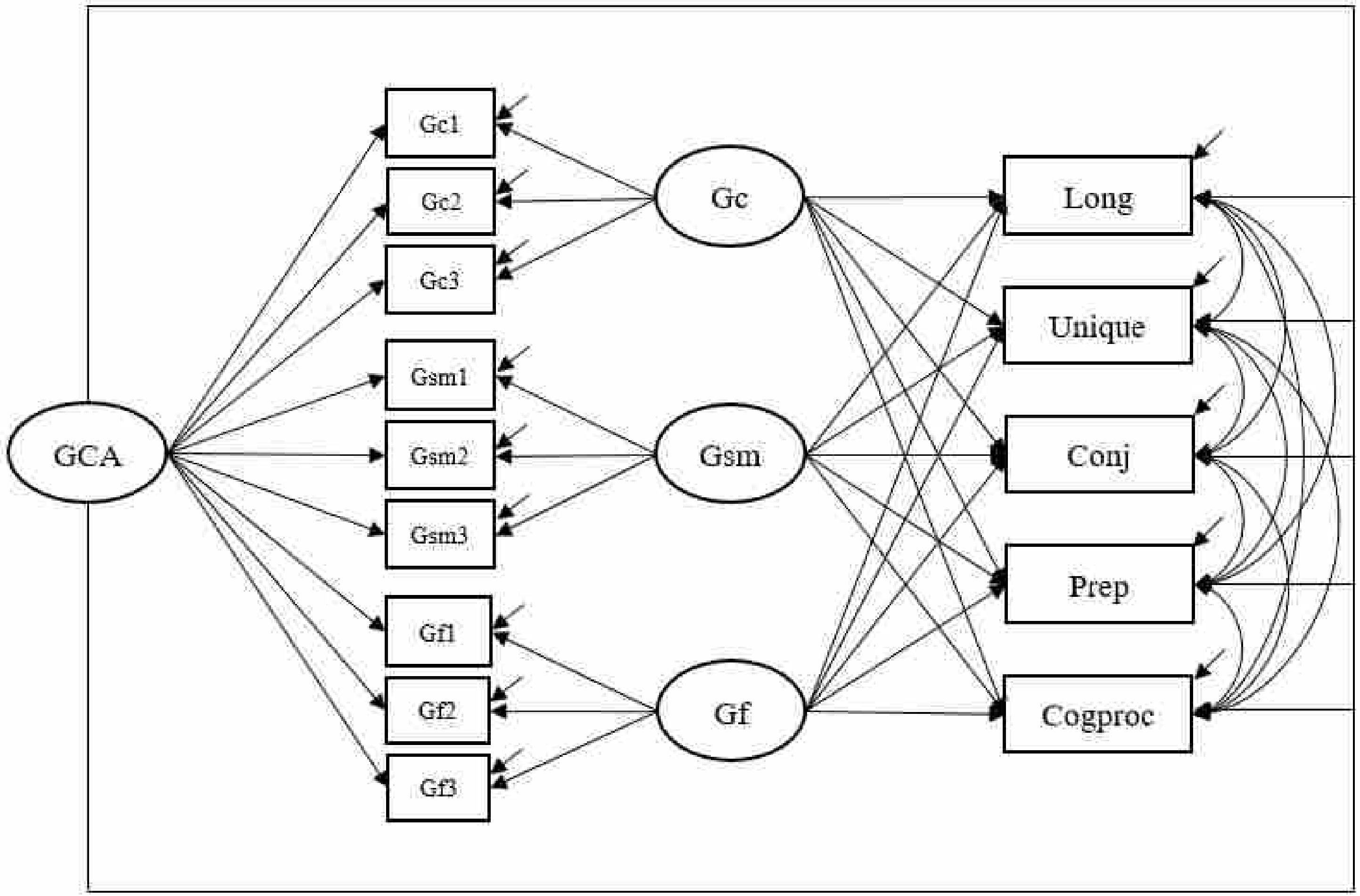


Figure 3. Exploratory Structural Equation Path and Measurement Model
Note: Squares indicate measured variables. Ovals indicate latent constructs. GCA = general cognitive ability. Gc = verbal ability. Gsm = short-term memory. Gf = fluid reasoning. Long = words with more than 6 characters. Unique = unique words. Conj = conjunctions. Prep = prepositions. Cogproc = cognitive process words. $\chi^2(43, N = 373) = 67.13$, $p = .011$; RMSEA = .039; CFI = .982; TLI = .961; SRMR = .027.

Table 11
Parameter Estimates and Bootstrapped Confidence Intervals for the Exploratory Model

Parameter	Standardized	Unstandardized	Lower 95% CI	Upper 95% CI
GCA → Long	-.04	-0.02	-0.33	0.11
GCA → Unique	-.23	-0.12	-0.32	0.01
GCA → Conj	-.14	-0.02	-0.08	0.03
GCA → Prep	.30	0.06	0.01	0.14
GCA → Cogproc	.12	0.03	-0.05	0.10
Gc → Long	.17	0.08	-0.08	0.56
Gc → Unique	.06	0.03	-0.15	0.44
Gc → Conj	.09	0.01	-0.04	0.25
Gc → Prep	-.19	-0.04	-0.23	0.02
Gc → Cogproc	-.20	-0.06	-0.23	0.07
Gsm → Long	.09	0.02	-0.02	0.08
Gsm → Unique	-.05	-0.01	-0.06	0.03
Gsm → Conj	.07	0.01	-0.01	0.02
Gsm → Prep	.03	0.00	-0.02	0.02
Gsm → Cogproc	-.03	-0.01	-0.03	0.02
Gf → Long	.09	0.04	-0.19	0.26
Gf → Unique	.09	0.05	-0.07	0.29
Gf → Conj	.05	0.01	-0.03	0.13
Gf → Prep	-.07	-0.01	-0.16	0.02
Gf → Cogproc	-.06	-0.02	-0.13	0.04
Long w/ Unique	.14	3.37	-14.71	11.00
Long w/ Conj	-.08	-0.53	-10.98	2.13
Long w/ Prep	.16	1.34	-2.59	11.25
Long w/ Cogproc	-.01	-0.10	-4.79	10.12
Unique w/ Conj	-.18	-1.57	-6.38	0.03
Unique w/ Prep	-.02	-0.19	-2.53	6.71
Unique w/ Cogproc	-.07	-0.98	-4.03	5.82
Conj w/ Prep	-.07	-0.22	-0.93	2.30
Conj w/ Cogproc	.14	0.60	-0.17	3.18
Prep w/ Cogproc	-.22	-1.17	-4.75	-0.04

Note. Bootstrapped confidence intervals are around the unstandardized estimates.

No exploratory paths were statistically significant at the $p < .05$ level. Thus, because inclusion of these exploratory paths decreased statistical power to test the theoretical model, the theoretical model was fitted without freeing any of the exploratory paths. In this final model, Mplus indicated that the residual variance of Gf2, the locations test, was negative. The negative residual variance was not statistically significantly different from zero, indicating that it may have been a sample fluctuation (Dillon, Kumar, & Mulani, 1987) or the test's true score correlation with Gf may have indeed been 1.00. Regardless of cause, to address this problem from a modeling perspective, the residual was set to zero. With this modification, a chi-square goodness-of-fit test indicated that the data slightly misfit the model, $\chi^2(54, N = 373) = 81.18$, $p = .010$, but relative fit indices indicated good model fit, RMSEA = .037; CFI = .979; TLI = .965; SRMR = .033. This final theoretical model as tested is presented in Figure 4. The standardized estimates, unstandardized estimates, and bootstrapped confidence intervals for the hypothesized model omitting exploratory paths is presented in Table 12. A summary of all hypothesis test results is presented in Table 13.

Table 12
Parameter Estimates and Bootstrapped Confidence Intervals for the Hypothesized Model

Parameter	Standardized	Unstandardized	Lower 95% CI	Upper 95% CI
GCA → Long	.05	0.03	-0.06	0.11
GCA → Unique	-.15	-0.09	-0.18	-0.01
GCA → Conj	.03	-0.02	-0.05	0.00
GCA → Prep	.21	0.05	0.01	0.09
GCA → Cogproc	.03	0.01	-0.03	0.06
Gc → Long	.09	0.03	-0.03	0.10
Gc → Unique	-.04	-0.02	-0.10	0.05
Gsm → Conj	.05	0.00	-0.01	0.01
Gsm → Prep	.04	0.00	-0.01	0.02
Gf → Cogproc	.00	0.00	-0.10	0.05
Long w/ Unique	.17	4.15	1.43	6.98
Long w/ Conj	-.04	-0.31	-1.08	0.44
Long w/ Prep	.10	0.85	-0.27	1.83
Long w/ Cogproc	-.05	-0.62	-2.06	0.66
Unique w/ Conj	-.15	-1.33	-2.31	-0.41
Unique w/ Prep	.10	-0.81	-2.01	0.43
Unique w/ Cogproc	-.10	-1.63	-3.44	-0.05
Conj w/ Prep	-.11	-0.36	-0.70	0.01
Conj w/ Cogproc	.09	0.44	-0.07	0.99
Prep w/ Cogproc	-.13	-0.75	-1.37	-0.20

Note. Bootstrapped confidence intervals are around the unstandardized estimates.

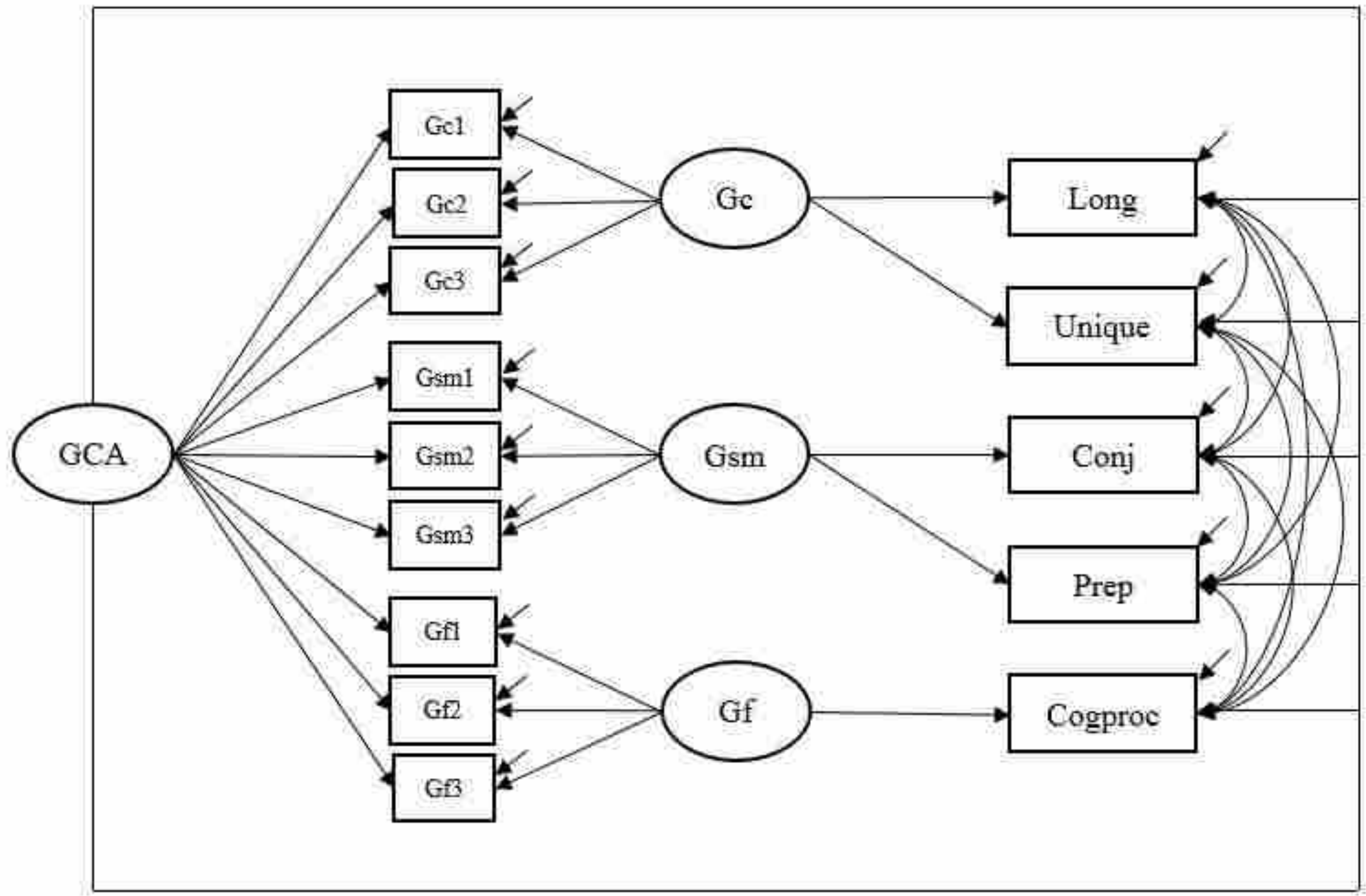


Figure 4. Hypothesized Structural Equation Path and Measurement Model.
Note: Squares indicate measured variables. Ovals indicate latent constructs. GCA = general cognitive ability. Gc = verbal ability. Gsm = short-term memory. Gf = fluid reasoning. Long = words with more than 6 characters. Unique = unique words. Conj = conjunctions. Prep = prepositions. Cogproc = cognitive process words. $\chi^2(54, N = 373) = 81.18$, $p = .010$; RMSEA = .037; CFI = .979; TLI = .965; SRMR = .033.

