Fall 1-31-2014

Using latent semantic analysis to detect non-cognitive variables of academic performance

Daniel Richard Aalderks
New Jersey Institute of Technology

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

USING LATENT SEMANTIC ANALYSIS TO DETECT NON-COGNITIVE VARIABLES OF ACADEMIC PERFORMANCE

by
Daniel Richard Aalderks

This thesis explores the possibilities of using latent semantic analysis to detect evidence of intrapersonal personality variables in post-secondary student essays. Determining student achievement based on non-cognitive variables is a complex process. Automated essay scoring tools are already in use today in grading and evaluating student texts based on cognitive domain traits, but at this time are not utilized to analyze non-cognitive domains such as personality. Could such tools be configured to detect non-cognitive variables in student essays? Key concepts in this proposal—personality traits, latent semantic analysis, automated essay evaluation, and online cinema reviews—are explored followed by a literature review to justify the research. As a proof of concept study, 43 writing samples written to a constructed response task are collected and analyzed by a test model specifically designed to evaluate sentiment in a movie review constructed response format. A test model is created using LightSIDE, a software tool for text assessment, to predict the sentiment of these essays with highly encouraging results. The thesis concludes with a path for future research in the largely unexplored area of automated assessment of non-cognitive variables.

USING LATENT SEMANTIC ANALYSIS TO DETECT NON-COGNITIVE VARIABLES OF ACADEMIC PERFORMANCE

by
Daniel Richard Aalderks

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Professional and Technical Communication

Department of Humanities

January 2014

APPROVAL PAGE

USING LATENT SEMANTIC ANALYSIS TO DETECT NON-COGNITIVE VARIABLES OF ACADEMIC PERFORMANCE

Daniel Richard Aalderks

Dr. Norbert Elliot, Thesis Advisor — Date
Professor of English, NJIT

Dr. Philip Klobucar, Committee Member — Date
Associate Professor of English, NJIT

Dr. Judith Redling, Committee Member — Date
Associate Provost for Academic Affairs**,** NJIT

Mr. Keith Williams, Committee Member — Date
University Lecturer, Information Systems Department, NJIT

BIOGRAPHICAL SKETCH

Author: Daniel Richard Aalderks

Degree: Master of Science

Date: January 2014

Undergraduate and Graduate Education:

- Master of Science in Professional and Technical Communication
  New Jersey Institute of Technology, Newark, NJ, 2014

- Bachelor of Science Computer Science
  Concordia College, Moorhead, MN, 2005

Majors: Professional and Technical Communication

To Mom and Dad,

All the stress and sleepless nights are because you said I could do it.

You were right.

ACKNOWLEDGMENT

This thesis would not have been possible without help and guidance of a number of people who have graciously contributed their time and expertise to not just completing this project, but completing it well.

First and foremost, my gratitude to Dr. Norbert Elliot, Professor of English at NJIT and my thesis advisor for seeing me through this rewarding journey. Developing this idea from initial inception to its current form has all been because of his guidance, patience, and skill in both the subjects explored and research mentored.

I would also like to thanks all the members of my advisory team Dr. Philip Klobucar, Dr. Judith Redling, and Mr. Keith Williams for lending their time and expertise.

Thanks as well to Dr. Irvin Peckham and Mr. Elijah Mayfield, who contributed time and resources to making this thesis possible.

Finally, I would like to thank Kristi Gandee, Bradley Landesm, and Penny Hanke in helping me review this project.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 Objective

The objective of this thesis is to provide evidence of the validity of using latent semantic analysis in conjunction with machine learning applications for detecting evidence of non-cognitive personality variables in post-secondary student essays. Specific attention will be paid to the Big Five personality factors; openness, conscientiousness, extraversion, agreeableness, and neuroticism, a broad set of domains used to describe human personality.

The supporting evidence of this objective will be presented in two methods. The first is in a literature review encompassing the Big Five factors of personality, automated essay evaluation software applications, latent semantic analysis, and online cinema reviews. By presenting contemporary research, it will be shown that there is a valid basis for the kind of research present in this thesis. The second set of evidence is in a proof of concept demonstration showing that the tools exist for performing further research on this topic and how they might be utilized to perform a more thorough research study in detecting Big Five personality factors. After this evidence has been presented and discussed, a direction for further study will be proposed as a follow-up to the research performed for this thesis. Based on the review of research and proof of concept demonstration, thesis proposal for further study will establish a program of research in this new area of writing studies.

1.2 Background Information

According to the National Center for Education Statistics, the ten colleges or universities with the highest enrollment in the United States averaged 86,465 students in 2010. If forty admissions officers spent only 5 minutes on each student's enrollment application, it would take four and a half business weeks to completely work through these applications. The example illustrates how much time and resources is required for processing the applications of prospective post-secondary students.

Methods have been developed to help sift through the large number of applications institutes like these handle on an annual basis. Various technological applications can be used to help identify prospective college students. Information systems are now designed to let students apply online and submit information such as academic records, test scores, and basic background information to help streamline the admissions process. As part of this process, admissions advisors look for measurements of cognitive variables (intelligence, aptitude, and memory) in applications based on high school grades and standardized test scores. Most of these types of criterion measures can be easily scored and processed for the admissions process.

These variables, however, present certain problems for identifying prospective students for enrollment. According to Sedlacek, there are a number of problems with strictly using cognitive variables in admissions (Why). For instance, tests such as the GRE are used only within the population for which it is intended, so the variance of test scores is limited. He states that the limits of what language and mathematical tests can tell when evaluating graduate level students have been reached. Sedlacek also points to studies conducted by the College Entrance Examination Board (CEEB) documenting

instances of grade inflation at all levels of education, which only further serves to make admissions decisions rather unclear for applicant evaluation (Why 2-4).

Cognitive variables, while certainly an important aspect of academic ability, are unable to represent the full spectrum of domains an academic institution needs to evaluate a potential student. To help fill the additional needed information that cognitive variables do not provide, non-cognitive variables have become valuable units of measurement for determining desirable applicants. Non-cognitive variables are used to understand non-traditional experiences, which can vary widely across a population. Some examples of non-cognitive variables include self-concept and appraisal, goals, leadership, and sense of community.

Research has been conducted about the advantages of using non-cognitive variables and cognitive variables for evaluating potential students[1]. Using non-cognitive variables aids in better evaluating multi-cultural students (Sedlacek, Why 4-6). Certain variables have also been shown to provide an accurate assessment of future academic success (O'Connor and Paunonen). Non-cognitive variables have also been found to increase the ability to identify students who will be retained by the institute they enroll in (Jackson and Strattner). This last point is a great advantage for post-secondary institutions as such assessment allows institutes to target students most likely to enroll and can also affect their enrollment versus graduation rates (Tracey).

Non-cognitive variables are difficult to infer from standard academic records, and utilizing them in the admissions process presents its own challenges. It can be difficult

[1] This identification of such domains is reflected in references in O'Connor and Paunonen; Sedlacek, "Employing Noncognitive Variables in the Admission and Retention of Nontraditional Students"; Gosling, Rentfrow, and Swann Jr.

and inefficient to determine how much or how little of any specific variable an applicant may possess with any degree of certainty. What makes one person's goal more indicative of a good college mindset than others? This is not a variable that can be easily deduced from an application or from test scores; interviews are not a practical solution on the scale that would be required for academic admissions. Turning to personality tests such as the Minnesota Multiphasic Personality Inventory (MMPI) and its subsequent incarnations can also present problems. The growing use of such tests has resulted in subjects becoming test wise, or being able to select an answer they feel will present them in the most acceptable way.

1.3 Proposal

It is not the intention of this research to devise a method for turning over all non-cognitive evaluation to machine learning applications. Instead, it is to propose a method that might work in support of current evaluations. Gallos notes how two different sources of data supporting the same hypothesis may be required to sustain the validity of psychological research in the face of multidimensional issues to be found in personality tests (425). Automated essay evaluation may act in such a supporting role in providing a second source of information. Since short answer tests like the MMPI personality test can result in fake responses, perhaps another format for measuring character would result in a profile provided by the student in which the student writes in some detail on attitudes toward selected subjects. From Stricker, it's known that longer responses such as role-playing produced more honest and less test wise reactions to personality prompts (13-16).

Latent semantic analysis (LSA), a technique of natural language processing, may present a possible supporting alternative. With the proper setup, LSA has been shown to

extract the semantic orientation and attitude of a text (Taboada et al.) and present limited imitation of human memory and processing (Dumais). It has even been shown to acquire information at a rate comparable to that of children (208-209). Could a process such as LSA be configured to automatically detect the presence and potency of a non-cognitive variable?

This thesis proposes that latent semantic analysis techniques in conjunction with machine learning applications are capable of detecting evidence non-cognitive traits such as the Big Five factors of personality. The research conducted was carried out under the broad assumptions of the following basic null hypotheses:

> $H_{0:}$ Machine learning applications and LSA algorithms *will not* be able to predict a positive or negative semantic response from students' writing reviews.
>
> $H_{1:}$ Machine learning applications and LSA algorithms *will* be able to predict a positive or negative semantic response from students' writing reviews.

While there are many more analyses to be undertaken regarding validation, this basic NHST test suggests that a combination of a literature review and a test case will be useful to determine if further study is feasible. If possible, this process will present several advantages within and beyond the field of education. The most obvious advantage is a more efficient and standardized method for identifying prospective enrollments for colleges and universities beyond current applications. Such a process will not be able to completely replace the scrutiny of trained staff, but it could act in a supporting role for verifying applications. The second possible advantage this system may provide is an expansion of the knowledge base in the fields of psychology and education. Are certain non-cognitive traits presented more robustly when produced in an extended written format? Which factors of personality are easier to detect in written responses? Here is an

exciting application of natural language processing that, at the present time, has seen very little exploration.

CHAPTER 2

KEY CONCEPTS

The following chapter provides information about key concepts in this thesis. These topics include the Big Five personality factors, latent semantic analysis, automated essay evaluation, and constructed response tasks.

2.1 Big Five Factors of Personality

The following section will provide a basic level of understanding of cognitive and non-cognitive variables, followed by information about the Big Five personality traits—a pivotal subject of this research.

Pellegrino and Hilton define cognition as "…types of knowledge and how they are structured in an individual's mind, including the processes that govern perception, learning, memory, and human performance (73)." These processes are different from emotional or volitional processes. Extensive research has been done on these mental processes and many people have taken aptitude tests in school to measure cognition. The SAT, GRE, and other admission tests are examples of tests seeking to analyze students' cognitive skills. Pellegrino and Hilton explain how the Committee on Defining Deeper Learning in the 21st Century Skills, in an effort to better define the skills required for deeper learning and its relationship to clusters of competency, identified three broad domains of competence; cognitive, intrapersonal, and interpersonal. Each of these domains contains a number of competencies, which in turn translate into various skills.

- Cognitive: reasoning and memory. Primarily dealing with information processing, the cognitive domain is the system that determines the flow of information, how it is gathered, and then how it is stored. Words and pictures are processed through the eyes and ears using sensory memory, then selected words and images are organized through working memory and is then integrated into long term memory which can then be pulled later for the additional processing of working memory procedure. The cognitive domain also includes problems solving skills employing methods like hill climbing, means-end, or trial and error to overcome obstacles. The timing and quality of information feedback also affects the acquisition of knowledge and skills. Ineffective feedback can hamper practice methods, motivation, and opportunities to correct mistakes. Competencies: cognitive process and strategies, knowledge, and innovation. Skills: skill acquisition, critical thinking, innovation, information literacy, and reasoning. (73-82).