Table 13
Summary of Hypothesis Test Results

No.	Hypothesis	Supported
1a	GCA will provide incremental prediction of long word use over Gc.	No
1b	GCA will provide incremental prediction of unique word use over Gc.	Partially
1c	GCA will provide incremental prediction of conjunction use over Gsm.	No
1d	GCA will provide incremental prediction of preposition use over Gsm.	Yes
1e	GCA will provide incremental prediction of cognitive process word use over Gf.	No
2	Gc will provide incremental prediction of long word use over GCA.	No
3	Gc will provide incremental prediction of unique word use over GCA.	No
4	Gsm will provide incremental prediction of conjunction use over GCA.	No
5	Gsm will provide incremental prediction of preposition use over GCA.	No
6	Gf will provide incremental prediction of cognitive process word use over GCA.	No

Hypothesis 1 stated that latent GCA would provide incremental prediction of (a) long word use, (b) unique word use, (c) conjunction word use, (d) preposition word use, and (e) cognitive process word use beyond broad cognitive abilities. Hypothesis 1a was not supported. The confidence interval around the path estimate contained zero, meaning that the estimate did not reach statistical significance (i.e., the null hypothesis could not be rejected). The partial correlation between latent GCA and long word use (i.e., the incremental contribution of GCA to predicting long word use after removing all variance and covariance associated with Gc) was very small (r = .05).

Hypothesis 1b was partially supported. While controlling for Gc, GCA had a statistically significant negative effect on unique word use. This finding was contrary to the hypothesized direction. However, latent GCA did add incremental prediction of unique word use over Gc. The partial correlation between latent GCA and unique word use (i.e., the incremental contribution of GCA to predicting unique word use after removing all variance and covariance associated with Gc) was small in size (r = -.15).

Hypothesis 1c was not supported. The confidence interval around the path estimate contained zero, meaning that the estimate did not reach statistical significance. The partial correlation between latent GCA and conjunction use (i.e., the incremental contribution of GCA to predicting conjunction use after removing all variance and covariance associated with Gsm) was small (r = -.14).

Hypothesis 1d was fully supported. While controlling for Gsm, GCA had a statistically significant positive effect on preposition use. Latent GCA added incremental prediction of preposition use over Gsm. The partial correlation between latent GCA and preposition use (i.e., the incremental contribution of GCA to predicting preposition use after removing all variance and covariance associated with Gsm) was medium in size (r = .30).

Hypothesis 1e was not supported. The confidence interval around the path estimate contained zero, meaning that the estimate did not reach statistical significance. The partial correlation between latent GCA and cognitive process word use (i.e., the incremental contribution of GCA to predicting cognitive process word use after removing all variance and covariance associated with Gf) was very small (r = .03).

Thus, latent GCA added incremental prediction of unique word use over latent Gc and incremental prediction of preposition use over latent Gsm. Latent GCA did not add incremental prediction of long word use over latent Gc, of conjunction use over latent Gsm, or of cognitive process word use over latent Gf.

Hypothesis 2 stated that latent Gc would provide incremental prediction of longer word use over latent GCA. This hypothesis not was supported. The confidence interval around the path estimate contained zero, meaning that the estimate did not reach statistical significance. The partial correlation between latent Gc and long word use (i.e., the incremental contribution of Gc to predicting long word use after removing all variance and covariance associated with GCA) was small (*r* = .09).

Hypothesis 3 stated that latent Gc would provide incremental prediction of unique word use over GCA and other broad abilities. This hypothesis not was supported. The confidence interval around the path estimate contained zero, meaning that the estimate did not reach statistical significance. The partial correlation between latent Gc and unique word use (i.e., the incremental contribution of Gc to predicting unique word use after removing all variance and covariance associated with GCA) was very small (*r* = -.04).

Hypothesis 4 stated that latent Gsm would provide incremental prediction of conjunction word use over GCA and other broad abilities. This hypothesis not was supported. The confidence interval around the path estimate contained zero, meaning that the estimate did not reach statistical significance. The partial correlation between latent Gsm and conjunction use (i.e., the incremental contribution of Gsm to predicting conjunction use after removing all variance and covariance associated with GCA) was very small (*r* = .05).

Hypothesis 5 stated that latent Gsm would provide incremental prediction of preposition use over GCA and other broad abilities. This hypothesis not was supported. The confidence interval around the path estimate contained zero, meaning that the estimate did not reach statistical significance. The partial correlation between latent Gsm and preposition use (i.e., the incremental contribution of Gsm to predicting preposition use after removing all variance and covariance associated with GCA) was very small (*r* = .04).

Hypothesis 6 stated that latent Gf would provide incremental prediction of cognitive process word use over GCA and other broad abilities. This hypothesis not was supported. The confidence interval around the path estimate contained zero, meaning that the estimate did not reach statistical significance. The partial correlation between latent Gf and cognitive process word use (i.e., the incremental contribution of Gf to predicting cognitive process word use after removing all variance and covariance associated with GCA) was zero (r = .00).

Research Question

As described earlier, Research Question 1 sought to understand the relationships between latent Gc and other word count outcomes, latent Gsm and other word count outcomes, and latent Gf and other word count outcomes while controlling for GCA and other broad abilities. None of the exploratory relationships reached statistical significance within the exploratory model. In light of these findings, the latent correlation matrix (Table 9) and the R^2 estimates for the LIWC outcome variables from the exploratory model (Table 14) were examined to draw conclusions about the previously unexplored relationships of latent broad abilities to LIWC outcomes.

Table 14
Percent of Variance Explained in Each LIWC Outcome

Outcome Variable	R^2
Long Word Use	.046
Unique Word Use	.064
Conjunction Use	.034
Preposition Use	.133
Cognitive Process Word Use	.058

When including latent GCA and all three latent broad cognitive abilities in the exploratory model, 4.6% of the total variance in long word use was explained. Latent Gsm uniquely accounted for 0.81% (i.e., the squared correlation between latent Gsm and long word use, see Table 9), while latent Gf also uniquely accounted for another 0.81% of the variance explained in long word use. In the exploratory model, latent Gc uniquely accounted for 2.89% of the variance in long word use and latent GCA uniquely accounted for 0.16% of the variance. Latent Gc accounted for more variance than any of the other predictors of long word use while latent GCA accounted for the least amount of variance among all predictors.

For unique word use, 6.4% of the variance was explained by all cognitive predictors. Latent Gsm uniquely accounted for 0.25% of the variance explained in unique word use, while latent Gf uniquely accounted for 0.81%. Latent Gc uniquely accounted for 0.36% of the variance and latent GCA uniquely accounted for 5.29% of the variance explained in unique word use. Latent GCA accounted for more variance than any of the other predictors of unique word use.

For conjunction use, 3.4% of the variance was explained by all cognitive predictors. Latent Gc uniquely accounted for 0.81% of the variance explained in conjunction use. Latent Gf uniquely accounted for 0.25% of the variance explained. Latent Gsm uniquely accounted for 0.49% and latent GCA uniquely accounted for 1.96% of the variance explained in conjunction use. Latent GCA accounted for more variance than any of the other predictors of conjunction use.

For preposition use, 13.3% of the variance was explained by all cognitive predictors. Latent Gc uniquely accounted for 3.61% of the variance explained in preposition use. Latent Gf uniquely accounted for 0.49% of the variance explained. Latent Gsm uniquely accounted for only 0.09% of the variance explained in preposition use. Latent GCA uniquely accounted for 9.00% of the variance explained, far more than any of the other predictors of preposition use. Latent Gsm accounted for less variance explained than any other predictor.

For cognitive process word use, 5.8% of the variance was explained by all cognitive predictors. Latent Gc uniquely accounted for 4.00% of the variance explained in cognitive process word use, while latent Gsm only uniquely accounted for 0.09% of the variance explained. Latent Gf uniquely accounted for 0.36% of the variance explained in cognitive process word use and latent GCA uniquely account for 1.44% of the variance explained. Latent Gc accounted for more variance than any other predictor of cognitive process word use.

Exploratory Analyses

To better understand cognitive ability expression in word counts given the mixed results for hypothesized relationships, two sets of exploratory analyses were also conducted. First, the observed correlation matrix from the exploratory model (Table 8) was examined to glean additional information about the relationships between cognitive ability tests and LIWC word count outcomes beyond planned tests. A composite score for total word count, the average word count across all three essay prompts, was included and examined more closely, given its strong negative relationship to unique word use (r = -.74). Word count positively correlated with conjunction use (r = .24) and preposition use (r = .14), suggesting that using these types of words often means including additional words beyond them. For example, any time that a preposition was used, a second word was included at a minimum (e.g., *under* there, *above* me, *through* the door, *over* the bridge). Any time that a conjunction is used, it was likely followed by an entire additional phrase, as previously discussed (e.g., "I like cats *and* I do not like mice."). Further, word count was positively correlated with the nine cognitive ability test scores to some extent. Word count had small to medium correlations with all three tests of Gc, r = .24-.28. Word count correlated less with Gsm tests, particularly the two memory span tests involving numerical digits (r = .06-.07). However, the correlation with Gsm3, the memory span test involving alphabetic letters, was stronger (r = .20). Gf1 and Gf3 had small to medium correlations with total word count (r = .21 and .14, respectively), whereas Gf2 correlated weaker (r = .07). These data suggest a positive manifold of cognitive ability tests with total word count in cognitive demanding essays, like GCA. Thus, GCA may possibly be expressed most directly through total word count in cognitively demanding essays.

Second, many of the observed correlations in Table 8 among the hypothesized and exploratory variable pairings were small in effect size, so the observed correlation matrix was also examined using two datasets constructed using different data cleaning standards to determine if cleaning strategy attenuated any observed relationships. In the first dataset, a larger sample size was examined (N = 393) by including all careless responders but still excluding the most problematic participants, including any participants that plagiarized essays, skipped essays, were located outside of the United States, or experienced technical errors. The observed correlation matrix for this sample is presented in Table 15. In the second dataset, a smaller sample size was examined (N = 298, presented in Table 16), which excluded the above participants in addition to all participants with any indicator of careless responding (i.e., only including participants passing 12/12 exclusionary criteria in Table 2). Comparing the observed correlation matrices for these two sample sizes would indicate effect size differences due to careless responding in relation to the final sample size for analyses (N = 373).

Most of the correlations in Tables 14 and 15 are of a similar magnitude, though a few differences are noteworthy. First, the correlations between total word count and all constructs are generally stronger for the group including careless responders. The relationships between tests of Gc and word count were about .10 higher in the larger sample, the largest difference for these constructs. Second, the relationships between tests of Gc and long word use were stronger for the group excluding careless responders (r = .14-.22 vs. r = .11-14). Third, the relationships between unique word use and tests of Gc were weaker for the group excluding careless responders (r = -

.05 to -.10 vs. r = -.26 to -.27). Fourth, the relationships between tests of Gsm and conjunction use, as well as with preposition use, was not consistently different across sample sizes, continuing to remain weak or non-existent in general. Fifth, the relationships between tests of Gf and cognitive process word use also tended to not meaningfully differ across sample sizes, remaining near-zero. Sixth, the relationships between Gc and preposition use were generally stronger for the group including careless responders (r = .13-.19 vs. r = -.02 to .05), suggesting that the expression of Gc through preposition use may be spurious in nature. Finally, no other clear patterns of difference emerged among the relationships of interest in Tables 14 and 15. A few additional correlations reached statistical significance and small to medium effect sizes, but not in a way aligned with hypotheses. For example, Gf1, the letter sets test, correlated somewhat with longer word use, conjunction use, and preposition use across both samples, though weaker in the sample excluding careless responders. Gf2 and Gf3 did not correlate in a similar fashion. Thus, the inclusion of slightly careless responders in the final sample (N = 373) may have slightly increased observed correlations but not at a meaningful magnitude. Overall, differences between the final dataset and these two exploratory datasets were minimal.

Table 15

Correlation Matrix of Observed Variables Including All Careless Responders

	WC	Long	Unique	Conj	Prep	Cogproc	Gc1	Gc2	Gc3	Gsm1	Gsm2	Gsm3	Gf1	Gf2	Gf3
WC	1														
Long	.05	1													
Unique	-.78	.05	1												
Conj	.24	-.02	-.17	1											
Prep	.23	.14	-.22	-.07	1										
Cogproc	-.01	.03	-.09	.08	-.13	1									
Gc1	.37	.14	-.26	.00	.17	.02	1								
Gc2	.38	.11	-.27	.01	.19	-.02	.76	1							
Gc3	.35	.14	-.26	-.03	.13	-.07	.76	.76	1						
Gsm1	.01	.08	.00	.01	-.02	-.03	-.04	-.01	.01	1					
Gsm2	.06	.17	-.02	-.03	.05	-.02	.04	.04	.07	.70	1				
Gsm3	.29	-.02	-.26	.03	.17	.02	.29	.30	.31	.43	.44	1			
Gf1	.29	.11	-.19	-.07	.22	.04	.45	.42	.43	.04	.20	.28	1		
Gf2	.12	.06	-.08	.00	.10	.02	.28	.23	.24	.02	.08	.15	.55	1	
Gf3	.17	.02	-.12	.04	.12	.05	.20	.10	.17	.03	.14	.19	.34	.33	1

Note. N = 393. All correlations greater than or equal to .11 in magnitude are significant at the $p < .05$ level. All correlations greater than or equal to .13 in magnitude are significant at the $p < .01$ level. WC = word count composite. Long = words with seven or more characters composite. Unique = unique words composite. Conj = conjunctions composite. Prep = prepositions composite. Cogproc = cognitive process words composite. Gc = verbal ability. Gsm = short-term memory. Gf = fluid reasoning.

Table 16
Correlation Matrix of Observed Variables Excluding All Careless Responders

	WC	Long	Unique	Conj	Prep	Cogproc	Gc1	Gc2	Gc3	Gsm1	Gsm2	Gsm3	Gf1	Gf2	Gf3
WC	1														
Long	.08	1													
Unique	-.73	.09	1												
Conj	.23	.01	-.12	1											
Prep	.08	.14	-.06	-.15	1										
Cogproc	.02	-.03	-.13	.12	-.14	1									
Gc1	.22	.22	-.05	-.04	-.02	-.03	1								
Gc2	.27	.14	-.10	-.03	.05	-.05	.69	1							
Gc3	.23	.21	-.09	-.07	-.03	-.11	.69	.71	1						
Gsm1	.06	.16	-.06	.06	.08	-.01	.04	.05	.07	1					
Gsm2	.05	.19	-.02	-.04	.08	-.02	.04	.00	.07	.76	1				
Gsm3	.20	-.01	-.15	.05	.07	-.01	.24	.24	.24	.43	.46	1			
Gf1	.22	.14	-.07	-.12	.12	-.01	.36	.33	.35	.06	.14	.17	1		
Gf2	.08	.07	-.01	-.03	.04	-.03	.23	.18	.19	.03	.05	.09	.54	1	
Gf3	.12	.01	-.07	.11	.11	.01	.14	.03	.12	.03	.10	.14	.28	.30	1

Note. $N = 298$. All correlations greater than or equal to .12 in magnitude are significant at the $p < .05$ level. All correlations greater than or equal to .15 in magnitude are significant at the $p < .01$ level. WC = word count composite. Long = words with seven or more characters composite. Unique = unique words composite. Conj = conjunctions composite. Prep = prepositions composite. Cogproc = cognitive process words composite. Gc = verbal ability. Gsm = short-term memory. Gf = fluid reasoning.

DISCUSSION

The present study sought to establish the strength of GCA and broad cognitive ability expression in a cognitively demanding essay context using an established word count approach to NLP. GCA was expected to incrementally predict word count outcomes while controlling for other broad cognitive abilities. Broad cognitive abilities were expected to incrementally predict word count outcomes while controlling for GCA and other broad abilities. Almost all of the hypothesized relationships were small or different than expected. GCA incrementally predicted two LIWC outcomes over broad cognitive abilities, but broad cognitive abilities did not incrementally predict LIWC outcomes over GCA. However, these findings suggest that GCA and verbal ability are expressed to some extent through word count proportions in cognitively demanding essays. Additionally, exploratory analyses were conducted, which will aid in establishing base rates for the relationships between word count categories and cognitive abilities in the literature.

This study contributes to the literature on psychological assessment via NLP in three additional ways beyond the findings related to GCA and LIWC outcomes. First, it provides empirical estimates of relationships between broad cognitive abilities and several LIWC word count outcomes. Previous research has examined GCA in relation to LIWC word outcomes (Weaver, 2017) and broad cognitive abilities to language variables (e.g., Kemper & Sumner, 2001), but this study was the first to examine broad cognitive abilities in relation to LIWC outcome variables specifically. Second, this study provides evidence against specific aptitude theory in the context of writing performance. GCA predicted writing behavior incrementally over broad cognitive abilities whereas broad cognitive abilities did not incrementally predict those same behaviors over GCA. Critics of specific aptitude theory claim that GCA is all that matters in predicting job performance (Schmidt & Hunter, 2004). In the present study, GCA was relevant to written communication, an aspect of job performance (Campbell et al., 1993). In an ever increasingly connected world, the assessment and prediction of written communication via NLP may play a critical role in the workplace. When communicating with others online, it is important to be detailed, clear, and tactful as many aspects of face-to-face communication are lost (e.g., tone and facial expressions). Automatic assessments of writing performance already factor in word length, uniqueness of words and content, conjunction use, and preposition use (Deane, 2013), and the present study demonstrates how GCA and Gc are expressed through some aspects of writing performance. Third, this study provides evidence that GCA and verbal ability are reflected through specific word categories in a cognitively demanding writing context. Tausczik and Pennebaker (2010) positioned several LIWC categories as markers of "cognitive complexity" rather than cognitive ability. Carroll (1993) identified a few studies measuring cognitive complexity, but did not find a relationship to GCA. Carroll concluded that cognitive complexity was a "cognitive style," a sort of miscellaneous individual difference related to cognition, but not exactly a cognitive ability. This study provides some evidence to suggest that longer word use, conjunction use, and cognitive processing word use may be types of cognitive styles, as they were not related to cognitive abilities in a consistently positive way.

Overall, GCA did not incrementally predict word count outcomes over other broad abilities entirely as hypothesized. Specifically, GCA did not incrementally predict long word use over Gc, conjunction use over Gsm, or cognitive process word use over Gf. However, GCA did add meaningful incremental prediction of unique word use over Gc (β = -.15, a small effect size, see Table 12) and other broad cognitive abilities (β = -.23, a small to medium effect size, see Table 11), although in the opposite direction of the hypothesis. Considering the strong inverse relationship found between unique word use and total word count (*r* = -.74, see Table 8), GCA may be positively related to total word count, which would suggest that people higher in GCA use more words in total in a cognitively demanding essay context. GCA also added meaningful incremental prediction of preposition use over Gsm (β = .21, a small to medium effect size, see Table 12) and other broad cognitive abilities (β = .30, a medium effect size, see Table 11). GCA accounted for more variance in preposition use than any other predictor. Prepositions signal increased complexity and detail in writing (Tausczik & Pennebaker, 2010) and GCA is critical to processing complex information (Gottfredson, 1997a), which may explain the strong overlap in these two constructs. Additionally, the correlational findings of the present study (i.e., Table 9) are somewhat contrary to the findings of Weaver (2017), who found small to medium positive correlations for GCA in relation to several of the same LIWC word count categories (i.e., long words, conjunctions, prepositions, and cognitive process words). Compared to Weaver's study, the cognitive battery used in the present study was much more comprehensive, which could be a reason for some of the differences in estimates. Weaver also examined word counts in the context of employment resumes, which may be an important contextual factor in how word use varies across prompts and formats.

None of the broad cognitive abilities examined incrementally predicted LIWC word count outcomes over GCA as predicted. All effects were small, very small, or near-zero and none of the effects reached statistical significance. It is possible that these effects were underpowered, a type II error. Many of the effects observed were lower in size than those used in the Monte Carlo simulation used for power analysis. Interpreting effect sizes may be useful in further explaining the relationships under observation. First, verbal ability uniquely accounted for 2.89% of the variance in long word use (i.e., the squared correlation between Gc and long word use, see

Table 9), a small to medium effect size. When all other cognitive abilities for a person are average, a person's vocabulary knowledge may play a role in the proportion of longer, more complex words that they use in a cognitively demanding essay. Second, verbal ability uniquely accounted for only 0.36% of the variance in unique word use, a very small effect. Referring back to the observed correlation matrix in Table 8, verbal ability tests did correlate positively with total word count, which appears to be the inverse of unique word use to some extent (r = -.74, see Pennebaker et al., 2007). People with larger vocabularies may use more words in general in writing. As a person uses more words, the proportion of unique words that person uses likely decreases (i.e., he or she is more likely to repeat the same words again). This is especially evident with words like pronouns. A person may say a word once, then use a pronoun repeatedly to represent that same thing. Sometimes a person will use the same pronoun to represent different things, which will attenuate the proportion of unique words used. Third, short-term memory uniquely accounted for only 0.49% of the variance explained in conjunction use, a very small effect. Although conjunctions do tend to be in more complex sentences, there may be a ceiling effect of conjunction use on sentence complexity. For example, if a person uses 1 conjunction, that person has probably combined two thoughts, phrases, or ideas together. However, if that person wanted to combine three thoughts together, he or she could do so and still only use one conjunction by creating a comma-separated list. If a person used many conjunctions in the same sentence, the sentence would probably look and sound odd (i.e., a run-on sentence), where punctuation marks would be replaced with conjunctions (e.g., "I like apples and I like oranges and I like bananas and I like grapes" versus "I like apples, oranges, bananas, and grapes"). Fourth, short-term memory uniquely accounted for only 0.09% of the variance in preposition use, a very small effect. People with better short-term memory may not have used more prepositions in writing. Although theory supports the notion that short-term memory is connected to syntactical complexity (Kemper & Sumner, 2001), it is possible that preposition use may not be a strong indicator of syntactical complexity. Fifth, fluid reasoning uniquely accounted for 0.36% of the variance in cognitive process word use, a very small effect. Even if cognitive process words represent cognition, anyone can think and solve problems to some extent, regardless of their actual fluid reasoning level, which may explain why this relationship was so weak.

Additional unhypothesized relationships were explored among broad cognitive abilities and word count outcomes as part of Research Question 1. None of these exploratory paths reached statistical significance, but some of the unique effects of broad abilities on LIWC outcomes were noteworthy. First, latent Gc uniquely accounted for more variance in preposition use (3.61%) than did the hypothesized construct, latent Gsm (0.09%; i.e., the squared correlations presented in Table 9). Verbal ability is generally considered to be relevant to lexical complexity, but not to syntactical complexity according to the research literature. The partial correlation between verbal ability and preposition use was negative (β = -.19), suggesting that when all other cognitive abilities for a person are average, prepositions are either used less by people with higher verbal abilities, or that people with lower verbal ability tend to use more prepositions in cognitively demanding writing contexts. Second, latent Gc uniquely accounted for more variance in cognitive process word use than did the hypothesized construct, latent Gf (i.e., 4.00% vs. 0.36%). The partial correlation between verbal ability and cognitive process word use was negative (β = -.20), suggesting that when all other cognitive abilities for a person are average, cognitive process words are either used less by people with higher verbal abilities, or that people with lower verbal ability use more cognitive process words in cognitively demanding writing contexts. Fluid reasoning may be relevant to problem-solving, but it is possible that cognition involving problem-solving does not require the use of cognitive process words. The effect of verbal ability on cognitive process words may lie in the lexical complexity of the word category. Some of the words in the cognitive process category are simple (e.g., “all,” “if,” and “doubt”) whereas others are notably complex (e.g., “definitive,” “notwithstanding,” and “supposition”). The more complex words may be positively related to vocabulary knowledge, but are likely to be used less in general writing. LIWC does not note which words were used most often. It only counts how many words in the category were used and how that relates to the other words used in the essay. Thus, a person with lesser verbal ability could have used many cognitive process words in general without ever using one of the more complex words in the cognitive process word category.

In summary, GCA was reflected most strongly in the proportion of unique words used and the proportion of prepositions used in a cognitively demanding essay context beyond other broad cognitive abilities. The broad cognitive abilities examined did not incrementally predict LIWC word count outcomes beyond GCA, but some LIWC word outcomes did reflect Gc to some extent. Latent GCA and latent verbal ability accounted for the most variance in LIWC word count outcomes among all predictors examined. In cognitively demanding writing contexts, LIWC word count categories may be partially explained by GCA and verbal ability, but the proportion of unexplained variance remaining in each LIWC outcome category is very large and likely due to other factors as well such as personality (Yarkoni, 2010; Schwartz et al., 2013, Park et al., 2015) and the situational context. Thus, future research should investigate personality, situational contexts, and other constructs simultaneously with cognitive abilities to estimate the role of each in word count outcomes. Examining the incremental prediction of word count outcomes added by one type of construct over another (e.g., personality over cognitive ability and vice-versa) would be a fruitful next step in exploring the expression of psychological constructs through NLP.

Limitations

The largest limitation to the present study may be a lack of statistical power to find statistically significant effects. Statistical significance does not guarantee meaningful effects, but it does rule out the possibility of a type I error (i.e., the null hypothesis was wrongfully rejected). Without statistical power, the possibility of committing a type II error is greater (i.e., failing to reject the null hypothesis when it should have been rejected). Interpreting effect sizes may inform conclusions and future research, but all non-significant effects must be interpreted with caution. A larger sample size may have increased the number of statistically significant paths, ruling out the possibility of alternative hypothesis outcomes due to chance. The Monte Carlo simulation population estimates were larger than the sample estimates obtained, suggesting that a larger sample may have been necessary for detecting the smaller effects with statistical significance. Given the effect sizes found in the exploratory model (see Table 11) and a larger sample size, GCA might also have been expressed through conjunction use and cognitive process word use to some extent. Also, verbal ability might have been expressed through long word use as predicted, as well as through preposition use and cognitive process word use.