- Intrapersonal: managing ones behavior and emotions. Simple beliefs about learning can greatly impact the learning process itself. If a student believes they do not have talent in a particular subject, they might be less inclined to overcome it. These preconceived notions can enhance or hamper the learning process without the skills to manage these assumptions. Metacognition, or awareness about how one thinks, has been identified as an important skill for experts. Monitoring ones own understanding and reacting to it can enhance memory performance. Self-regulation and setting and pursuing goals despite challenges are linked to the intrapersonal domain of conscientiousness. The level of self-regulation in an individual has been shown to aid or hinder in such milestones as graduating from high school. Competencies: intellectual openness, work ethic and conscientiousness, and positive self-evaluation. Skills: flexibility, personal and social responsibility, self-direction, perseverance, self-evaluation, and physical and mental health. (88-95)

- Interpersonal: expressing ideas and interpreting and responding to others. This domain is less defined than others, but Pellegrino and Hilton see it as the learning associated with unique social situations and various communities. This is a skillset primarily developed through interaction and discourse with others through participation. These participation skills are very important for an interactive learning processes and communication development. Competencies: teamwork and collaboration and leadership. Skills: communication, cooperation, interpersonal skills, empathy, responsibility, assertiveness, and social influence. (95-97)

The three domains above include cognitive functions that Pellegrino and Hilton state are malleable (25-26). They are subject to forces such as how much effort is put forth, motivation, and intrapersonal competencies. These competencies are the ability to

meet complex demands drawing upon various psychological resources (23). Research has thus focused on identifying these various traits that persist throughout a person's life. From this research, identifying traits have been identified; and while there is discussion about the exact combination of traits that are most influential, it is generally agreed that a set of latent indicator variables are influential in the learning process and are termed reflective latent variables because their ability to reflect other traits based upon correlation with other indicator variables (25-26).

These reflective latent characteristics, such as non-cognitive variables or specifically the Big Five factors of personality are examples of the intrapersonal domain of Pellegrino and Hilton. The Big Five contain a number of shared characteristics between personality traits and the domains and competencies above. While the domains are referenced in this thesis, focus will be on the intrapersonal domain.

Non-cognitive variables, according to Sedlacek, are everything cognitive traits are not. Where cognitive variables are used to measure information processing, non-cognitive variables are used to represent personality aspects such leadership, self-confidence, community service, field knowledge (unusual or cultural ways of acquiring knowledge) and self-appraisal (Sedlacek). Universities have begun making use of non-cognitive variables in student admissions as a way of better sorting applicants. Personality tests have become one such tool for measuring non-cognitive traits, alongside interviews, personal history evaluations, and other academic records.

The Big Five factors of personality are examples of non-cognitive variables. They have proven to be indicative of several areas of lifestyle and development. Oliver notes that certain combinations of Big Five personality traits are indicators of risk for

subsequent maladjustment in adolescents as well as juvenile delinquency, childhood psychopathology, and academic performance (Oliver 35; O'Connor and Paunen). The conscientiousness trait has been shown to be a proven indicator of job performance (35-36). Oliver does concede than defining personality in such a method may be more simplistic that it really is and alludes to several studies on the matter; however, he does state that the Big Five personality traits have also stimulated research into personality testing that has improved the level of knowledge available on the subject (36-37).

Such personality research has led to the creation of the Big Five. In fact, McCrae details these personality factors into a hierarchal model of five basic domains. These domains have been distilled by psychologists from decades of analysis of natural language terms people used to describe themselves. Systems for describing personality had existed prior to the creation of the Big Five, but one advantage to this taxonomy is that it can serve as an integrative function for representing those other systems by putting them into a common framework (Oliver 5).

The Big Five are made up of the personality traits listed below and are described using the same outline as Raad describes. O'Connor and Paunonen have further tested these facets of personality as to how they can be used to deduce future academic success. Further defining characteristics are assigned to each factor. These non-cognitive variables are generally recognized by the field of psychology as an accurate representative interpretation of personality. They also have proven to have a relative ease in being applied to testing procedures relevant to psychological procedures.

- Openness: inventive/curious vs. consistent/cautious. This trait reflects 'open-mindedness and an interest in culture. People who rate high in this Big Five trait tend to be imaginative, creative, and to seek out cultural and educational

experiences. People who rate lower are more down-to-earth, less interested in art, and more practical in nature.

- Conscientiousness: efficient/organized vs. easy-going/careless. This trait reflects how organized and persistent people are in pursuing our goals. Those who are high in this trait are methodical, well organized and dutiful. Low scorers are less careful, less focused, and more likely to be distracted from their goals.

- Extraversion: outgoing/energetic vs. solitary/reserved. This trait reflects preference for, and behavior in, social situations. People who rate high in extraversion are energetic and seek out the company of others while people at the other end of the scale tend to be more quiet and reserved.

- Agreeableness: friendly/compassionate vs. cold/unkind. This trait reflects how people tend to interact with others. People high in agreeableness tend to trust, friendly and cooperative. Lower scorers tend to be more aggressive and less cooperative socially.

- Neuroticism: sensitive/nervous vs. secure/confident. This trait reflects the tendency to experience negative thoughts and feelings. People more prone to neuroticism are insecure and emotionally distressed. Those found to be lower in this trait are more relaxed, less emotional, and less prone to distress.

But can personality really be concentrated into only five overall traits? In 1981 in a symposium in Honolulu, four prominent researchers—Goldberg, Takamoto-Chock, Comrey, and Digman—reviewed the personality tests and research available at the time and determined that most of the tests available which held any promise in gauging personality seemed to measure a subset of five common personality factors (Big Five). These five factors were formalized on this testing basis and became the Big Five personality factors, or simply Big Five. Srivastava considers the arrival of these five personality factors a simple extension of the Lexical Hypothesis (9), which Goldberg agrees with (26), establishing that the personality characteristics that are the most important in peoples' lives will eventually become a natural part of their language and that the more important characteristics are more likely to be contained within a single

word. Ashton and Lee express a similar view. They provide evidence from previous researchers that support a view that those personality markers can be represented as personality-descriptive terms used in communicative language (7-9). The application of words and how they are structured leads to common groupings for describing language that are encompassed by the domain of the personality traits laid out by the Big Five.

This is not to say that the debate about adding personality traits to the Big Five is not an ongoing process, as McCrae and Costas as well as Perrigrino and Hilton's work indicates. As an example, honesty-humility has been proposed as an addition to these five personality factors (Ashton, Lee, and Son). Variations of the Big Five that propose additional underlying forces of personality have been proposed such as Cattell's 16 Factor Model, Eysenck's Big Three Factors of psychoticism, extraversion, and neurotisism, and the Big Six in which the honesty-humility trait was proposed in addition to the Big Five (Linden, Nijenhuis, and Bakker). Other researchers have explored the possibilities of adding other categories to the Big Five; however, most of the proposed traits are already built into other personality frameworks outside the Big Five. Other examples of possible additions to the Big Five can be found in O'Connor and Paunonen, as well as Oliver.

A hierarchy division of personality such as the Big Five is not without its problems. Dimitri et al., as well as Paunonen and Jackson, detail how the Big Five personality traits are not necessarily independent of each other or might be related to a higher order personality trait. While Dimitri showed that there existed some correlation between personality factors, it was largely negligible within the bounds of their study. Paunonen and Jackson analyzed several previous studies from Saucier and Goldberg and

argued that based on the criterion presented in the studies they analyzed, some of the Big Five personality traits are not strictly orthogonal in nature. What this translates into for the proposed research is that there might be some overlap between personality factors that LSA is unable to account for.

While use of natural language processing techniques such as LSA is still in the early stages of research, personality tests specifically designed to measure the prevalence of personality traits are well established. For example, there are a number of basic online personality tests—some more validated than others—that are designed to test the level of the five factors of personality[2]. These tests ask the participant to answer forty to eighty questions by using a 5-point Likert scale, true and false, and some simple answer. Probably the most well known Big Five personality test is the Revised NEO Personality Inventory developed by Paul Costa and Robert McCrae. This test, also called the NEO PI-R, is a 240-point test that measures the Big Five personality traits as well as six subordinate facets of each trait. The abbreviated NEO Five-Factor Inventory Form S personality test has been shown to be successful in determining factors of personality and has also shown that these testing results can point toward academic success (Conrad).

There is criticism of reliance on personality tests such a NEO and that inherent flaws accompany these tests. Donaldson and Grant-Vallone argue that outside factors can influence the motivation of a person to accurately answer these tests. One such influence is a tendency for self-reporters to be biased. People tend to want to answer the questions in ways that make them appear to have advantageous qualities either because of their belief that they truly act in such a manner or out of fear of reprisal from an employer or

[2] Roberts lists two such tests on his biography website, one for conscientiousness and a second for narcissism. Buchanan of the University of Westminster also maintains an active Big Five personality test.

administrator (247). Another concern Donaldson and Grant-Vallone present is that almost all personality tests share the same methods of testing. Any significant findings may then be sullied by shared method variance problems (247-248). However, they acknowledge that this last issue is still a hotly debated topic.

2.2 Latent Semantic Analysis

Automated essay scoring technology is currently used in various assessment and evaluation tasks in numerous educational instructions across the country. Of the two most widely used systems for automated essay analysis are in use today, E-rater uses natural language processing while Intelligent Essay Assessor (IEA) using latent semantic analysis (Burstein 2). Both systems are used for assessment of the cognitive domain of writing, but the principles behind the systems also hold potential for non-cognitive assessment. Both systems have their advantages, but this thesis will be focusing on latent semantic analysis as the engine driving this research.

Latent semantic analysis is a computational analysis algorithm derived from natural language processing—a field of computer science, artificial intelligence, and linguistics—to analyze relationships between sets of documents, paragraphs, sentences, and words. The process finds the average uses for each word, sentence, and paragraph it processes to determine relationships between words and presents these findings in a quantitative representation of a semantic domain. This data can be measured in sentences, paragraphs, or pages and the data points created from these computations are used to determine associations or semantic similarities between word-word, word-passage, and passage-passage matrixes being built. In this way, LSA is able to determine relationships and the meaning of words based on how the data is used and how the data around it is

used. The semantic similarity of any word is determined by the resemblance of the words around it and how those words are being used in a similar context. It might help to think of LSA finding the average meaning and usage of the words and passages in the text it is given to analyze. Landauer et al. point out that LSA can be viewed as both a model of the underlying representation of knowledge and its acquisition or as a practical method for estimating aspects of similarities in meaning.

The mathematics behind LSA, vector space modeling, also called vector algebra, allows for representing the text in data being analyzed as identifiers based on the data's contextual usage. The ultimate goal of these calculations is to reduce the data to matrices that are then used to compare different sections of the text with adjoining units to determine if there is a semantic relationship. The larger the size of the communication LSA has to process, the more space the vector mathematics has to grow and form an understanding of the text it is analyzing. Remember that space is the size of the matrix generated by the mathematical process. The more words, sentences, and paragraphs that are analyzed, the larger the matrix will be and the greater the need for space. After a large enough corpus set of data has been processed, the similarity and usage of the words in the sample data can be analyzed to determine the relationships between word usage and meaning based on the scores assigned by the LSA methodologies (Landauer, Foltz, and Laham 3-4). A further explanation of the mathematics behind LSA can be found in the APPENDIX.