A second limitation to the present study lies with the motivation to perform on the essay writing prompts. The GRE essay prompts were realistic high-stakes essay prompts, but it is generally difficult to make the stakes feel high in an online research study. Participants may not have felt pressure to perform at a maximum level, only exerting enough effort to finish the task and receive payment. This limitation was combatted by advertising a $50.00 bonus to the five best essay writers. This should have increased the stakes to some degree, but it might not have worked for all participants. One participant posted online that he or she did not believe the bonus was real. This participant may not have been alone in these beliefs. To investigate these possibilities a bit further, a post-hoc analysis was conducted on motivation to perform using two motivation-related items from the demographic survey (see Appendix A). Participants reported being very motivated to write by the possibility of earning the bonus payment for good writing (M = 3.98 on a scale of 1 to 5, SD = 1.12). Participants were slightly less motivated to write when not considering the bonus (M = 3.82, SD = 1.03). Correlations between the two items and LIWC word count outcomes are presented in Table 17. Neither item was strongly correlated with any of the LIWC word count outcomes, although total word count was slightly positively correlated with motivation to obtain the bonus, suggesting that motivated participants tended to write longer essays to an extent. Given this analysis, motivation did not appear to affect results. However, motivation issues cannot be ruled out completely, as participants may have been motivated to report higher motivation to ensure payment for their work.

Table 17
Motivation to Perform and Correlates with Outcomes

Motivation	Bonus	General	Long	Unique	Conj	Prep	Cogproc	WC
Bonus	1	.44	-.02	-.09	.08	.00	-.06	.09
General	.44	1	.00	-.04	.10	-.03	.02	.01

Note: N = 373. Bonus = motivation by bonus payment. General = motivation to write not counting bonus payment. Long = longer word use, Unique = unique word use, Conj = Conjunction word use, Prep = preposition word use, Cogproc = cognitive process word use, WC = total word count.

A third limitation to the present study and online testing in general was that the protocol was weak to cheating. The essay prompt and cognitive tests were timed, which should have helped deter cheating to some extent. When tests are timed, test-takers do not have enough time to acquire every answer and still finish the test in time. Although it is unknown to what extent cheating may have occurred in the protocol, none of the participants' memory tests received perfect scores before dropping items that no one answered correctly. This evidence suggests that cheating did not occur on the memory test. Those tests were not timed and participants had the ability to replay the stimulus audio and video files repeatedly. Participants also could have taken paper or digital notes of the number and letter sequences in order to answer all items correctly. Although there was not much reason for participants to cheat on the cognitive tests, some participants were caught cheating on the essays. When two participants' responses were similar or identical in phrasing, the text in question was searched online. When the response or parts of the responses were found online, both participants' submissions were rejected in MTurk (see Table 1, Rules 7 & 8). Although several participants were caught plagiarizing, it is possible that other participants re-phrased others' work slightly before submitting. If a participant copied a response that no other participant copied, it was undetected. In the future, a software solution checking for online plagiarism (e.g., SafeAssign) may help prevent cheating in a higher-stakes assessment context.

Future Research Directions

Broadly, there is much potential for future research in the domain of NLP in assessment contexts. Many constructs have been studied to some extent, including personality, cognitive abilities, leadership skills, and communication skills (Park et al., 2015; Campion et al., 2016; Weaver, 2017). However, these constructs and more may be studied and assessed with more than one NLP methodology. For example, latent Dirichlet allocation (Blei et al., 2003), a type of topic analysis, may be useful in examining cognitive abilities in writing. It is currently unknown whether people higher in GCA talk or write about different topics than people lower in GCA. Latent Dirichlet allocation would enable the clustering of writing or speech samples into various topics, which could cause individual or class word use. Another alternative approach to NLP and assessment may involve latent semantic analysis (Landauer, Foltz, & Laham, 1998), where writing samples are scored based on how similar they are to a target text sample. Latent semantic analysis has been used to assess student knowledge in automatic essay scoring (Rehder et al., 1998), which could be applied in pre-employment assessments of job knowledge. Another approach might involve machine learning regression, where individual words or phrases could be examined as markers of GCA or broad abilities rather than broader categories in a closed-vocabulary approach such as LIWC. Other pre-packaged software and theoretical approaches such as Coh-Metrix (Graesser et al., 2004) may be useful for examining the expression of cognitive abilities as well. Coh-Metrix is able to assess additional aspects of sentence complexity, such as counting the proportion of subordinate, left-, and right-branching clauses in sentences in a text sample. It is currently unknown whether these markers would be tied to GCA or broad cognitive abilities in the same way as conjunctions and prepositions.

Additionally, future research should investigate other text sources, verifying the generalizability of this methodology to other writing or speech contexts. Weaver (2017) examined LIWC outcomes using resumes from an online panel. Campion and colleagues (2016) analyzed the accomplishment records written by real job applicants. Other sources of text may be more or less useful than these or the cognitively demanding essay context of the present study. Social media data, cover letters, biodata, and interview transcriptions are often readily available text sources which could be analyzed for additional information about applicant KSAOs. It is possible that different constructs may manifest themselves in different ways across contexts. For example, in a free response outlet such as a blog post or social media post, personality markers may be more readily available than in a specific writing prompt. Specific writing prompts such as essays or accomplishment records may be better at identifying markers of mental abilities, job knowledge, or skills. Future research should aid in the mapping of psychological constructs to the most amenable writing or speech contexts. It is clear from the LIWC2015 manual (Pennebaker et al., 2015) that word count categories do vary across different genres and outlets (e.g., blogs, newspapers, novels, and social media).

As the literature on NLP and assessment develops further, future research should also explore NLP in relation to other selection outcomes such as applicant reactions. Part of the appeal of NLP-based assessment is the increased efficiency of analysis without additional testing of the applicants. However, if applicants do not feel that NLP-based assessment is a face valid or fair method for assessing their KSAOs, NLP-based assessment may do more harm than good, as poor applicant reactions can lead to job offer rejections or possible litigation (Hausknecht et al., 2004).

Conclusion

In the present study, GCA was expressed through unique word use and preposition use. To a lesser extent, GCA was also expressed through conjunction use and cognitive process word use. Among broad cognitive abilities, verbal ability was expressed through long word use (i.e., words with seven or more characters), preposition use, and cognitive process word use to a small extent. Short-term memory and fluid reasoning were not expressed through word count categories. Although these findings are helpful for the exploration of cognitive ability expression in NLP through word counts, the theoretical justification for some of these exploratory findings (i.e., the expression of Gc through preposition and cognitive process word use) remains unclear. The zero-order correlations between several cognitive abilities and word count categories were negative. Thus, the findings of the present study did not totally align with the theory available in the research literature. New theory in this domain should focus on the LIWC word count categories not explored in the present study, which may provide other outlets for the expression of GCA and broad cognitive abilities.

Assessing cognitive abilities from word count categories in practice is not advised at this time. Although the data collected in the present study fit the proposed model, theory linking the expression of GCA and broad cognitive abilities in word counts of cognitively demanding essays needs to be refined. Other areas of NLP (e.g., latent Dirichlet allocation, latent semantic analysis) may be more useful in assessing cognitive abilities, but this is left to future research. Although applying the present findings to practice is not advised, it is too early to close off this research, especially considering some of the small to medium effect sizes. Future NLP-based assessment research should focus on the expression of GCA and verbal ability through language, as these constructs were more noticeably expressed through word count outcomes than short-term memory and fluid reasoning. Traditional assessment of cognitive abilities are probably still a better method for assessment than NLP-based methods, but NLP-scored assessments take much less time to analyze and score. Such assessments generally require less development effort by the test developer as well as less effort and time by participants (i.e., 15 minutes versus 90 minutes). With some refinement and on a large enough scale, NLP-based assessments could be useful for triangulating job applicant or employee GCA or possibly as an early pre-employment selection hurdle.

REFERENCES

Ackerman, P. L. & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121*(2), 219-245.

Alpers, G. W., Winzelberg, A. J., Classen, C., Roberts, H., Dev, P., Koopman, C., & Taylor, C. B. (2005). Evaluation of computerized text analysis in an internet breast cancer support group. *Computers in Human Behavior, 21*, 361-376. doi:10.1016/j.chb.2004.02.008

Armstrong, M. B. and Landers, R. N. (2017). An evaluation of gamified training: Using narrative to improve reactions and learning. *Simulation & Gaming, 48*(4), 513-538. doi:10.1177/1046878117703749

Attali, Y. & Burnstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment, 4*(3), 3-30.

Baddeley, A., Emslie, H., & Nimmo-Smith, I. (1993). The Spot-the-Word test: A robust estimate of verbal intelligence based on lexical decision. *British Journal of Clinical Psychology, 32*, 55-65.

Bentler, P. M. & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Boot, P., Zijlstra, H., & Geenen, R. (2017). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics, 6*(1), 65-76. doi:10.1075/dujal.6.1.04boo

Boyd, R. L., Wilson, S. R., Pennebaker, J. W., Kosinski, M., Stillwell, D. J., & Mihalcea, R. (2015). Values in words: Using language to evaluate and understand personal values. In

Proceedings of the International Conference on Weblogs and Social Media (pp. 31-40). Palo Alto, CA: AAAI Press.

Brown, K. G., Le, H., & Schmidt, F. K. (2006). Specific aptitude theory revisited: Is there incremental validity for training performance? *International Journal of Selection & Assessment, 14*(2), 87-100.

Bunderson, C. V. (1967). Transfer of mental abilities at different stages of practice in the solution of concept problems. *ETS Research Bulletin Series, 1967*(1), i-113. doi:10.1002/j.2333-8504.1967.tb00369.x

Cambria, E. & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, *9*(2), 48-57. doi:10.1109/MCI.2014.2307227

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *New Developments in Selection and Placement* (pp. 35-70). San Francisco, CA: Jossey-Bass.

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101*(7), 958-975. doi:10.1037/apl0000108

Carlisle, J. F. & Stone, C. A. (2005). Exploring the role of morphemes in word reading. *Reading Research Quarterly*, *40*(4), 428-449. doi:10.1598/RRQ.40.4.3

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

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, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*(1), 219-251. doi:10.1111/j.1467-6494.2011.00739.x

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, *41*(2), 189-225. doi:10.1207/s15327906mbr4102_5

Cormier, D. C., Bulut, O., McGrew, K. S., & Frison, J. (2016). The role of the Cattell-Horn-Carroll (CHC) cognitive abilities in predicting writing achievement during the school-age years. *Psychology in the Schools, 53*(8), 787-803. doi:10.1002/pits.21945

Crocker, L. M. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, & Winston.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*, 7-24. doi:10.1016/j.asw.2012.10.002

Dillon, W. R., Kumar, A., & Mulani, N. (1987). Offending estimates in covariance structure analysis: Comments on the causes of and solutions to Heywood cases. *Psychological Bulletin*, *101*(1), 126-135.

Dowling, G. R. & Kabanoff, B. (1996). Computer-aided content analysis: What do 240 advertising slogans have in common? *Marketing Letters, 7*(1), 63-75.

Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements. *Organizational Research Methods, 10*(1), 5-34. doi:10.1177/1094428106289252

Dunham, J. L. & Bunderson, C. V. (1969). Effect of decision-rule instruction upon the relationship of cognitive abilities to performance in multiple-category concept problem. *Journal of Educational Psychology*, *60*(2), 121-125.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Feldman, S. (1999). NLP meets the jabberwocky: Natural language processing in information retrieval. *Online*, *23*(2), 62-72.

Flower, L. & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition & Communication, 32*(4), 365-387.

Floyd, R. G., McGrew, K. S., & Evans, J. J. (2008). The relative contributions of the Cattell-Horn-Carroll cognitive abilities in explaining writing achievement during childhood and adolescence. *Psychology in the Schools, 45*(2), 132-144. doi:10.1002/pits.20284

Frey, M. C. & Detterman, D. K. (2004). Scholastic assessment or *g*? The relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science, 15*(6), 373-377.

Furnham, A. & Monsen, J. (2009). Personality traits and intelligence predict academic school grades. *Learning and Individual Differences, 19*, 28-33. doi:10.1016/j.lindif.2008.02.001

Gignac, G. E., Stough, C., & Loukomitis, S. (2004). Openness, intelligence, and self-report intelligence. *Intelligence, 32*, 133-143. doi:10.1016/j.intell.2003.10.005

Gottfredson, L. S. (1997a). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, *24*(1), 13-23.

Gottfredson, L. S. (1997b). Why *g* matters: The complexity of everyday life. *Intelligence, 24*(1), 79-132.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223-234. doi:10.3102/0013189X11413260

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, & Computers, 36*(2), 193-202.

Gustafsson, J.-E. & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28*(4), 407-434.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*, 639-683.

Hill, B. D., Foster, J. D., Elliott, E. M., Shelton, J. T., McCain, J., & Gouvier, W. D. (2013). Need for cognition is related to higher general intelligence, fluid intelligence, and crystallized intelligence, but not working memory. *Journal of Research in Personality, 47*, 22-25. doi:10.1016/j.jrp.2012.11.001

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Reseach Methods*, *6*(1), 53-60.

Horton, J. J., & Chilton, L. B. (2010, June). The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce* (pp. 209-218). New York, NY: ACM. doi:10.1145/1807342.1807376

Hough, L. M. (1984). Development and evaluation of the "accomplishment record" method of selecting and promoting professionals. *Journal of Applied Psychology*, *69*(1), 135-146.

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American Journal of Psychology*, *120*(2), 263-286.

Kemper, S. & Sumner, A. (2001). The structure of verbal abilities in young and older adults. *Psychology and Aging, 16*(2), 312-322. doi:10.1037//0882-7974.16.2.312

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2&3), 259-284.