When it was originally patented, LSA was designed as an improvement to lexical, patent, and keyword matches, such as what a user would encounter when using a search engine on the Internet for example. LSA would also allow for a more efficient data

retrieval method in patent searches (Deerwester et al.). Users do not generally take into account the possible synonymy and polysemy that can be found in a simple keyword search. A user typically performs an information search based on word meaning but they do not always use the best word to express the information they are looking for. To demonstrate the variability of the differential in word meaning, it has been shown that two people, searching for a same well-known topic, will only use the same keyword approximately 20% of the time (Deerwester et al.). Dumais uses the example that a person looking for a document on the human-computer interaction will not find any meaningful responses using only the phrase man-machine studies or human factors in their searches (215). These are some of the original problems that LSA was created to resolve.

While improving database search performance shows that LSA can act as an enhancement to the human/machine interface, LSA has also been shown to interpret word meaning in a manner similar to a human mind. LSA applications are capable of interpreting the meaning of words as demonstrated by actually taking vocabulary tests such as the *Test of English as a Foreign Language*, or TOEFL. Dumais references a test conducted by herself and Landauer in 1996 and 1997 to compare the results of word interpretation by human beings and LSA. In this test, a latent semantic process was used to analyze over five million words from *Grolier's Academic American Encyclopedia;* after completing this analysis, the process was used to perform the TOEFL. The LSA software tool then interpreted the meaning derived from the words in the dictionary and compared them to the multiple-choice selection available in order to make its own selection using a similar process to human test takers. LSA's performance on the TOEFL

proved to have a 64 percent accuracy rating, which was the same as students who were also taking the TOEFL at the time (Dumais).

In order to better understand what LSA is capable of doing, a real world example may be required to more clearly demonstrate the concepts of this thesis. Dittmer and Parr examined of specific media sources to determine if US newspapers legitimized or undermined the sovereignty claims of Kosovo and South Ossetia during their respective conflicts. This is based on the narrative of the coverage each conflict received by US media sources (124-125). Using LSA, Dittmer and Parr were able to successfully determine what overriding themes were used in western media coverage during the reportage of each conflict. They also determined that when certain themes such politics or casualties were the primary subject of individual news articles, other opposing themes such as refuges and aid where typically not cited in conjunction with this main theme.

The researchers used the following LSA test to determine if a news bias did in fact exist. First, they used LSA to measure the semantic relation between the significant pieces of text of over one thousand different articles pulled from the Lexus Nexus database about each conflict. They then put them in three datasets: articles covering primarily Kosovo, Ossetia, and both parties. After the values of each set were indexed by LSA, a Person correlation was used to associate words and articles that then exposed common wordings and phrases from the media sources. From this process, highlighted themes from each data set emerged based on the relative strength of the terms and relations found by LSA. For example, in the dataset containing the entire corpus of media articles, positively skewed articles reporting on themes involving children, villages, and families were shown to be less likely to also report on negotiating or political and

military forces in the same article (133-136). The researchers concluded that the Kosovo conflict was narrated as a humanitarian intervention incident, while the Russian intervention in South Ossetia was narrated by western news as an imperialist intervention (124, and 138-139).

There are limitations to using LSA for processes such as essay grading for cognitive domains. There is no information for a clearly defined size of the data corpus used by LSA. Also, what kind of text should be used in a corpus? Experts suggest for many studies that a minimum of 500 samples be collected for a data corpus with double or even triple that being closer to ideal (Mayfield Interview). When setting up a control corpus of data, it is best to populate with as many examples as possible that represent comparable cross section to what it will be analyzing. LSA also has trouble with certain syntax. Weimer-Hastings points out that LSA does not take word order into account when reviewing text (8). LSA does well with longer strings of text and even single word responses (Landaur and Dumais), but it has trouble with small, short sentences. Finally, as a possible consequence of dropping stop words from the data corpus, LSA can have trouble with negation (Weimer- Hastings 9). Words such as non, no, or doesn't can be dropped by LSA as supposedly not important to the data corpus. Such are issues that need to be kept in mind while performing an LSA based research study.

2.3 Automated Essay Scoring of Cognitive Domains

With the application of standardized testing—set to move away from the traditional pencil and paper format to computer-based methods in 2014 (Tkacik)—a corresponding rise for computer-based solutions has occurred. Collaborative learning techniques, such as virtual environments, computer-supported collaborative learning, and learning

management systems are now available to enhance the learning and teaching experience. These are all relatively new paradigms in the academic community. Understanding the design of AEE in the cognitive domain allows identification of the potential and challenges of applications in the non-cognitive domain.

Automated essay evaluation (AEE) is a process for using computer programs to evaluate and score written text. AEE is a complex topic that is most evident when one realizes that it incorporates a wide range of fields such as applied linguistics, psychometrics and psychology, computer and information science, educational measurement, businesses administration and management, and rhetoric and writing studies (Shermis, Burstein, and Bursky; Elliot and Klobucar). Just like there are a number of disciplines that are poised for research in into AEE, an equally good number of applications beyond essay scoring come from it. Some examples of these applications include the following: determining a summative assessment of the development of learners at determined intervals from evaluated essays (Rich, Schneider, and D'Brot); monitoring reader performance and possibly detecting reader drift in essays (Lottridge, Schulz, and Mitzel); and establishing components for grammatical error detection and evaluation to improve language usage, grammar, and mechanics in succeeding essays (Gamon et al.). Again, these applications of AEE technology are only used in evaluation of cognitive traits. The research conducted for this thesis indicates that the use of AEE applications in identifying intrapersonal traits is limited to research investigation only (Burstein et al., Automated).

Automated essay evaluating and scoring applications have already been implemented in widespread academic programs. AEE technology has already been

applied in evaluating the Graduate Record Examination (GRE), the Test of English as a Foreign Language (TOEFL), and Graduate Management Admissions Test (GMAT); with further funding expanding the range of these tools today (Elliot and Klobukar 19). Applications of AEE have also been promoted in the state of West Virginia as a tool for teachers to perform various writing assessment evaluations (Chanhua). Finally, organizations such as ETS have emerged that are capable of scoring and evaluating over fifty million tests annually (ETS Fastfacts).

If applications of AEE are to be successful in evaluating essays, then they must have a sufficient corpus of text to draw upon for comparison—a demand that is also true for non-cognitive assessment. Simply put, AEE tools compare unevaluated text with text that has been appraised as having characteristics that are defining of the criteria being used to judge other corpuses of text. For example, if an evaluation tool such as LightSIDE ™, IntelliMetric ™, or E-rater ™ is grading a set of 12th grade English essays on a traditional A, B, C, D, F system, and then it must use a corpus of text to compare them to that has already been evaluated by human scorers and assigned these letter grades. If the evaluation of an essay more closely matches the traits found in the B set of text, then the AEE tool will assign it this grade. Such a system necessitates a sufficient corpus of control text to draw upon, but more importantly, this text must accurately represent what makes the scoring or evaluation system being applied. If the data used as the scoring model is faulty, then faulty results will be returned when other essays are evaluated.

Another potentially large application for AEE tools comes from the recent Common Core State Standards Initiative (CCSSI) that seeks to develop a common set of

standards in language arts and literacy and mathematics at each grade school level. Hakuta proposes that natural language processing, and subsequently AEE applications, stand to advance heavily from this policy by carving out a niche in this standardization process by "…being able to flag words and features that signal logical argumentation, sentiment, and other features of the text related to argumentation (349)." This policy shift in the standardization of education, and the subsequent understanding of the nature of language itself, offers many opportunities for AEE and natural language process experts to add their expertise to this debate and shape how education is evaluated.

This is not to say that AEE applications will soon be found in every academic institute in the country. The reaction to AEE has ranged from encouraging to distrust or worse from various groups. Page first predicted the use of AEE applications being performed within academic grading in 1966 and was generally met with skepticism for this statement (Shermis, Burstein, and Bursky 6). Yet there are currently eight commercial vendors of AEE applications and one open source entry from Carnegie Mellon University currently available today (Reich). Various studies have also been performed that validate the application of AEE in a controlled setting and in a working production environment such as the vendors mentioned earlier. Despite this, evidence of distrust can be found in places such as an online petition (humanreaders.org/petition/) that was launched on March 12, 2013. As of September 16 of the same year, this petition has accumulated 4087 signatures. Yang et al. also notes that an "overreliance on surface features of responses, the insensitivity to the content of responses and to creativity, and the vulnerability to new types of cheating and test-taking strategies" while reviewing a framework for validating computer-automated scoring methods in 2002 (393). Another

source of distrust comes from writing professionals, such as the Conference on College Composition and Communication which is quite vocal in its criticism of AEE practices (Attali 181).

Despite these stances on AEE, there are similarities between automated essay evaluation and human essay evaluation. As Williamson points out, both methods rely on the evaluators (human and machine) receiving appropriate levels of training to accurately evaluate their subjects (174). Just like how humans would need to know what constitutes a good essay by possibly viewing samples, an AEE model must be trained using a corpus of data for comparison as well. If either evaluator, human or automated, is trained with bad data, then they will issue bad evaluations. Also, in order to demonstrate the unbiasedness of these applications and the ability to meet various standards such as evaluation certification, AEE scores have been shown to compare to human scorers who have taken the same certification tests.

The exploration of AEE applications to this point has largely focused on their uses in cognitive evaluation. However, the use of AEE in evaluating intrapersonal traits in media such as student essays presents a more complex array of challenges. The evaluation of sentiment in student writing has been a largely unexplored field (Burstein et al. Automated). An opinion expressed in a constructed argument can be expressed in many different formats, each with its own criteria of success. This makes it difficult to develop a natural language processing application for evaluating sentiment in this constructed response. Burstein et al. Part of the difficulty in analyzing a sentiment is that words expressing opinion are not simply associated with one polarity or another (positive or negative). Words can be associated by varying degrees too a sentiment. Words such as

irritated and enraged carry a very different degree of intensity while still pointing to a negative sentiment. If essays were merely judged on their overall sentiment, an essay that lists a number of minor irritants but an overall positive evaluation expressed in a short statement, versus one exposition of impassioned anger would be evaluated as equal by an AEE application simply because both shared the same sentiment. In order to better meet the challenge of evaluating sentiment, Burstein et al. built a family of lexicons containing words associated with a certain sentiment polarity and evaluated them based on their performance in identifying sentiment by type and intensity. This was a step in their goal of building a system that can identify portions of essays that use sentiment to contribute the overall quality of the essay.

2.4 Constructed Response Tasks

The term constructed response is a superordinate classification for a broad range of tasks (Bennett). A common scheme for a categorization of item types is useful for its six types: selection/identification (the task of deleting extraneous information from a paragraph); reordering/rearrangement (ordering and sequence of information); substitution/correction (sentence combining); completion (sentence completion); construction (production of a total unit of thought), and presentation (a performance). In the kinds of robust construct representation required for non-cognitive assessment, constructed response tasks were created that result in demonstrations of writing performance are preferred. That is, constructed response writing tasks, linked to a specific construct model, allow deeply considered construct model to be employed.

In order to evaluate intrapersonal traits like the Big Five, a constructed response that presents opportunities personal traits and opinions should be analyzed. Movie

reviews are a genre of literature used to analyze and evaluate films for an audience. Film reviews typically are crafted in the literary review format from which it originated. These reviews can be broadly slotted into three categories: academic reviews, journalistic reviews, and fan reviews. Academic film reviews can be written to better understand why a film works, how it works, the message it carries, and how it affects the audience. Journalistic reviews analyze a movie for the audience the journalist represents from a professional point of view. Online reviews are written to be helpful to the peers of the writer, and are written in the similar way to a friend telling someone what they thought of a movie. A star rating of one to five stars typically summarizes online movie reviews, although some reviews borrow the signature thumbs up and thumbs down rating of Gene Siskel and Roger Ebert. Other metrics for evaluation ratings are used, these are just some examples.