Landers, R. N., Armstrong, M. B., & Collmus, A. C. (2017, April). Empirical Validation of a General Cognitive Ability Assessment Game. In R. N. Landers (Chair) and B. Hawkes (Discussant), *Assessment Games and Gamified Assessment*. Symposium presented at the 32nd Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Landers, R. N. & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8*(2), 142-164. doi:10.1017/iop.2015.13

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436-444. doi:10.1038/nature14539

Lemke, E. A., Klausmeier, H. J., & Harris, C. W. (1967). Relationship of selected cognitive abilities to concept attainment and information processing. *Journal of Educational Psychology*, *5*(1), 27-35.

Liddy, E. D. (1998). Enhanced text retrieval using natural language processing. *Bulletin of the American Society for Information Science, 24*(4), 14-16. doi:10.1002/bult.91

Marsh, H. W., Hau, K.-T., & Balla, J. R. (1998). Is more ever too much: The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*(2), 181-220. doi:10.1207/s15327906mbr3302_1

McCarthy, D. (1933). Language development. In C. Murchison (Ed.), *A handbook of child psychology* (2nd ed., Vol. 1, pp. 329-373). New York, NY: Russell & Russell/Athneum Publishers. doi:10.1037/11552-000

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary Intellectual Assessments: Theories Tests and Issues* (pp. 151-179). New York, NY: The Guilford Press.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (2nd ed., pp. 136-181). New York, NY: The Guilford Press.

McIntosh, C. N. (2007). Rethinking fit assessment in structural equation modeling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences, 42*, 859-867. doi:10.1016/j.paid.2006.09.020

Meade, A. W. & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437-455. doi:10.1037/a0028085

Miller, G. A., Newman, E. B., & Friedman, E. A. (1958). Length-frequency statistics for written English. *Information and Control, 1*, 370-389.

Mount, M. K., Oh, I.-S., & Burns, M. (2008). Incremental validity of perceptual speed and accuracy over general mental ability. *Personnel Psychology, 61*(1), 113-139.

Murray, A. L. & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, *41*, 407-422. doi:10.1016/j.intell.2013.06.004

Muthén, L. K. & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*(4), 599-620. doi:10.1207/S15328007SEM0904_8

Muthén, L. K. & Muthén, B. O. (2012). Mplus User's Guide (7th Ed.). Los Angeles, CA: Muthén & Muthén.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*(5), 665-675. doi:10.1177/0146167203251529

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Park, G., Schwartz, A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., . . . Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934-952. doi:10.1037/pspp0000020

Pennebaker, J. W. (1993). Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour Research and Therapy*, *31*(6), 539-548.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. Austin, TX: University of Texas at Austin.

Pennebaker, J. W. & Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition and Emotion, 10*(6), 601-626. doi:10.1080/026999396380079

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count: LIWC2001*. Mahwah, NJ: Lawrence Erlbaum Associated.

Pennebaker, J. W. & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296-1312.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*, 547-577. doi:10.1146/annurev.psych.54.101601.145041

Ramirez-Esparza, N., Pennebaker, J. W., Garcia, A. F., and Suria, R. 2007. La psicología del uso de las palabras: Un programa de computadora que analiza textos en espaniol [The psychology of word use: A computer program that analyzes texts in Spanish]. *Revista Mexicana de Psicologia 24*, 85-99.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48. doi:10.1006/cogp.1999.0735

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology, 79*(4), 518-524.

Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes, 25*(2&3), 337-354.

Rosenberg, S. & Abbeduto, L. (1987). Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics, 8*, 19-32.

Rosenberg, S. D., Schnurr, P. P., & Oxman, T. E. (1990). Content analysis: A comparison of manual and computerized systems. *Journal of Personality Assessment, 54*(1 & 2), 298-310.

Rude, S., Gortner, E.-M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion, 18*(8), 1121-1133. doi:10.1080/02699930441000030

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*(2), 262-274.

Schmidt, F. L. & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*(1), 162-173. doi:10.1037/0022-3514.86.1.162

Schneider, W. J. & McGrew, K. S. (2012). The Cattell-Horn-Carrol model of intelligence. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests and Issues* (3rd ed., pp. 99-144). New York: The Guilford Press.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., …Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE 8*(9): e73791. doi:10.1371/journal.pone.0073791

Sewell, W. H. & Shah, V. P. (1968). Parents' education and childrens' educational aspirations and achievements. *American Sociological Review, 33*(2), 191-209.

Short, J. C., McKenny, A. F., & Reid, S. W. (2018). More than words? Computer-aided text analysis in organizational behavior and psychology research. *Annual Review of Organizational Psychology and Organizational Behavior, 5*, 415-435. doi:10.1146/annurev-orgpsych-032117-104622

Spearman, C. (1970). *The abilities of man: Their nature and measurement*. New York, NY: AMS Press.

Stanek, K. C. & Ones, D. S. (2018). Taxonomies and compendia of cognitive ability and personality constructs and measures relevant to industrial, work and organizational psychology. In D. S. Ones, N. Anderson, C. Viswesvaran, & H. K. Singangil (Eds.), *The SAGE Handbook of Industrial, Work, and Organizational Psychology: Personnel Psychology and Employee Performance* (pp. 366-407). London, UK: SAGE Publications. doi:10.4135/9781473914940.n14

Tausczik, Y. R. & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24-54. doi:10.1177/0261927X09351676

Traub, R. E. (1970). A factor analysis of programmed learning and ability measures. *Canadian Journal of Behavioral Sciences*, *2*(1), 44-59.

Weaver, J. D. (2017). *Predicting employee performance using text data from resumes* (Doctoral dissertation). Seattle Pacific University. Retrieved from http://digitalcommons.spu.edu/iop_etd/9/

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality, 44*, 363-373. doi:10.1016/j.jrp.2010.04.001

Yuseph, R. L. & Vanderploeg, R. D. (2000). Spot-the-Word: A measure for estimating premorbid intellectual functioning. *Archives of Clinical Neuropsychology, 15*(4), 319-326.

Zhao, N., Jiao, D., Bai, S., & Zhu, T. (2016). Evaluating the validity of simplified Chinese version of LIWC in detecting psychological expressions in short texts on social network services. *PLoS ONE, 11*(6), 1-15. doi:10.1371/journal.pone.0157947

APPENDIX A

Full Item Lists for Each Measure

Writing Sample Task

Overview: On the next three pages, you will be presented with an issue. You will have 5 minutes to plan and compose a response to that issue before moving on to the second and third issues. There are no right or wrong answers for any of the issues, but your response should be written according to the specific instructions and support your position on the issue with reasons and examples drawn from areas such as your reading, experience, observations, and/or academic studies. A response to any other issue besides those given will receive a score of zero.

You may spend no more than 5 minutes per essay before you will be automatically directed to the next page. A timer will display how many minutes and seconds remain on each essay. **The "next page" button will not appear until after 1 minute has past since the start of each essay.**

You must write a minimum of 5 sentences per essay, but no more than 25 sentences at a maximum per essay **in order to receive any form of payment for this HIT.**

After the completion of this research study, the **five participants with the strongest and best written essays overall will be awarded a bonus of $50.00 each.**

You must enter your Amazon MTurk Worker ID at the end of this study survey along with your demographic information in order to qualify for the bonus.

Proceed to the next page to begin the first essay. The 5-minute timer will begin automatically.

Issue 1: As people rely more on technology to solve problems, the ability of humans to think for themselves will surely deteriorate.

Instructions: Write a response in which you discuss the extent to which you agree or disagree with the statement and explain your reasoning for the position you take. In developing and supporting your position, you should consider ways in which the statement might or might not hold true and explain how these considerations shape your position.

[essay text box]

Issue 2: To understand the most important characteristics of a society, one must study its major cities.

Instructions: Write a response in which you discuss the extent to which you agree or disagree with the statement and explain your reasoning for the position you take. In developing and supporting your position, you should consider ways in which the statement might or might not hold true and explain how these considerations shape your position.

[essay text box]

Issue 3: Scandals are useful because they focus our attention on problems in ways that no speaker or reformer ever could.

Instructions: Write a response in which you discuss the extent to which you agree or disagree with the claim. In developing and supporting your position, be sure to address the most compelling reasons and/or examples that could be used to challenge your position.

[essay text box]

Cognitive Ability Test Battery

1. Verbal Ability
    a. Extended Range Vocabulary Test
    b. Advanced Vocabulary Test I
    c. Advanced Vocabulary Test II
2. Short-Term Memory
    a. Auditory Number Span Test
    b. Visual Number Span Test
    c. Auditory Letter Span Test
3. Fluid Reasoning
    a. Letter Sets Test
    b. Locations Test
    c. Figure Classification Test
4. Visual Processing
    a. Form Board Test
    b. Paper Folding Test
    c. Surface Development Test

Verbal Ability (Gc)

EXTENDED RANGE VOCABULARY TEST -- V-3

This is a test of your knowledge of word meanings. Look at the sample below. One of the five numbered words has the same meaning or nearly the same meaning as the word above the numbered words. Mark your answer by putting an X through the number in front of the word that you select.

jovial

1-refreshing
2-scare
3-thickset
4-wise
X-jolly

The answer to the sample item is number 5; therefore, an X has been put through number 5.

Your score will be the number marked correctly minus a fraction of the number marked incorrectly. Therefore, it will <u>not</u> be to your advantage to guess unless you are able to eliminate one or more of the answer choices as wrong.

You will have <u>6 minutes</u> for each of the two parts of this test. Each part has one page. When you have finished Part 1, STOP. Please do not go on to Part 2 until you are asked to do so.

1. cottontail
   1-squirrel
   2-poplar
   3-boa
   4-marshy plant
   5-rabbit

2. marketable
   1-partisan
   2-jocular
   3-marriageable
   4-salable
   5-essential

3. boggy
   1-afraid
   2-false
   3-marshy
   4-dense
   5-black

4. gruesomeness
   1-blackness
   2-falseness
   3-vindictiveness
   4-drunkenness
   5-ghastliness

5. loathing
   1-diffidence
   2-laziness
   3-abhorrence
   4-cleverness
   5-comfort

6. bantam
   1-fowl
   2-ridicule
   3-cripple
   4-vegetable
   5-ensign

7. evoke
   1-wake up
   2-surrender
   3-reconnoiter
   4-transcend
   5-call forth

8. unobtrusive
   1-unintelligent
   2-epileptic
   3-illogical
   4-lineal
   5-modest

9. terrain
   1-ice cream
   2-final test
   3-tractor
   4-area of ground
   5-weight

10. capriciousness
    1-stubbornness
    2-courage
    3-whimsicality
    4-amazement
    5-greediness

11. maelstrom
    1-slander
    2-whirlpool
    3-enmity
    4-armor
    5-majolica

12. tentative
    1-critical
    2-conclusive
    3-authentic
    4-provisional
    5-apprehensive

13. placate
    1-rehabilitate
    2-plagiarize
    3-depredate
    4-apprise
    5-conciliate

14. surcease
    1-enlightenment
    2-cessation
    3-inattention
    4-censor
    5-substitution

15. apathetic
    1-wandering
    2-impassive
    3-hateful
    4-prophetic
    5-overflowing

16. paternoster
    1-paternalism
    2-patricide
    3-malediction
    4-benediction
    5-prayer

17. opalescence
    1-opulence
    2-senescence
    3-bankruptcy
    4-iridescence
    5-assiduity

18. lush
    1-stupid
    2-luxurious
    3-hazy
    4-putrid
    5-languishing

19. curtailment
    1-expenditure
    2-abandonment
    3-abridgment
    4-improvement
    5-forgery

20. perversity
    1-adversity
    2-perviousness
    3-travesty
    4-waywardness
    5-gentility

21. calumnious
    1-complimentary
    2-analogous
    3-slanderous
    4-tempestuous
    5-magnanimous

22. illiberality
    1-bigotry
    2-imbecility
    3-illegibility
    4-cautery
    5-immaturity

23. clabber
    1-rejoice
    2-gossip
    3-curdle
    4-crow
    5-hobble

24. sedulousness
    1-diligence
    2-credulousness
    3-seduction
    4-perilousness
    5-frankness

Careless Responding Item: Happy

1 – Sad; 2 – Angry; 3 – Afraid; **4 – Joyful**; 5 – Disgusted

ADVANCED VOCABULARY TEST I — V-4

This is a test of your knowledge of word meanings. Look at the sample below. One of the five numbered words has the same meaning or nearly the same meaning as the word above the numbered words. Mark your answer by putting an X through the number in front of the word that you select.

jovial

1-refreshing
2-scare
3-thickset
4-wise
~~5~~-jolly

The answer to the sample item is number 5; therefore, an X has been put through number 5.

Your score will be the number marked correctly minus a fraction of the number marked incorrectly. Therefore, it will <u>not</u> be to your advantage to guess unless you are able to eliminate one or more of the answer choices as wrong.

You will have <u>4 minutes</u> for each of the two parts of this test. Each part has one page. When you have finished Part 1, STOP. Please do not go on to Part 2 until you are asked to do so.