Taboada defines genre as the "structurally-determining characteristics of texts" (249). This analysis of online movie reviews identifies five characteristics, or stages, in their creation; they are subject matter, plot, characters, background, and an evaluation. These stages appear to be generally interchangeable in their order and, with the exception of the evaluation stage, can be left out of an online review at the desire of the author.

The evaluation stage of online reviews is the most important part. Since the purpose of a review is to present the opinion of the complete film, the genre will have various lengths. Every other statement of a film review should be in support of this phase of the review. Depending on the total word count of the review, the evaluation may be one sentence, several paragraphs, or simply amount to a simple statement such as "Do not see this movie." Various other stages of a film review may contain an evaluation of

specific points such as characters or plot points, but an evaluation of the film as a whole is always present. Every review analyzed by Taboaba contained an evaluation stage of a certain length (252-255).

This topic brings us back to semantics. Sentiment analysis is used to determine the polarity of the sentiment of the author. Movie reviews are a perfect example of a writing format that allows the author to express his feelings, or sentiment, in his own words. This constructed expression of sentiment is the reason that movie reviews are the chosen genre of our LSA data analysis. Further information about the reasoning behind this decision can be found in the literature review presented in the following sections.

CHAPTER 3

LITERATURE REVIEW

3.1 Introduction and Purpose

The following section is a literature review of the source material that is the basis for this thesis. There are a number of sources that can be used to justify the proof of concept study performed in this thesis.

3.2 The Big Five in Predicting Academic Performance

The Big Five variables are indicators of academic performance and can be found in written texts. There have been a number of research studies performed to determine the relationship between various personalities constructs and academic performance (O'Connor 339). An example of a test of the Big Five influencing academic performance can be found in an article by a researcher Maureen Conard. In this test, 300 full-time graduate students took the NEO Five Factor Inventory (Form C) to measure their respective personality traits. The data from this test was then compared to their course performance. The results of this study confirmed that there is an incremental validity of the Big Five traits over academic ability such as the SAT performance test. Of the Big Five, conscientiousness was shown to predict three academic outcomes (GPA, course performance, and attendance) over the other four personality traits. In fact, for every one standard deviation increase of conscientiousness, GPA increased by 0.11 percent (344) on a 0-4.0 point scale.

O'Connor and Paunonen conducted their own literature review of the Big Five personality factors influence on post-secondary academic performance and found that

Conrad's test falls in line with many other research tests. This verification was achieved by comparing the correlating Big Five personality traits to academic performances in 23 such academic research tests between 1991 and 2006. In the final tally, 20 of 23 significant correlations were found in conscientiousness whereas the other four traits each had between three and eight significant correlations (975). O'Connor goes on to state that conscientiousness has often been tied to level of motivation to perform well. Traits such as openness and agreeableness have been tied to academic ability and GPA scores (O'Connor 975-978). Expanding the literature of O'Connor and Paunonen, Figure 3.1 presents a basic variable model in which the Big Five personality factors may be examined for their relationship to academic performance.

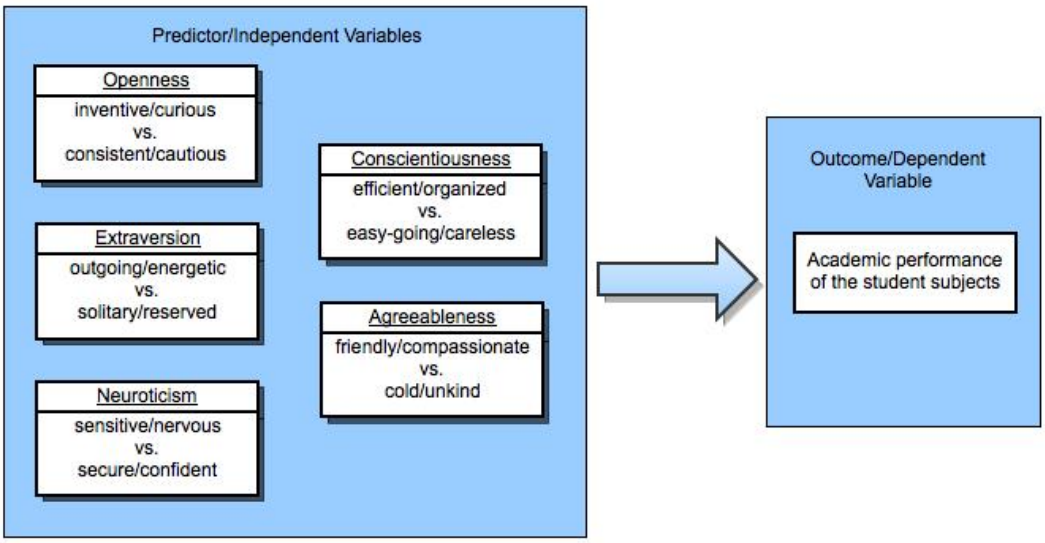

Figure 3.1 Predictor Chart of Dependent/Independent Personality Variables

Are the Big Five personality traits related to academic performance? Yes, with some caveats. Not all of the Big Five have been shown to predict academic performance. Neuroticism and extraversion were not shown to significantly contribute to academic

performance and the contributions of openness and agreeableness seem to be too unstable to accurately use. If any future testing for Big Five personality traits in students' essays is going to show success, it will most likely find it by searching for evidence of conscientiousness, as this seems to be the most important personality factor for academic success.

Much of the work of reviewing how the Big Five can be interpreted from student texts comes from how language can be interpreted to build a personality structure. Part of the basis for assuming personality traits can be deduced from written media is derived from Srivastava, who considers the arrival of these five personality factors a simple extension of the Lexical Hypothesis (9): the personality characteristics that are the most important in peoples' lives will eventually become a natural part of their language and that the more important characteristics are more likely to be contained within a single word. Ashton and Lee defend this approach to personality structure building and provide reference to how the Big Five were originally built using a lexical approach to personality. The lexical hypothesis does not itself specify the parts of speech in which personality attributes will be determined; therefore, the attributes of a personality lexicon might include nouns, verbs, adjectives, and adverbs which could be found within, say, an set of student writing. While some parts of speech may prove more valuable than others for detaching these traits, the fact they might be present in the written media is enough to satisfy the basis of analyzing texts using AEE technology.

3.3 Reviewing the Validity of AEE

Before AEE applications can be used to analyze text, the validity of using such tools must be established first; however, doing so can be complex. There is no monolithic support of

the validity of AEE or are only just beginning. Each application of AEE technology must be supported by its own set of validity. Comprehensive independent studies of AEE, such as the work by Shermis and Hamner, are either rare or still in their initial stages of research (313). It must also be kept in mind that as AEE grows, as Hakuta predicts will happen, the claims about the validity of each method will develop as well and each claim must be met individually.

Certain AEE technologies, especially those using LSA, are designed to analyze a corpus of texts and assign scores based on the model design of the previously annotated and categorized data (Landauer et al.) They have been able to achieve this with a certain amount of success. In fact, AEE technologies are currently able to assign single score grades to essays comparable to human scorers and perform certain evaluations (Tkacik). Shermnis and Hamner might have some reservations to this stance as they conclude that current AEE tools are sufficient for low-stakes assessment or for a second evaluator in high-stakes assessment in a general approach to scoring and evaluating essays. They are not disregarding AEE as a flawed technique, but they do caution that it might function best as a method of providing a second evaluation to a human scorer.

Elliot and Klobucar, as well as Shermis and Hamner, share a somewhat similar view of AEE applications in evaluating cognitive skills. While an interesting and developing field, AEE is a supporting tool in its various applications—one that should be carefully researched to better understand its development and use. Elliot and Klobucar recommend that while there is a place in academic learning for AEE, certain strategies should be followed in implementing it. First, as a growing field, AEE technology and its researchers should be looked at as investigators rather than managers. Thinking as

researchers encourages a dispersal of knowledge allowing for others to contribute their own findings and add to the AEE body of knowledge. Second, barriers to innovation should be recognized and evaluated in developing a base of validation. Finally, while focusing on the novelty provided by AEE, care must be taken in deciding reasonable capabilities of the technology and its applications (30-31).

The gold standard for determining, developing, and evaluating the performance of AEE engines has traditionally been with how they compare to human scorers (Bridgeman 229). However, he goes on to state that this approach is to simplistic. He quotes an article by Bennet and Bejar that AEE scoring validity includes “inter- play among construct definition and test and task design, examinee interface, tutorial, test development tools, automated scoring, and reporting-for in the development process these components affect one another” (229). While human scoring should not be the only standard upon AEE is evaluated, it will remain a strong consideration in the near future. That conclusion coincides with the research conducted in this thesis and its desire to determine if LSA can mimic the evaluation of student texts in the same manner as a human scorer.

In the end, Williamson, Xi, and Breyer as well as Bennet point to this meaning, that the responsibility for establishing the validity of the AEE application being used is largely left up to the researchers. A framework established by Williamson, Xi, and Breyer provides a solid foundation upon which that validity can be built. While this framework is targeted at the ETS E-rater application, this framework can be used for other AEE applications. This framework is divided into five areas of emphasis. These points are listed below with their corresponding inference within a validity framework and information about the guidelines and criteria inherent within each area. These parts are

what will be used to establish the basis of the research (both conducted and proposed) in this thesis.

- Construct relevance and representation: evaluating the fit between the capability and the assessment (explanation). Do the goals and design of the task AEE is being used to perform fit with the application itself? This point is establishing the content validity of the design. This is done through evaluating the construct, task design, scoring rubric, and reporting goals (6).

- Empirical performance: association with human scores (evaluation). Williamson, Xi, and Breyer as well as Bridgeman consider human scorers to be the gold standard upon with AEE must measure up against (7; 221) with some caveats. Williamson et al. suggests a number of different criteria for ensuring the criterion of the performance of AEE applications in relation to human scorers. Some examples include verifying that the process for scoring by the human raters sufficient for the training of the automated models. What is the threshold for human adjudication? And evaluating the task type and reported score level for any difference when the human scorer is replaced by the AEE application (7-8). It is noted that these listed criteria are sufficient for when AEE is acting as a supplement for a human evaluator. If AEE is totally replacing the human factor, a more stringent set of criteria should be established.

- Empirical performance: association with independent measures (extrapolation). This area examines the relationships between automated scoring and other possible external criteria beyond the simply human/AEE relationship. Identifying points of variance and the relationship that spawned it may lead to improvements in design or validity. It is very important here to note that current AEE practices leave very little unknown outside influences. However, when AEE is applied to areas that have not had much examination, such as those proposed in this thesis, divergent patterns of evaluation may be found and then analyzed.

- Empirical performance: generalization of scores (generalization). This area examines how the scores produced by AEE are generalizable in comparison to human scores and can these generalities across various tasks be used to provide any insights into consistency. Also, can alternate forms of the human/AEE interaction be applied to other forms? This point is significant in the present study as it uses the same application for detecting five different personality variables.

- Score use and consequences: impact of decisions and consequences (utilization). This area deals with the results and consequences of using AEE applications on its consequences. What is the impact of using AEE?

> What type of claims and disclosures will be used? And what will happen by replacing even one human element from the process AEE is being applied to.

These evaluation points of validity were created for use in high stakes AEE application evaluations (11). While the study being performed in this thesis is a descriptive, baseline proof of concept, it represents a proposed process that does have a higher risk and higher ceiling in its application. As such, if the conceptual process done in this study does not carry a sufficient level of validity, then the follow-up study will not have the required level validity for authentication. The conclusion of this line of reasoning is that a high level of validity evidence will have been obtained in both the performed research and the proposed future study in this thesis.

3.4 Personality Indicators in Online Film Reviews

It has been established that personality does have an effect on academic performance (some traits more than others) and that automated methods of evaluating essays are viable. It must now be shown that aspects of personality can be demonstrated in a written text format. For this thesis, that written format means movie reviews. Studies have shown that there are links between an individual's morality and the Big Five personality traits (Williams et al.). This research is even more pertinent to this study in that conscientiousness has been shown to be a Big Five trait heavily linked to morality (Williams et al. 3-4). Research by Yarkoni is able to show limited evidence of Big Five personality factors in blog postings. This idea can link personality traits to writing. However, an argument could be made that there may be a variance to these studies because the age of the sampling size may be varied (an age range of the sample

population is not given) and it has been shown that personality is subject to change over time (McCrea et al).

As described earlier, movie reviews contain an evaluation stage that is determined by the opinions of the author of the review. This indication of semantic polarity is then conveyed to the reader who can be influenced by this opinion. A study by Wyatt and Badger was set up so understand how a positive, negative, or neutral film review can influence an audience in comparison to a neutral presentation of information about a movie. The researchers designed an experiment to determine how reviews containing only neutral information of a movie, reviews containing low levels of information and positive to negative reviews, and reviews containing both high levels of neutral information and positive to negative evaluations can affect movie attendance. What they found was that while reviews high in only neutral information and no sentiment polarity caused a limited increase in movie attendance, positive reviews with a high amount of neutral information increased attendance the most while movies with negative sentiment and with both high and low amounts of information were able to decrease attendance by a measurable amount. This study shows that evaluation sentiment of movie reviews can be imparted to an audience, detected by natural language processes, and shows indications of personality traits.

Another study by Barriga sought to understand the level of morally relevant comments found in movie reviews and then how readers reacted to those reviews. Since personality has been tied to moral character (Arvan; Barrio, Aluja, Garcia; Walker) the results from this study show that movie reviews are a viable format from which it may be possible to use automated essay analysis tools such as LSA to pull Big Five character

traits that then could lead to predicting academic performance. As Barriga points out, "…in order to actually produce conscious moral thoughts, people would have to recognize that there are moral elements present in the movie" (5). Barriga analyzed the comments reviews left for 14 contemporary films of various moral ambiguities at the Internet Movie Database (IMDb.com) and then analyzed the responses to those reviews (2-4). The research showed that movies with higher levels of moral ambiguities produced reviews with more morally pertinent content (12). The implications of this study on the proposed research in this thesis are that any reviews being analyzed by an automated essay evaluation tool should be constructed so that the writer expresses their views about a morally ambiguous matter. A larger corpus of text about morally uncertainties should allow for more aspects of the author's personality to be expressed in the text. The value of larger corpuses of text has been substantiated in previous LSA research (Layfield). This, in turn, will provide more opportunities for automated evaluation tools like LSA to detect personality traits.

3.5 Using LSA to Detect Non-Cognitive Personality Traits

Latent semantic has been used for analyzing gender stereotypes (Lenton et al.), analyzing song lyrics for cognitive components (Petersen et al.), and predict psychological phenomena (Wolf and Goldman). However, very little has been done for using LSA methodologies to analyze any facets of personality.

As of publication of this thesis, only two studies have been identified detailing research using LSA or natural language processing for detecting non-cognitive personality traits from any corpus of text. There are some key differences between the methods being used and the goals of the research but these differences are not enough to

discourage further research on this topic. Jon Oberlander and Scott Nowson used support vector machine learning to analyze weblog entries for evidence of certain Big Five personality traits. Using the WMatrix tool, Nowson and Oberlander apply Support Vector Machines and Naive processing tools to their corpus of weblog entries. The researches' results showed they were relatively successful in finding evidence of personality traits in their weblog text, more than enough to go forward with further testing. With their application of natural language processing, they speculate that they should be able to identify the author of future weblogs and the type of personality the author has with some success.

While Oberlander and Nowson did not use LSA in their research, LSA has specifically been shown to be able to handle identifying personality traits in the work of Bates, Neville, and Tyler. Using LSA algorithms, the researchers analyzed whether chatting, face-to-face speech, or written communication produces better results in predicting the gender of the communicator, their political affiliation, and the level of aggressiveness of the communicator. These variables are not the Big Five factors of personality, but they would qualify as related non-cognitive variables. The authors show that when they were able to make a prediction about any of the above fields, a written corpus of data, rather than dictated spoken words, is more effective at revealing useful information about the author but dictation of spoken conversation produces a larger corpus. While Oberlander and Nowson showed that personality factors can be detected with natural language processing, Bates et al. demonstrated that LSA can be used to detect non-cognitive factors. Combining these two studies produces encouraging results about the viability of the proposed research.

What the previous two studies show is there is some basis for using LSA in the detection of personality traits such as the Big Five, Nowson and Oberlander are procedurally very close to the research proposed in this thesis but do not use LSA. In addition, the content vector analysis process they use is not as sensitive to identification of certain words as LSA, and thus may not be as robust as process in document comparisons based around textual meaning (Burstein 4-6). While both content vector support and LSA use a similar branch of vector mathematics, they are different enough to warrant disregarding them in a direct comparison. In comparison, Bates, Neville, and Tyler use a direct application of LSA algorithms based on Landaeur et al. on non-cognitive aspects of personality with some success. Obviously the procedures between this thesis and Bates, Neville, and Tyler vary, but the spirit of the study is as close to a direct application of using LSA to determine specific personality traits as could be found to date.

3.6 Literature Review Conclusion

A number of diverse topics from various backgrounds have been discussed to this point and their relation to the main topic of this thesis is presented. From this review, sufficient evidence has been presented that it can safely be assumed that using LSA to detect personality variables is viable. Indeed, a chain of evidence has emerged. Big Five personality traits can be influential in academic success, and these personality traits mirror a lexical approach found in communication. As well, it is clear that AEE applications are valid methods of evaluating text, although the responsibility for demonstrating this validity is falls upon the researcher. In such validation research, online reviews have been shown to be acceptable constructed formats for writers to display

personality-defining characteristics. It does thus appear that these non-cognitive personality variables have legitimate expectations of being detected using latent semantic analysis methodologies.

CHAPTER 4

METHODOLOGY

The following sections describe the method used to test a proof of concept procedure that will demonstrate how future testing in the theories described in this thesis may be carried out. This chapter describes the sampling plan, descriptions of validation design, and calibration of the tools used to produce scores.

4.1 Population Sample

For the purposes of this study, there are two sampling populations that must be defined. The first population is the large corpus of data that is assessed by human hands that the AEE tools use to determine the traits that all these samples have in common. This set of data is then applied to the second sample population, the essays being evaluated.

LightSIDE Labs provided a set of annotated test data that could be used to create the machine-learning model that was created for this research. It contained 600 movie reviews containing 300 positive reviews and 300 negative reviews (Mayfield, CSV). The movie review samples were of various lengths, typically between 700 and 1200 words, and had already been annotated as either positive or negative in their sentiment polarity. While it is convenient that this data was already annotated, these annotations had to reflect the same scoring methods as what would be applied to the student essays from a human scorer and from there used to make predictions on the student essays. To verify the provided reviews matched the human scorer's evaluation, every fifth review was read and evaluated using a modified Bales Interaction Process Analysis chart (IPA) to determine if the sentiment of the review coincided with how the student essays would be

evaluated by the human scorer. There were no discrepancies found in using this method between the two sets of dates.

The second set of sample data being analyzed contained essays from 15 New Jersey Institute of Technology graduate and undergraduate students from Dr. Norbert Elliot's STS 307: Fundamentals of Research in Science, Technology, and Society; PTC 604: Communication Theory and Research; and Dr. Andrew Klobucar's COM 303 Video Narrative class. These essays were representative of the types of writing assignments that were performed over the course of the pertaining classes and were constructed so that the writers' attitudes about writing would emerge. Participants were instructed to write essays at least one thousand words in length. As Layfield demonstrated, the size of the items analyzed with LSA algorithms does influence the results in evaluating writing. Requirement lead to choosing a minimum number of words that was on the larger end of what could typically be found on the Internet. Finally, the students were tasked to construct reviews of their attitudes toward writing—and, in that case of STS 307 their attitudes toward a specific university-wide test they had taken—in a specific constructed manner using the following structure outline:

- *Background*: The participant was prompted to write about the course being taken and the kinds of writing being performed for it.

- *Classification:* In this section, the writer wrote about the kinds of writing tasks being performed and how these tasks fell into their own classifications of different types of writing.

- *Plot:* This section asked that the writer describe the class's writing assignments as a narrative with a beginning and an end and include the feelings they experienced while performing these assignments.

- *Evaluation:* This would be the focus of the review and hopefully the largest section. The writer would compare this writing class against others they have

enrolled in and how they felt about them. It would also provide an overall evaluation of the assignments taken in the class.

- *The Take Away:* This section was the final course evaluation; this section prompted the writers to give a thumbs up or thumbs down evaluation of the class.

In an effort to increase the population of text being analyzed, writing samples from 28 students from Dr. Irvin Peckham's English 3301 class at Louisiana State University were added to the original set of data. These samples were collected with Dr. Peckham's permission from *Writing Ourselves in to Each Other's Lives*. This work collected the attitudes about writing shared by the students in the class. Many of the writing samples pulled from this work contain similar elements as those described in the constructed response evaluation such as the students' attitudes toward writing and an evaluation of the work performed in this or other writing classes. More importantly, there is a polarity in the readers work about writing that can be used as a source for analysis in the research for this thesis. Almost all the student participants from this set had at least 1000 words of text that included their feelings about writing, so the length of the text is in line with that of their NJIT counterparts in this study.

In total, 43 writing samples were collected for analysis in this research. Upon analyzing the sentiment of these studies using the same method as the movie reviews detailed earlier, a human rater determined that 42 were positively inclined in their attitudes toward the subject at hand while one was negatively inclined. For most sentiment studies, such a strong bias toward one polarity in the analyzed set would be detrimental for making sound judgments. However, for a proof of concept test such as the one being performed in this thesis, the number of samples being analyzed is a more important consideration.

4.2 LightSIDE Procedure

The engine driving the research in this thesis is the open source, machine-learning application called LightSIDE (Mayfield and Rosè). This software application was designed to allow a non-technical level audience to perform machine-learning applications using complex statistical models rarely available to them. LightSIDE provides numerous algorithms for the feature extraction of text and for performing machine learning with the Weka toolkit (Hall et al.). There were three main areas used in LightSIDE for performing this basic proof of concept test: feature extraction, model building, and predicting labels. While LightSIDE itself does implement NLP features, it can employ natural language processing artificial intelligence feature sets for the purposes of research and development (Shermis, Burstein, and Bursky 11). It is important to keep in mind that LightSIDE is not a LSA application, but instead a machine learning application that can employ natural language process algorithms, such as latent sematic analysis, for evaluating text.