1. mumble
   1-speak indistinctly
   2-complain
   3-handle awkwardly
   4-fall over something
   5-tear apart

2. perspire
   1-struggle
   2-sweat
   3-happen
   4-penetrate
   5-submit

3. gush
   1-giggle
   2-spout
   3-sprinkle
   4-hurry
   5-cry

4. massive
   1-strong and muscular
   2-thickly populated
   3-ugly and awkward
   4-huge and solid
   5-everlasting

5. feign
   1-pretend
   2-prefer
   3-wear
   4-be cautious
   5-surrender

6. unwary
   1-unusual
   2-deserted
   3-incautious
   4-sudden
   5-tireless

7. veer
   1-change direction
   2-hesitate
   3-catch sight of
   4-cover with a thin layer
   5-slide

8. orthodox
   1-conventional
   2-straight
   3-surgical
   4-right-angled
   5-religious

9. stripling
   1-stream
   2-narrow path
   3-engraving
   4-lad
   5-beginner

10. salubrious
   1-mirthful
   2-indecent
   3-salty
   4-mournful
   5-healthful

11. limpid
   1-lazy
   2-crippled
   3-clear
   4-hot
   5-slippery

12. procreate
   1-sketch
   2-inhabit
   3-imitate
   4-beget
   5-encourage

13. replete
   1-full
   2-elderly
   3-resentful
   4-discredited
   5-restful

14. frieze
   1-fringe of curls on the forehead
   2-statue
   3-ornamental band
   4-embroidery
   5-sherbet

15. treacle
   1-sewing machine
   2-framework
   3-leak
   4-apple butter
   5-molasses

16. ignominious
   1-inflammable
   2-elflike
   3-unintelligent
   4-disgraceful
   5-mysterious

17. abjure
   1-make certain
   2-arrest
   3-renounce
   4-abuse
   5-lose

18. duress
   1-period of time
   2-distaste
   3-courage
   4-hardness
   5-compulsion

ADVANCED VOCABULARY TEST II — V-5

This is a test of your knowledge of word meanings. Look at the sample below. One of the four numbered words has the same meaning or nearly the same meaning as the word at the left. Indicate your answer by writing, in the parentheses at the right, the number of the word that you select.

attempt 1-run 2-hate 3-try 4-stop . . . . . ( )

The answer to the item is number 3; you should have a "3" written in the parentheses.

Your score will be the number marked correctly minus a fraction of the number marked incorrectly. Therefore, it will not be to your advantage to guess unless you are able to eliminate one or more of the answer choices as wrong.

You will have 4 minutes for each of the two parts of this test. Each part has one page. When you have finished Part 1, STOP. Please do not go on to Part 2 until you are asked to do so.

1. rancor	1-forbearance 2-ridicule 3-malice 4-bravery	( )
2. raucous	1-empty 2-quiet 3-smooth 4-harsh	( )
3. gargoyle	1-oil 2-medicine 3-carved waterspout 4-ugly building	( )
4. recrudesce	1-purify 2-renew activity 3-lack refinement 4-crush	( )
5. specious	1-plausible, but not genuine 2-noteworthy 3-class or variety 4-roomy	( )
6. bauble	1-bubble 2-showy plaything 3-idle talk 4-confusion	( )
7. prolific	1-scarce 2-producing abundantly 3-reckless 4-speaking profanely	( )
8. opulent	1-party 2-wealthy 3-happy frame of mind 4-semiprecious stone	( )
9. coercion	1-conspiracy 2-strategy 3-restraint 4-attraction .	( )
10. hiatus	1-animal 2-calamity 3-dread 4-gap	( )
11. germane	1-microbe 2-contagious 3-relevant 4-different . . .	( )
12. perfunctory	1-fundamental 2-formal 3-superficial 4-careful . .	( )
13. diverge	1-reveal 2-chant 3-distract the attention of 4-differ or turn off from	( )
14. evoke	1-take away 2-anger 3-connect 4-bring out	( )
15. pertinent	1-relevant 2-lying next to 3-necessary 4-bold . . .	( )
16. holocaust	1-entirety 2-destruction 3-saintly 4-price	( )
17. piquant	1-mellow 2-fish 3-pungent 4-cloth	( )
18. firmament	1-foundation 2-heavens 3-strong 4-glue	( )

Careless Responding Item: Mother

1 – Uncle; 2 – Aunt; **3 – Mom**; 4 – Dad

Short-Term Memory (Gsm)

AUDITORY NUMBER SPAN TEST — MS-1

This is a test of your ability to remember series of numbers. The examiner will call out the numbers. After he finishes, you are to write down the numbers in the exact order in which they were called out. Please do not write any numbers until the examiner has finished the whole series.

Some of the series will be too long for you to remember all of the numbers. If you do not remember some of them, leave a blank space for them and write down all the numbers you do remember. Try to remember all the numbers if possible, and be sure to write them down in the exact order in which they were called out.

For example, the examiner might call out, "Series One. 7 2 4 Begin."

When he says "Begin" (showing that the series is complete), write the numbers on the answer page in this manner:

1. 7 2 4

It is very important that you do not write numbers while a series is being called out, since this is a test of your memory for numbers.

Your score on this test will be the number of series you remember correctly.

Auditory Number Span Test -- MS-1

1. 8, 1, 9, 5, 7, 2
2. 4, 6, 2, 9
3. 3, 7, 1, 4. 9, 2, 5, 8, 1, 6
4. 9, 2, 6, 2, 8, 6
5. 7, 9, 5, 3, 8
6. 5, 2, 9, 4, 1, 6, 8, 3, 7
7. 2, 6, 3, 1, 5
8. 2, 4, 8, 5, 1
9. 6, 8, 2, 4, 1, 3, 9, 7, 2, 5, 3
10. 9, 2, 8, 5, 7, 1
11. 7, 4, 2, 9, 3, 5, 8, 6
12. 4, 3, 7, 2, 3, 9
13. 5, 7, 3, 1 , 6, 9, 4, 8, 5, 1, 7, 2
14. 6, 2, 5, 9, 7, 1, 8, 3
15. 4, 7, 9, 3, 6, 1, 5, 8, 4, 2, 7
16. 5, 1, 8, 7, 2, 3, 1
17. 8, 2, 6, 9, 1, 7, 3, 8, 5, 9, 6, 4
18. 5, 1, 9, 2, 7, 4, 8, 3, 6
19. 7, 5, 2, 6, 4, 9, 1
20. 3, 2, 1, 8, 1, 4, 6, 5
21. 3, 8, 1, 7
22. 9, 4, 6, 3, 5, 1, 8
23. 2, 8, 4, 9, 6, 2, 7, 5, 9, 3, 1
24. 6, 2, 8, 7, 2, 3, 6, 1

VISUAL NUMBER SPAN TEST — MS-2

This is a test of your ability to remember series of numbers. For each series the examiner will show you several numbers on cards, one after the other. After he finishes, you are to write down the numbers in the <u>exact order</u> in which they were shown to you. Please do <u>not</u> write any numbers until the examiner has finished the whole series.

Some of the series will be too long for you to remember all of the numbers. If you do not remember some of them, leave a blank space for them and put down all the numbers you do remember. Try to remember <u>all</u> the numbers if possible, and be sure to write them down so that they will be in the <u>exact</u> order in which they were written on the cards.

For example, the examiner might show you a card like this:

$\boxed{7}$, then $\boxed{2}$, then $\boxed{4}$, and then say "Begin."

When he says "Begin" (showing that the series is complete), write the numbers on the answer page in this manner:

1. 7 2 4

<u>It is very important that you do not write numbers while a series is being shown to you, because this is a test of your memory of numbers.</u>

Your score on this test will be the number of series you remember correctly.

Visual Number Span Test -- MS-2

1. 4, 1, 5, 2, 3, 7
2. 7, 2, 5, 6
3. 8, 9, 6, 1, 3, 7, 2, 4, 5, 7
4. 1, 2, 5, 2, 7, 4
5. 2, 1, 6, 8, 5
6. 1, 2, 4, 9, 5, 6, 3, 8, 7
7. 2, 6, 5, 3, 4
8. 8, 5, 1, 2, 4
9. 4, 3, 7, 2, 1, 6, 8, 9, 7, 2, 5, 3
10. 9, 1, 8, 6, 4, 3
11. 5, 4, 8, 6, 7, 1, 3, 2
12. 9, 3, 2, 5, 3, 6
13. 5, 3, 9, 7, 1, 6, 8, 4, 2, 5, 7, 3, 2
14. 3, 6, 5, 4, 1, 9, 2, 8
15. 6, 1, 5, 8, 4, 9, 7, 3, 6, 2, 1
16. 3, 4, 7, 1, 8, 9, 5, 4
17. 1, 4, 7, 3, 5, 2, 8, 1, 9, 3, 7, 6
18. 3, 2, 8, 5, 6, 4, 7, 9, 1
19. 5, 1, 3, 2, 7, 6, 4
20. 2, 6, 5, 1, 3, 2, 7, 4
21. 2, 5, 1, 6
22. 8, 5, 1, 3, 7, 4, 2
23. 3, 6, 1, 9, 8, 5, 2, 7, 9, 4, 3
24. 8, 9, 7, 4, 2, 6, 3, 9

Careless Responding Item: 3, 4, 5

AUDITORY LETTER SPAN TEST — MS-3

This is a test of your ability to remember series of letters. The examiner will call out the letters. After he finishes, you are to write down the letters in the exact order in which they were called out. Please do not write any letters until the examiner has finished the whole series.

Some of the series will be too long for you to remember all of the letters. If you do not remember some of them, leave a blank space for them and write down all the letters you do remember. Try to remember all the letters if possible, and be sure to write them down in the exact order in which they were called out.

For example, the examiner might call out, "Series One. H R L Begin."

When he says "Begin" (showing that the series is complete), write the letters on the answer page in this manner:

1. H R L

Only the following letters will be used: C, F, G, H, K, L, P, R, S, W, Y.

It is very important that you do not write letters while a series is being called out, because this is a test of your memory for letters.

Your score on this test will be the number of series you remember correctly.

Auditory Letter Span Test -- MS-3

1. K, F, C
2. H, S, L, Y, G
3. P, F, R, C, W, S, G, K, Y
4. P, L, S, C, W, K, R, F, H, G
5. R, G, S
6. L, W, C, Y, K, R, P
7. F, S, Y, L, C, H
8. S, C, F, K, W, L, P
9. Y, C, G, P, W, L, S, K, H, R, F
10. W, Y, S, C, L
11. P, G, L, F, H, K, Y, W, C, R
12. C, F, G, W, K, S, R, L, P
13. H, Y, R, W, S, P, K
14. S, F, L, H
15. C, P, R, K, H, S, W, F
16. S, F, W, G, P, L, K, H
17. K, W, G, Y, L, R
18. F, P, S, G, L, H, W, K, R, C
19. G, R, H, P, C, S, F, Y, W
20. Y, C, W, S, P, R, F
21. R, P, F, K
22. L, G, K, S, Y, C, R, F, W, H, P
23. C, S, P, G, R, Y, H, L
24. F, C, G, H, L, P, S, K

Fluid Reasoning (Gf)

LETTER SETS TEST -- I-1 (Rev.)

Each problem in this test has five sets of letters with four letters in each set. Four of the sets of letters are alike in some way. You are to find the rule that makes these four sets alike. The fifth letter set is different from them and will not fit this rule. Draw an X through the set of letters that is different.

NOTE: The rules will not be based on the sounds of sets of letters, the shapes of letters, or whether letter combinations form words or parts of words.

Examples:

A.	NOPQ	~~DEFL~~	ABCD	HIJK	UVWX
B.	NLIK	PLIK	QLIK	~~THIK~~	VLIK

In Example A, four of the sets have letters in alphabetical order. An X has therefore been drawn through DEFL. In Example B, four of the sets contain the letter L. Therefore, an X has been drawn through THIK.

Your score on this test will be the number of problems marked correctly minus a fraction of the number marked incorrectly. Therefore, it will not be to your advantage to guess unless you are able to eliminate one or more of the letter sets.

You will be allowed 7 minutes for each of the two parts of this test. Each part has 1 page. When you have finished Part 1, STOP. Please do not go on to Part 2 until you are asked to do so.

1.	QPPQ	HGHH	TTTU	DDDE	MLMM
2.	BCDE	FGHI	JKLM	PRST	VWXY
3.	BVZC	FVZG	JVZK	PWXQ	SVZT
4.	BCEF	FGIJ	STWX	CDFG	PQST
5.	BCCB	GFFG	LMML	QRRQ	WXXW
6.	AAPP	CCRR	QQBB	EETT	DDSS
7.	ABDC	EGFH	IJLK	OPRQ	UVXW
8.	CERT	KMTV	FHXZ	BODQ	HJPR
9.	PABQ	SEFT	VIJW	COPD	FUZG
10.	CFCR	JCVC	CGCS	CLXC	KCWC
11.	XDBK	TNLL	VEGV	PFCC	ZAGZ
12.	CAEZ	CEIZ	CIOZ	CGVZ	CAUZ
13.	VEBT	XGDV	ZIFX	KXVH	MZXJ
14.	AFBG	EJFK	GKHM	PSQT	RWSX
15.	KGDB	DFIM	KIFB	HJMQ	LHEC

Careless Responding Item:

NNNN	NNNN	NNNN	**MMMM**	NNNN

LOCATIONS TEST — I-2

Each problem in this test consists of five rows of small dashes separated into groups by blank spaces. In each of the first four rows one dash is replaced by an "x". In the fifth row five of the dashes are replaced by numbers. In each problem there is a rule guiding the placement of the "x" in each of the first four rows. You are to figure out what that rule is and to use the rule in deciding where the "x" should come in row 5. When you have picked the number in row 5 which appears where the "x" belongs, draw an X through it.