The feature extraction functionality of LightSIDE employs several optional plugins for establishing accurate text representation. For extracting data from text, LightSIDE uses TagHelper (Rosè et al.), which is built on the Weka toolkit, to turn a set of text into a set of feature vectors in a table (Mayfield, Adamson, and Rosè). This feature extraction function was used to parse the set of annotated movie reviews. The basic configuration features selected in the extraction phase of this research were unigrams, punctuation, POS (parts of speech) bigrams, and binary N-grams.

- N-grams: Consisting of unigrams, bigrams, trigrams mark single words, pairs of words, and three consecutive words. This allows words that would possibly have their meaning changed by their neighbors to be processed.

- Punctuation: Punctuation could be included or excluded from being processed. This option is typically removed too if trimming the dataset is required. It was not required to be removed for this study and punctuation might provide further insight to gathering personality traits in later studies.

- POS bigrams: Including parts of speech as an evaluation field in LightSIDE extracts bigrams that have been abstracted to level of parts of speech. This option proved especially useful in this study for increasing the accuracy of the test model.

- Binary N-grams: Rather than have LightSIDE reduce each feature to a true or false value, there are sometimes cases where the number of times words are used affect the evaluation. This feature is used when the size of the essays is variable and employs stop words.

- Remove Stop words: Stop words (it, a, the, etc.) are words that typically do not carry any significant meaning. This feature is not very useful in longer texts such as the essays used in this thesis. This feature might be used when analyzing text in instant messages, but not for longer essays. However, this feature was still employed so as not to remove possible negation words and prefixes, also to give LightSIDE the largest bag of words available to work with.

During the research and test setup process for this thesis, there was no determination about the choice of whether N-grams would work best for this test. Bigrams and trigrams might present more meaningful relationships for the LSA algorithm to analyze. However, the use of these options would also cause fewer instances for LSA to analyze. Unigrams present a bag of words approach for parsing data, which creates more instances for analysis.

There are some notable extraction options that were not employed in this research. LightSIDE allows users to include stop words in its extraction. While including stop words might increase word count available for an LSA analysis, stop words are typically removed from natural language processing methods (Bates et al. 2) and are not considered useful for longer sets of text being analyzed (Mayfield, Adamson, and Rosè 13). Line length, used when LightSIDE includes the number of words in the document as

a feature, was also an option. This feature was disregarded, however, because all the text being used in this research study fell within a similar length. Finally, stem N-grams were not included in this research. Stemming reduces words to their base forms so that words like drive, driving, and drives would all represent drive. This would cause LightSIDE to lose inflection, but enhance its generalization. While this proof of concept test might get away with this feature, it does not seem like a beneficial trait in future training when LightSIDE is analyzing text for non-cognitive variables.

After the LightSIDE feature settings were selected, the model using these extracted features was built. This process began with selecting a machine learning plugin. The Weka engine contains the classifier needed to incorporate latent sematic algorithms, specifically the weka.classifier.meta.AttributeSelectedClassifier, allowing LSA to be applied. This classifier reduced the dimensionality of the training and test data attributes created in the model by applying an evaluator to the data before it was passed off to the machine learning plugin. This evaluator is where the latent semantic analysis algorithm was applied. It was configured to apply data reduction to any sets in the matrix ranked below 0.95 kappa rating. A ranker application was then applied that ranks the remaining attributes by their individual evaluations. It was recommended that the number of attributes retained by the ranker (numToSelect) should be explored with a number of different figures (Mayfield Student Inquiry). It was determined that the best results were achieved when all attributes had been retained. This might be because once the LSA algorithm had identified the most meaningful relationship attributes in the training corpus, cutting them down further might negatively impact the model prediction. The size of the test was probably also a factor in this outcome. The larger corpuses of test data

LSA was designed with function with allow it to remove less significant relationships, yet still leave a large corpus of test data.

Finally, a classifier has to be chosen. This classifier is what will be used to create the model from our control data corpus. For this thesis, the Naïve Bayes learning plugin was selected. This classifier deduces the probability of each possible label and assigns a label based on what it found most frequently in the text it analyzed. It then determines the prospects of the observed features of text occurring, which have occurred in the data being analyzed (Mayfield and Rosè). Naïve Bayes has seen a great deal of success in email spam filtering and is considered a good option for basic text classification. It is also considered to work well with weak predictor indicators and with multiple labels. This means it will have an advantage in processing texts for evidence of multiple types of personalities and their indicators. With the machine learning plugin configured, the model based on these settings was created.

The task of making predictions of the student essays once the model is created was a rather simple one to implement. LightSIDE prompts the user to select the model it has processed and then the data it is being applied to. Since the training data being analyzed by the model has already been annotated, a side-by-side comparison between the results of the LightSIDE model and human scores was plainly visible, which can be viewed in Chapter 5.

CHAPTER 5

RESULTS AND VALIDITY OF THE STUDY

5.1 Results and Discussion

Once the training model was compiled, it was applied to the student review essays. LightSIDE then issued predictions of the positive or negative semantic polarity of the documents that were analyzed. In the 43 student review essays that were collected, it was determined by a human scorer that 42 were positively inclined while only one was negatively inclined. The test model that was compiled analyzed these essays and predicted that 39 essays were positively inclined while three were negatively inclined. Both the human scorer and the test model agreed on the negative essay. However, the human and test model evaluations disagreed about three of the positively inclined essays. If it is assumed that the human scorer is correct, and then the training model was 93.02% accurate.

Percentage is not quite the number that needs to be used to properly determine the validity of this research, however. A kappa, or the measure of the degree to which two judges, A and B, concur in their respective scoring, is a much more appropriate figure to determine this. It is also the type of figure LightSIDE uses to measure its training models and, more importantly, this is the figure Williamson, Xi, and Breyer use to help determine validity in their own framework.

To that end, a confusion matrix was created that illustrates the instances that the human scorer and the LightSIDE training model agreed. Table 5.1 below shows that the human and test model agreed on thirty nine positive essays and one negative essay, while

the test model disagreed about three negative essays the human scorer believed to be positive.

Table 5.1 Confusion Matrix of the Testing Results

		LightSIDE Test Model Scores		
		Positive	Negative	Total
Human Scores	Positive	39	3	42
	Negative	0	1	1
	Total	39	4	43

From this information, Cohen's Quadratic Weighted Kappa value, standard error, and the upper and lower limit of kappa with a 0.95 Confidence Interval were calculated (Lowry). These numbers can be seen in Table 5.2.

Table 5.2 Quadratic Weighted Kappa Results after Running the Test Model

Observed Kappa	Standard Error	0.95 Confidence Interval	
		Lower Limit	Upper Limit
0.38	0.23	0.00	0.82

As Table 5.2 demonstrates, the observed kappa value is 0.3768. This is not a strong agreement score between the human scorer and the test model. In a normal kappa rating scale, this number would be in the high end of a fair agreement. Unfortunately, for

achieving validity in an AEE procedure, this number does not approximate the minimum .70 mark recommended by Williamson, Xi, and Breyer in cognitive domain scoring(7).

This outcome would imply that the procedure being used is not a valid one and would have to be redesigned; however, another explanation may be posed that leaves this process valid. As described in section 4.1, of the 43 student essays collected, the human scorer labeled only one essay as negative. This population is not a very balanced cross section for the test model to demonstrate its abilities with. While this is not a great deterrent to a proof of concept test, the results were explored to establish how such a problem might influence future studies.

This process began with determining if this imbalance led to such a low kappa value; to determine if this was the case, calculations exploring other possible results within the expectations of this study were conducted. A second calculation altered the number of positive and negative essays agreed on by the human scorer and the test model so that five more negative essays were evaluated and five less positive essays were evaluated. This altered the confusion matrix in Table 5.1 to what is shown in Table 5.3 below.

Table 5.3 Hypothetical Confusion Matrix with More Negative Reviews

		LightSIDE Test Model Scores		
		Positive	Negative	Total
Human Scores	Positive	34	3	37
	Negative	0	6	6
	Total	34	9	43

The total number of essays scored and the number of divergent negative/positive remained the same; only the number of agreed upon positive and negative essays changed. This change alters the number of positive and negative essays in the population sample of student essays, but it preserves the number of instances in which the human scorer and test model agreed with each other as well. From this hypothetical situation, quadratic weighted kappa value was recalculated. These recalculated values can be found below in Table 5.4.

Table 5.4 Recalculated Quadratic Weighted Kappa Values Based on Table 5.3

Observed Kappa	Standard Error	0.95 Confidence Interval	
		Lower Limit	Upper Limit
0.76	0.11	0.55	0.97

As this chart demonstrates, if even a small number of additionally negative essays had been collected, and accurately labeled by the test model and human scorer, then the observed kappa value of this process will have been pushed high enough to clear the kappa level proposed by Williamson, Xi, and Breyer for ensuring the validity of this process.

So what does this mean for the research in this thesis? The relatively small size of data resulted in this unforeseen issue and it was only evident once the kappa score was calculated. Obviously the value of the observed kappa calculation might be affected when only one attribute is being deduced by machine learning rather than having some

represented in the two polarities equally represented in the data sample. It could be said that the agreement between two scorers is difficult to deduce when they only agree on one facet of evaluation. For purposes of evaluating the validity of a model, this type of research might also benefit from having more attributes to test against to ensure a more accurate kappa score.

A second possible method for making an improvement in this process is if the test model can be calibrated to better match that of the human tester. If the three point score of positive and negative disagreement between the human and AEE scorers was reduced to one, then the observed kappa would increase to 0.66, almost to the validity threshold demonstrated in Table 5.5. One possible way to further increase the accuracy of a test model is to increase corpus of data used to train it. Latent semantic analysis only benefits from an increase in data and the only limitation on how much it can handle is the processing speed of the machine exercising it.

Table 5.5 Second Recalculation of the Quadratic Weighted Kappa if only the Test Model Mislabeled One Essay Instead of Three

Observed Kappa	Standard Error	0.95 Confidence Interval	
		Lower Limit	Upper Limit
0.66	0.26	0.14	1.00

The expectations of those involved in this thesis were largely met. Previous studies have shown that latent semantic analysis has had success with detecting non-cognitive variables (Wolfe; Bates, Neville, and Tyler; Connolly, Veksler, and Gray), and LightSIDE is a tool that allows for developing research processes in AEE applications

and algorithms. The interpretations of these results and how they represent the test model that was created to evaluate them, it was decided that the null hypothesis being tested is false and that the alternative hypothesis is valid and further research of this topic is warranted.

Part of the expectations going into this project was to determine what kinds of obstacles and limitations might be faced upon further research into this topic and using this methodology. The kappa calculation in the testing results show that determining how well a test model will work will require that there is a range of material for it to be tested against before the model can be said to be of a sound valid framework. Fine-tuning the settings under which this model is built will also be important. As this research was being designed, it was originally thought POS n-grams would not be a very consequential setting for parsing the test model data. This idea proved to be in error as the hit rate for agreement between the human and AEE software evaluations improved by six points, a not insignificant margin considering the size of the test data being employed.

5.2 Process Validity

As explained in Section 3.3, establishing the validity of AEE applications is largely the responsibility of those applying it. To that end, this section will be following the framework created by Williamson, Xi, and Breyer to establish the validity of the research process and the subsequent results generated from them.

A number of academic sources have recognized LightSIDE in the discussion of AEE applications, and Tkacik states that LightSIDE is comparable to other AEE

applications[3]. LightSIDE is also unique among AEE tools in that it is an open source application, the backbone of this application being the machine learning application designated as Weka and its supporting algorithms (Hall et al.). As an open source program, algorithms can be created by anyone with sufficient mathematical and Java programming skills. Such sources should also be verified as having been created by reputable sources, however. In this case, the LSA algorithm being used in this thesis was created by Napolitano. A Microsoft Academic search revealed him to be an author in 31 publications in the fields of software engineering, data mining, and artificial intelligence.

After showing that the LightSIDE and Weka tools are valid for this project, the most important piece of this research left to validate is the training model created from these tools and the training text. The gold standard described by Bridgeman is echoed in the validity processes of Williamson, Xi, and Breyer. A way for determining the validity of a test model is to compare the evaluation results of this test model against that of a human scorer. For this research, the test model using LSA was attempting to match the scoring data of a human scorer. This scorer had verified their own annotated data matched with what had been assigned to the evaluation data. As the results in Table 5.2 show, the kappa value of this research test did not clear the 0.70 needed to meet the validity standard in an AEE application, but with some adjusting and hypothetical adjustment, it could easily have met this requirement.

The second way of assessing this model is found in LightSIDE itself. This tool provides the ability for testing the model using cross validation while creating the model. Cross validation is a model validation technique that assesses how the results of a statistic

[3] These sources include Shermis, Burstein and Bursky; Elliot and Klobucar; Shermnis and Hamner; and Mayfield and Rosè.

will generalize to an independent data set. In the case of LightSIDE, that data set is found in the text being used to create the model. In cross validation used LightSIDE, if the number of folds being used to validate the model is set to N, then the text being used to train the model is divided into N parts. One part will then be held independently of the rest and the majority of the sets will be used to predict the data in the remaining hold out data set. This process will be repeated N more times with each hold out data set being unique.

From this cross validation process, a sense of the accuracy of the model can be deduced (Mayfield, Adamson, and Rosè 27-28). In the creation of the LSA model, cross validation was set to use ten folds, or tenths of the testing data, to make predictions on the final fold. This cross validation method produces a kappa value of 0.673, a little low on the good range of kappa scores and an accuracy rating of 34.7%. While the accuracy rating is somewhat low, it is possible that because this cross validation uses one tenth of the original data set available multiple times over. The processes may be stunting this accuracy number. A larger set of documents would likely generate a more precise accuracy rating as well.

Beyond explanation, evaluation, and extrapolation areas of Williamson, Xi, and Breyer's validity framework, the final two categories of validity evidence—generalization and utilization—may seem somewhat difficult to establish the validity of, simply because of the narrow confines of this particular research study. However, there are vast possibilities available with the proper application and modification of the procedures outlined in this thesis. In this thesis, students were asked to construct their thoughts on writing and their writing class in the form of a movie review. The overall

positive response indicates that the handling of this subject is on the right track and perhaps could be reinforced. If the constructed response were to feature the students' opinion on a different subject as their attitudes toward cooperative learning, it might be possible to determine the sentiment behind that topic and how and why it generates the recorded sentiment. The utilization of these scores might be used to identify possible methods of enhancement to the topic being analyzed. If the research in this thesis found that there was an overall negative view toward writing, this would spur further analysis of the reasons for such sentiment. Williamson, Xi, and Breyer might be concerned about the over reliance on automated essay evaluation triggering such an investigation, but while the AEE tools may state that something is wrong with student sentiment toward writing, it cannot identify what is wrong or suggest how to change it. This would require expert analysis and the final call would rely upon them. Even the students themselves could be polled for information. If the constructed response includes suggestions for improvement, then an additional source of information could be polled.

CHAPTER 6

DIRECTIONS OF FURTHER RESEARCH

Now that the proof of concept methodology in Chapter 4 has been shown to have potential in detecting non-cognitive semantic polarity, the next step is to illustrate an example of how the tools and concepts from other chapters of this thesis can be combined to detect the Big Five personality traits. The following chapter will seek to design a research study with the goal of detecting the Big Five personality traits in post-secondary student writing. As components to this study are described, issues related to their validation will be detailed as well. A diagram of the proposed research is shown in Figure 6.1.

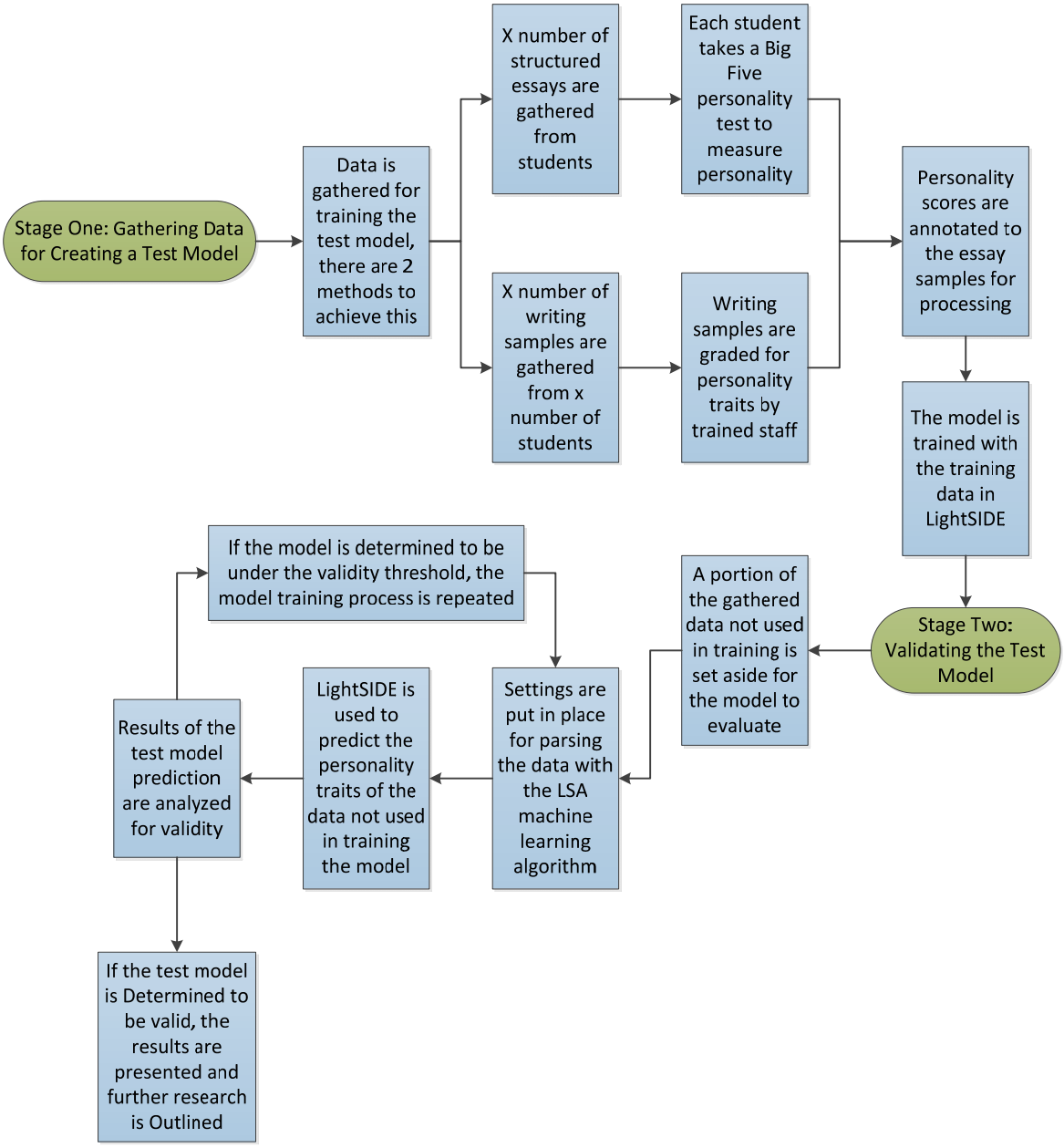

Figure 6.1 A Flow Chart Outlining the Recommended Procedure for Implementing Future Testing Methods of Detecting Big Five Personality Traits Using LightSIDE

6.1 Detecting the Big Five in Student Writing

The first step in study design is determining what personality traits will be targeted. Should all of the Big Five be included in this research, or just a certain few? As O'Connar points out, neuroticism and extraversion were not shown to significantly

contribute to academic performance and the contributions of openness and agreeableness might be too unstable to accurately use. Conscientiousness was deemed to be the personality trait most indicative of future academic success. From this information, it might be prudent to attempt to target only those three traits that have shown to have some influence in academic performance. Depending on the success resulting from this research proposal, it might be worth looking into neuroticism and extraversion. However, since this thesis is focusing on traits that determine academic performance, openness, agreeableness, and conscientiousness will be proposed traits that LightSIDE will be trained to try and detect evidence.

Now that the specific personality traits being searched for have been determined, what kind of structured situation can be presented that would prompt a student to have their personality be imprinted on a sample of written media? Using a review constructed response to critique academic written work seemed sufficient to the proof of concept test used earlier and no reason was presented in the results of this test for not using it again.

Using information from Stricker et al., a new constructed response will have to be created. Using a constructed response format that prompts students to critique their high school experiences and the academic admission process, enough material may be presented for the purposes of this essay. It might also be worthwhile to consider adding a question of some type of morally ambiguous question related to academics. Barriga showed how morality can be more readily expressed when morally questionable subjects were responded to. Adding something like this to essay instructions may help to increase the material indicating personality traits. This is only a suggestion based on current

research. Psychology experts will be consulted in order to deduce how such a procedure can be best accomplished.

The length of these essays is another matter to consider. Ideally, these essays will be a minimum of 1500 to 2000 words in length. Sentiment polarity is a much simpler concept to pull from a piece of text. Personality markers, however, are a much more complex idea that might need the additional word count in order to accurately represent a personality. This might be a difficult essay size to acquire so the idea should be left open for possibly making this part of a two or three part assignment. At the very least, the essays used to create the test model should be larger than the essays that this model is being tested against (Layfield).

With a set of instructions for prompting a personality revealing response established, a sample plan must be created for the test model being created and for the test model is being verified against. Obviously the number of documents used to create the test model it must be more than the number it is being tested against, but by how many and at what ratio? These types of numbers seem to come in the form of recommendations. With a study such as this, from 500 to 1000 would be the recommended corpus size of data with more included if possible (Mayfield Interview). It would be incredibly difficult to get 1000 unique essays from the same number of students. Multiple essays, such as what was suggested earlier in Chapter 4, from a lesser number of students might suffice as well. If a repository of post-secondary student essays becomes available, this might work as well, depending upon the method used to score personality traits in these sources.

When collecting essays to use in this study, the age of the writers will be another piece of information that has to be taken into account when organizing a sampling population. Maturation is capable of confounding the results of this type of study. As people grow older and more experienced, their age impacts their personality. The level of neuroticism, openness, and extraversion traits in college age individuals declines toward middle adulthood, while agreeableness and conscientiousness increase in the same span of time (McCrea et al.). This phenomenon would imply that the data being used to create the test model may need to be of a similar age to those the model would be applied to in order to assure an accurate evaluation. It may also mean that if certain personality factors are not as well represented as others, a different age group may be mined to increase the corpus of data.

How would these essays be annotated or graded on personality traits? Two options present themselves. The first is a panel of experts or trained staff who look for words or markers that indicate a personality factor and what polarity of this personality trait it indicates. As an example, a scoring method could be devised based on the phrases used, tone of the writer, and even specific words being used. The link between personality and language use has been studied before and Yarkoni has added to this body of research and even identified certain words that are associated with certain personality types in web blogs. This method would require more personnel and time to implement and might cause a problem with the variance found in human scorers. It would also require a review and quality control process to assess the value of the scores being assessed. A second possible method could be to have each participant in the study who contributes essays to take a basic Big Five personality test. Such tests can be found on the

Internet or a test such as the MMPI test can be used as well. This second method might prove to be more cost efficient and lack some of the problems found in human scorers, but it might not be as easy to validate and would also might require that the scoring method used by LightSIDE employ the same scale as the personality test.