<u>Example A</u>:

```
Row 1   --------------- --x -----
Row 2   ----- --x-- ------------
Row 3   ------------- --x ------
Row 4   ----- --x---- ----------
Row 5   --------1-- --X--3-4 5---
```

Example A has been correctly marked. In the first four rows the "x" always replaces the <u>third</u> dash from the left of a group. The group is always the <u>second</u> group in the row. Therefore the correct answer is 2 because the number 2 replaces the <u>third</u> dash of the <u>second</u> group in row 5.

<u>Example B</u>:

```
-------- x------ --------
-- --------- -- x-- ----
x------- ---------------
----------- x------ -----
1-------2 3---- 4-- -5---
```

In the first four rows of example B the "x" replaces the <u>first</u> dash in a group. The group with the "x" is always the <u>next to last</u> group in the row. Therefore the correct answer is 4, since the number 4 replaces the <u>first</u> dash in the <u>next to last</u> group in row 5.

You should expect to find any kind of relation or rule to explain the position of the x's.

Your score on this test will be the number marked correctly minus a fraction of the number marked incorrectly. Therefore, it will <u>not</u> be to your advantage to guess unless you are able to eliminate one or more of the answer choices as wrong.

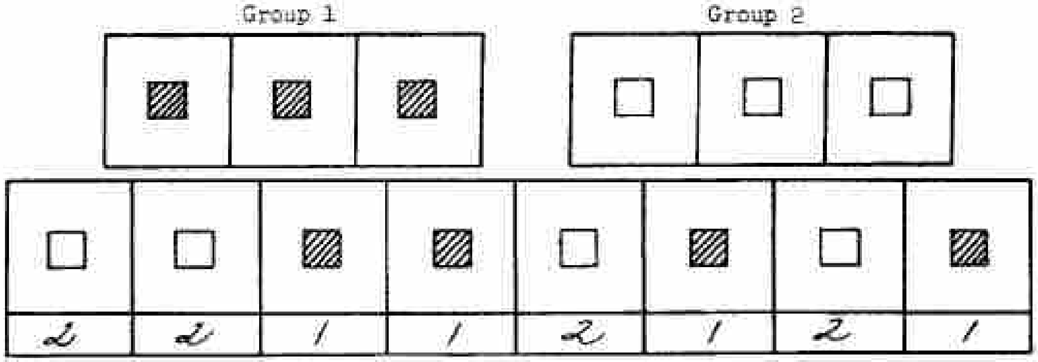
You will have <u>6 minutes</u> for each of the two parts of this test. Each part has one page with 14 items. When you have finished Part 1, STOP. Please do not go on to Part 2 until you are asked to do so.

```
1.  ------- x------- ---- --
    ---- -x--- -- --- ------
    --------------- --x-----
    -------- ---x----------
    ----1 2---3-- 4---5-----

2.  --- ---- x--- ----------
    ------ x--- ------ -----
    --------- -------- x- --
    ----- x--- ------- -----
    1-- 2--- 3----- 4-------5

3.  --x- ---- ---- ---- ----
    ---- -x-- ---- ---- ----
    ---- ---- x--- ---- ----
    ---- ---- --x ---- ----
    --1- -2-- 3--- --4- ---5

4.  ----------- -- -x-------
    -- -x-------------------
    --------- ------- -x----
    --- -------- -- -x------
    -1---- 23- -4 -5--------

5.  ------- -x- ------ -----
    -x- ------ --------- ---
    ----- ----- -x- --------
    --- ------ --------- -x-
    --1 -2- ---3 - -4- -5---

6.  ---- ----- ---x--- -- --
    -x- --------- --- --- --
    --------- --x-- --------
    -- - ---- ----x---- ----
    --1-- --2-- -34 --5-- --

7.  ----- --- ----- ---x ---
    -----------x -----------
    ----- ------ -------x --
    ------ ----x -----------
    ---1---2 ----34 ----5 --

8.  ------------ - x--------
    - x------ ---- ---------
    ------- ----- - x-------
    --- - x----- ---- ------
    1-- - 2-----3 4-- -5----

9.  ---- --- x ---- --- - --
    ---- - ---- --- x ------
    ------ -------- --- x --
    -- ------ --- x - ------
    ---1 ---2 --3 4 --- 5 --

10. --- ----x-- ------- ----
    -------x-- -- ----- ----
    - --- ---x -------- ----
    -- -- -- -x ----- ------
    -- - 12-34- -5- --- ----

11. ----- x-- -------- -----
    -- x----- ---------- ---
    ------- x-------- ------
    --- x--- --- --- --- ---
    1-2 3----- 4-------5 ----

12. --x ---- -------- ------
    ----x---- ------- ------
    -- - -- -x---- -- ------
    ---------x--- ------ ----
    --1 -- 2 --3 - 4--5-----

13. x---- --------- -- -- --
    ------- x-- -------- ---
    --- --- x--------- -----
    ---- ------ ----- x-----
    1-- 2-- 3-- 4------ 5---

14. ------ --- --- ---x-----
    ---x------ ------- -----
    --- ------- ---x----- --
    -- - ---x---------------
    -1- ---2---3 ---4---5---
```

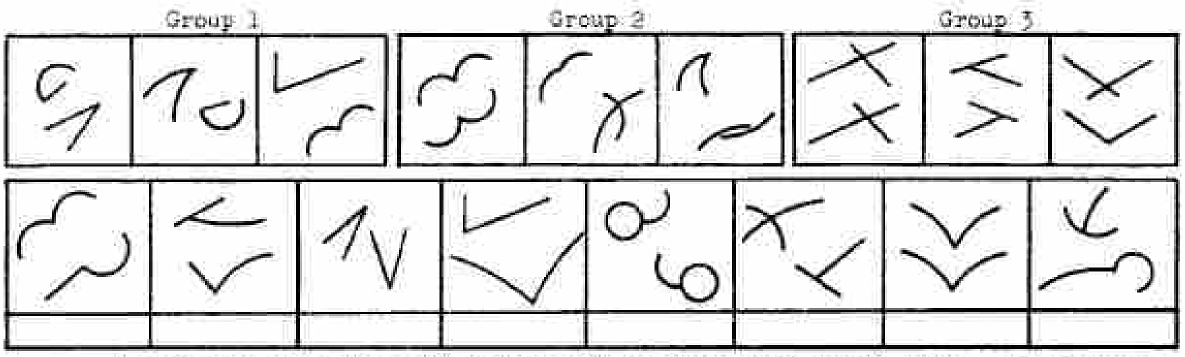
FIGURE CLASSIFICATION — I-3

This is a test of your ability to discover rules that explain things. In each problem on this test there are either two or three groups, each consisting of three figures. You are to look for something that is the same about the three figures in any one group and for things that make the groups different from one another.

Now look at the sample problem below. In the first line, the figures are divided into Group 1 and Group 2. The squares in Group 1 are shaded and the squares in Group 2 are not shaded. In the second line a 1 has been written under each figure that has a shaded square as in Group 1. A 2 has been written under each figure with an unshaded square as in Group 2.



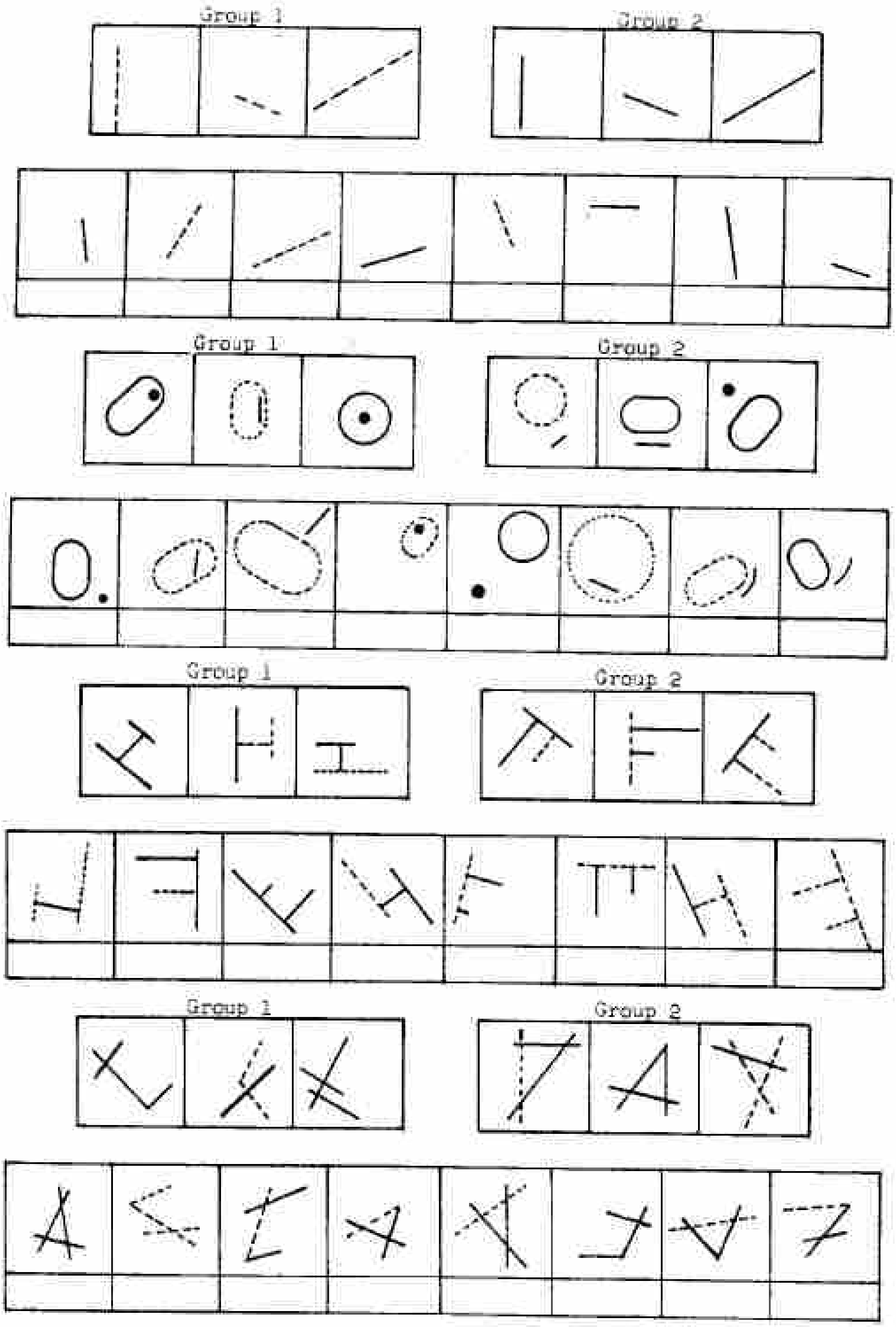
Now try this more difficult sample, which has three groups:

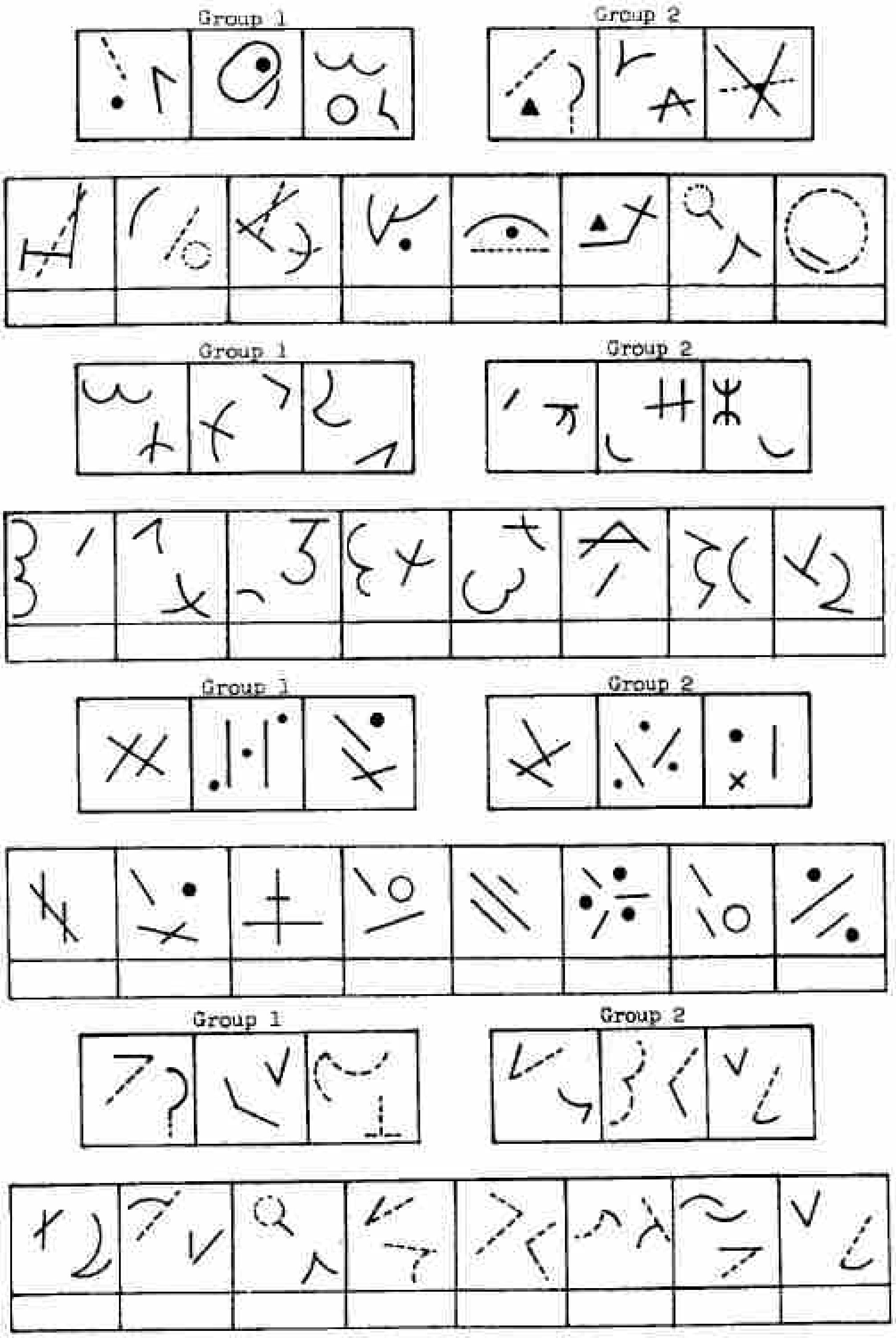


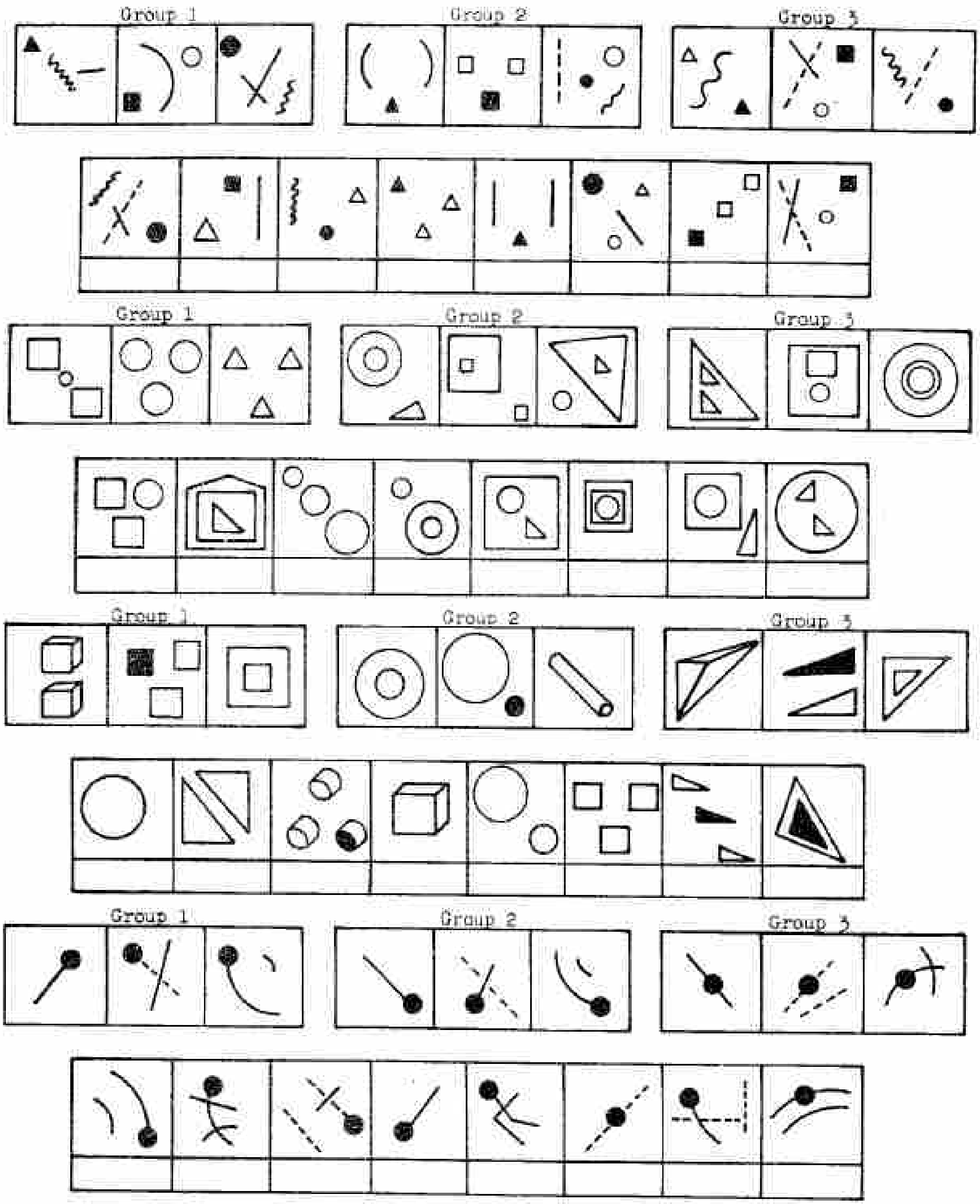
The figures in Group 1 consist of both straight and curved lines. The figures in Group 2 consist of curved lines only. The figures in Group 3 consist of straight lines only. As you can see, there are other details that have nothing to do with the rule. The answers are: 1, 1, 3, 1, 2, 1, 2, 2.

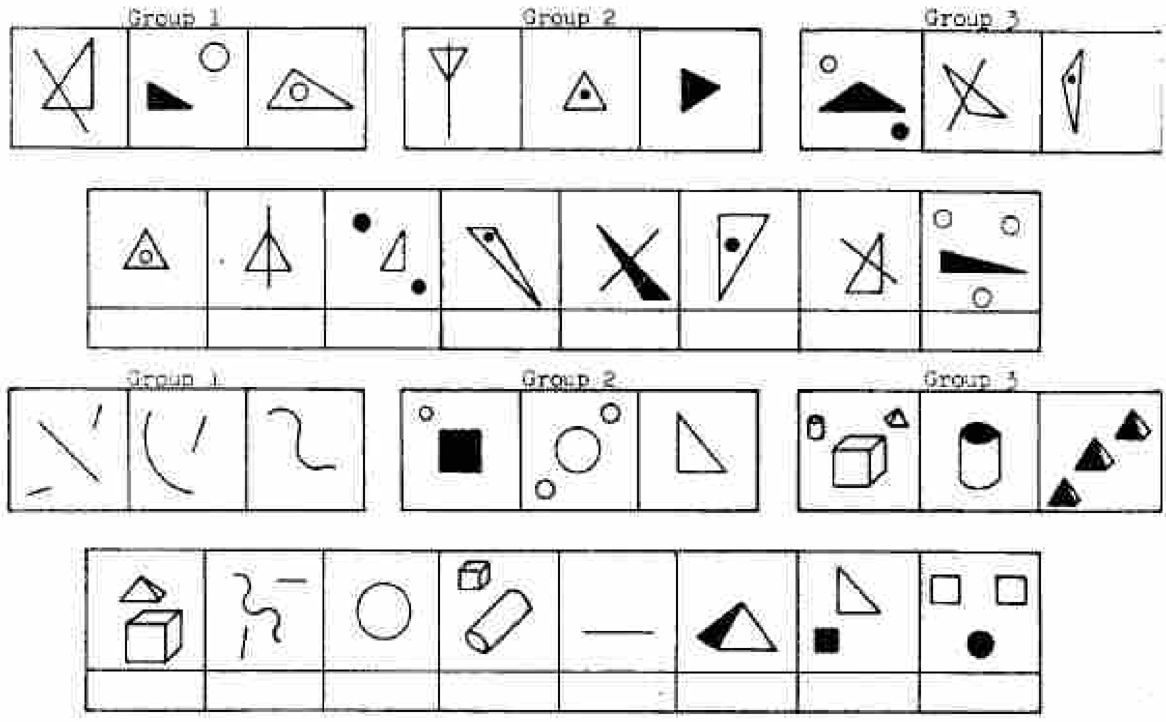
Your score on this test will be the number of figures identified correctly minus a fraction of the number marked incorrectly. Therefore, it will not be to your advantage to guess unless you have some idea of the group to which the figure belongs.

You will have 8 minutes for each of the two parts of this test. Each part has 4 pages. When you have finished Part 1, STOP. Please do not go on to Part 2 until you are asked to do so.


Group 1
Group 2
Group 1
Group 2
Group 1
Group 2
Group 1
Group 2


Group 1
Group 2
Group 1
Group 2
Group 1
Group 2
Group 1
Group 2


Group 1
Group 2
Group 3
Group 1
Group 2
Group 3
Group 1
Group 2
Group 3
Group 1
Group 2
Group 3


Group 1
Group 2
Group 3
Group 1
Group 2
Group 3

Demographic and Careless Responding Questionnaire

- What is your age?
  - [drop down menu with ages 18-64, under 18, and 65+]
- What is your gender?
  - Male
  - Female
  - Other [blank]
- What is your ethnicity?
  - Hispanic
  - Non-Hispanic
- What is your race?
  - African American or Black
  - Asian American
  - Caucasian or White
  - Native American or Native Alaskan
  - Pacific Islander or Native Hawaiian
  - Other single race [blank]
  - Two or more races (select all that apply)
  - Not American
- What is your native language?
  - Arabic
  - Bengali
  - English
  - Hindi
  - Japanese
  - Mandarin
  - Portuguese
  - Punjabi
  - Spanish
  - Russian
  - Other [blank]
- Besides Amazon MTurk, are you currently employed?
  - Yes, full time
  - Yes, part time
  - No
- If yes to the above:
  - How long have you held this job? (years, months)
  - In what type of business are you employed?
    - Business Services
    - Education
    - Finance
    - Health Care
    - Insurance
    - Manufacturing
    - Retail

        - Wholesale
        - Other [blank]
    - On average, how many hours do you work each week? (dropdown menu with choices 1-79, 80+ hours)
- Have you earned the Amazon MTurk "Master Worker" certification (i.e., are you a Master Worker)?
    - No
    - Yes
- I gave an honest effort at all of these assessments.
    - False
    - True
- In all honesty, you should not use my data for research purposes because I did not respond completely honestly or to the best of my ability.
    - You should use my data.
    - You should not use my data.
- The five workers with the best written essays overall will receive $50.00 bonuses each. On a scale of 1 (not at all motivating) to 5 (extremely motivating), how motivating was this bonus for you when writing your essays?
    - 1 – Not at all motivating
    - 2 – Slightly motivating
    - 3 – Moderately motivating
    - 4 – Very motivating
    - 5 – Extremely motivating
- Without considering the bonus, and with the same scale as above, how motivated were you when writing your essays in general?
    - 1 – Not at all motivated
    - 2 – Slightly motivated
    - 3 – Moderately motivated
    - 4 – Very motivated
    - 5 – Extremely motivated
- Copy and paste your Amazon MTurk Worker ID number (located in the top left of the Amazon MTurk website while logged in as a worker) here to ensure you receive payment and to be entered in the contest for the bonus $50.00 for best written essays: [blank]

VITA

Michael Beaumont Armstrong
Old Dominion University
Department of Psychology
Mills Godwin Building, Room 250
Norfolk, VA 23529

Education

Bachelor of Arts, Psychology, Western Kentucky University, May 2013
Master of Science, Psychology, Old Dominion University, December 2015
Concentration: Industrial/Organizational Psychology

Selected Publications

Armstrong, M. B. & Landers, R. N. (2018). Gamification of employee training and development. *International Journal of Training & Development, 22*(2). doi:10.1111/ijtd.12124

Armstrong, M. B. & Landers, R. N. (2017). An evaluation of gamified training. Using narrative to improve reactions and learning. *Simulation & Gaming, 48*(4), 513-538. doi:1046878117703749

Armstrong, M. B., Ferrell, J. Z., Collmus, A. B., & Landers, R. N. (2016). Correcting misconceptions about gamification of assessment: More than SJTs and Badges. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 9*(3), 671-677.

Landers, R. N. & Armstrong, M. B. (2017). Enhancing instructional outcomes with gamification: An empirical test of the technology-enhanced training effectiveness model. *Computers in Human Behavior, 71*, 499-507. doi:10.1016/j.chb.2015.07.031

Selected Presentations

Armstrong, M. B. (Co-Chair) & Landers, R. N. (Co-Chair). (2018, April). *Using natural language processing to measure psychological constructs.* Symposium presented at the 33rd Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Poeppelman, T. & Armstrong, M. B. (2018, April). Learning technology and gamification. In E. Sinar (Chair), *Paving two-way streets: Digital dyads to drive technology research and practice*. Alternative session presented at the 33rd Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Armstrong, M. B. (Co-Chair), Sanchez, D. R. (Co-Chair), Bauer, K. N. (Co-Chair), & Kraiger, K. (Discussant). *Gaming and Gamification IGNITE: Current Trends in Research and Application.* Symposium presented at the 32nd Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Armstrong, M. B. (Co-Chair) & Landers, R. N. (Co-Chair). (2015, April). *Game-thinking in assessment: Applications of gamification and serious games*. Symposium presented at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, Philadelphia, PA.