After a process has been established for annotating the student essays with the scoring methodology chosen earlier, the test model will then be created. As this thesis has demonstrated, LightSIDE seems quite promising. As long as the scoring of the essays has been consistent, a valid model should be built that will mimic that human scoring (or test scoring if that route was done instead). This portion of the research design will have fulfilled the construct evaluation and task design of the validity framework. Determining if the scoring rubric is valid will mostly be determined upon completion of the research, but the design of the study is sound to this point.

Most of the evaluation validity of the model and larger AEE process will have to be done during the actual setup of the proposed research study. Safeguards in validity can be suggested in this thesis, but it is recommended that validity should be determined from Williamson, Xi, and Breyer's framework for AEE validity.

Unlike the polarity test created earlier in this thesis, a greater chance exists for independent measures to affect the validity of this AEE personality model, much more than the model created earlier to determine polarity. There are a number of influences that can affect the model are numerous just from the human elements. For example, halo effects, fatigue, an inclination to overlook details, and consistency in scoring (scorer A may score differently than scorer B) all can affect the creation of the model (9). These are not the only outside variables that might influence the model but they are good examples.

Unlike parts of the evaluation and extrapolation areas, the utilization area of a validity-based framework is something that can be speculated on at this time, at least in a hypothetical sense. This means that consequence protocols can be established for questions about what type of impact this process might have when implemented. These questions include the impact automated scoring can have on other decision making processes, how will claims and disclosure of utilizing this process be handled, and the consequences of using automated scoring for non-cognitive domains (10).

The first question to ask in the utilization of automated essay evaluation in detecting student personality traits is in what way should this system be implemented? Williamson, Xi, and Breyer recommend a number of combinations ranging from purely human scoring (something that will not work because that is what is being avoided here) to a purely automated scoring method (5). A purely automated method of detecting non-cognitive variables in a mid to high stakes assignment such as post-secondary admissions should be avoided until such a process has proven itself to be at least compatible to a human scorer. If the proposed process proves to be viable for a rollout on a small scale, then it might best serve as a supplemental or quality control method of human evaluation. If this method is implanted in an admission role, it might be used as a preliminary evaluation method for identifying retention students. It would require much more research, let alone success in the pilot study, to consider this process as viable for a production environment.

Claims and disclosure forms, like any other social research project, need to be included as with any study using outside volunteers. This is done to inform and protect the rights of the participants of any study. Williamson, Xi, and Breyer recommend that

such disclosure include the extent of the study, the strengths of automated scoring, and some general statements about improved scoring statements be included (10).

Finally, what kinds of consequences are inherent in implementing an AEE process such as this? There are a number of learning benefits that might be gleaned from this process such as a greater understanding about the communication in a written medium, how personality plays into the writing process, and the role personality can have in improving the education experience. It might also change students' opinions on the admissions processes involved in post-secondary education. Non-cognitive variables will not supplement cognitive factors such as school records, but such information may be useful in admissions.

CHAPTER 7

CONCLUSION

This thesis proposed that latent semantic analysis would be able to detect evidence of non-cognitive variables such as the Big Five factors of personality in student writing.

In order to verify this hypothesis, three different methods were employed. The first method, a literature review was given using published research in topics such as latent semantic analysis, automated essay analysis, and the Big Five personality factors to present an argument in support of this process. The second method was to use a proof of concept research procedure to show that the tools and methods needed to determine the presence of non-cognitive variables exist and can be implemented. This research showed the procedures and methods can be employed in a system that can be configured to the needs of this thesis. The final method used in this thesis is to propose a design plan of a research process that should help to further determine just how valid such a process is. From these methods, it was successfully determined that further research is justifiable and advisable in order to further develop these ideas.

There are potentially larger implications beyond the hypothesis of this thesis and proposed research it recommends. If it proves to be a viable option for latent semantic analysis and automated essay evaluation tools to detect non-cognitive personality traits, further understanding of personality, academic performance, and communication in writing may be generated. Tools such as LightSIDE offer methods of analysis that can pinpoint the areas of agreement in creating a test model—specific points of language can be analyzed to possibly help determine where and what particular words and parts of

speech are indicative of personality traits. Landauer pointed out that LSA mimics word sorting and category judgments, as well as simulates word–word and passage–word lexical data in a similar manor as a human does (2). So by building a test model designed to function in a similar manor as a human scorer will opperate, researchers could gain further understanding of the human brain's language processing and learning capabilities as well as a unique perspective on how personality effects communication.

Future research of this subject depends on further explorative examination. Using the tools and techniques outlined in this thesis, a more comprehensive research study targeting the Big Five personality traits should be possible.

Additional research is needed in this subject. The effects of such research are beneficial for both theoretical and practical use.

APPENDIX

VECTOR MATHEMATICS IN LATENT SEMANTIC ANALYSIS

LSA is an opinion mining application used to determine the contextual-usage meaning of words by vector-based representations of text. Any meanings or relationships discovered by LSA are then applied to a larger group of text (Landauer, Foltz, and Laham 2). To find these relationships, LSA uses vector algebra to convert documents into semantic space using the number of occurrences of unique words, sentences, and paragraphs that vectoring algebra can then use to determine the semantic similarity of documents or terms being applied. In LSA, vectorial algebra begins this process by using a document-term matrix to organize the frequency of terms in each piece of text being analyzed such as the one shown below in Table A.1. This table shows that each unique word in both sentences is used to populate the top row while a count of the frequency of each word is totaled in the cell corresponding to the text it is contained in. The goal of using such a model with vector mathematics is to represent the topic of a text by the frequency of semantically significant terms.

S1: I love reading.

S2: I hate hate hate reading.

Table A.1 An Example of a Document-term Matrix Illustrating the Distribution of Data (Landauer)

	I	love	hate	reading
S1	1	1	0	1
S2	1	0	3	1

The data in the analyzed corpus can be sentences, paragraphs, or pages that are then assigned points based on their contextual usage and applied to a mathematical

matrix for processing. Singular value decom mathematics follows to reduce the possible size of this matrix and better enable it to be processed by Eigen analysis, factor analysis, principal component analysis, and linear neural networks (Dumais 191-193). This dimensional reduction in the LSA process is the most important step and is relatively different from applications of vector-based mathematics. It cuts the matrix data down to only the relevant values that are required for an analysis. This is also the basis for the multi-dimensional vector space needed for LSA of sentiment.

After the document-term matrix has been created from the target body of text, the vector math process begins by creating an *m* by *n* matrix (equal to *A*) where *m* is the number of unique terms in the set of documents being examined by LSA and *n* is the number of documents. In essence, the process starts with a matrix like what is shown below. Remember that each row and column can hold X number of items, the 2 x 2 matrix below is just representing what will be a much larger matrix. This process is shown in the equations below.

$$\text{A} = \begin{matrix} \text{m} \\ \begin{bmatrix} X_{1,1} & X_{1,a} \\ X_{b,1} & X_{a,b} \end{bmatrix} \\ \text{n} \end{matrix} \qquad \text{A.2}$$

Common words such as *is, it, are, a,* etc. (also called stop words) are typically left out of latent semantic analysis. Vectors can now be formed from the rows and columns of this matrix. Every row in the matrix represents a unique term and its representation in a document, while every column represents a document and all the terms it contains.

Once the matrix has been created, a matrix decomposition technique known as Singular Value Decomposition (SVD) is used to create three additional matrices that separate~~s~~ the meaningful data. These new matrices are represented as:

$$A = TSD^T \quad \text{A.3}$$

Where: *A* is the term by document matrix, *T* is the left singular vectors in the matrix, *S* is the diagonal matrix of singular *R* values (rank), and *D* is the right singular vectors. The *S* value is where the reduction of the size of the original matrix *A* takes place. The lowest ranking values are removed from this value leaving a dimensional approximation of *A*, shown as:

$$A \approx \widetilde{A_k} = T_K S_K D_K^T \quad \text{A.4}$$

This equation turns each vector representing a document or term into an approximate dimensional (k). A key function of this process is the assumption that there is a structure or relationship to be found in the set of documents (Deerwater et al.)

Now that text and terms have been analyzed, it can then be compared to the comparison matrix. This comparison matrix is created in the same manner as our text matrix above, but the comparison matrix is what is being used to determine if specific features and relationships are found in *A*. These relationships are found through comparing both matrices by testing the angle of two vectors, one from each matrix. The value (cosine) is created from this comparison will tell us to what degree a relationship exists between a term in one set of documents to the same term in the comparison set of documents. This process is shown mathematically by the following expression:

$$\cos(\theta) = \frac{\overrightarrow{d_1} \cdot \overrightarrow{d_1}}{\|\|d_1\| \, \|d_1\|\|} \quad \text{A.5}$$

The larger the value of cosine as it approaches a value of one indicating more semantically similar documents, paragraphs, sentences, or words (127-128). This lengthy process is typically the reason for machine learning software performing the large number of calculation required by a representative document sample size.

REFERENCES

Ashton, Michael, and Kibeom Lee. "A Defence of the Lexical Approach to the Study of Personality Structure." *European Journal of Personality* 19 (2005): 5-24. Print.

Attali, Yigal. "Validity and Reliability of Automated Essay Scoring." *Handbook of Automated Essay Evaluation*. Ed. Mark Shermis and Jill Burstein. NY: Routledge, 2013: 181-98. Print.

Arvan, Marcus. "New Study Confirms and Extends Earlier Finding Linking Socially-Conservative Value Judgments to Anti-Social Personality Traits." *PRWeb*, 6 Mar. 2012. Web. 1 Sept. 2013.

Barriga, Claudia. *Morality and Movies: What Are People Thinking? A Content Analysis of Informal Movie Reviews Online*. Thesis. Cornell University, NY, 2007. Cornell University Library, *Thesis and Dissertations*. Web. 10 Aug. 2013.

Barrio, Victoria Del, Anton Aluja, and Luis F. García. "Relationship Between Empathy And The Big Five Personality Traits In A Sample Of Spanish Adolescents." *Social Behavior and Personality: An International Journal* 32.7, 2004: 677-81. Print.

Bates, Jordan, Jennifer Neville, and Jim Tyler. *Using Latent Communication Styles to Predict Individual Characteristics*. West Lafayette, IN: Psychological Sciences Department, Purdue University, Portable Document Format (PDF) File, Report.

Bennett, Randy E. "On the Meanings of Constructed Response." *Construction versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment*. Ed. Randy E. Bennett and William C. Ward. Hillsdale, NJ: L. Erlbaum Associates, 1993. 1-27. Print.

Bernard, H. Russell. *Social Research Methods: Qualitative and Quantitative Approaches*. Los Angeles, CA: Sage Publications, 2013. Print.

"Big Five Personality Traits." *Seven Counties Services Inc*. CenterSite, LLC, Web. 2 Sept. 2013.

Buchanan, Tom. "Five Factor Personality Test." University of Westminster, 2009. Personality Test. Web. 2 Sept. 2013.

Burstein, Jill, Beata Beigman-Klebanov, Nitin Madnani, and Adam Faulkner. "Automated Sentiment Analysis for Essay Evaluation." *Hand of Automated Essay Evaluation*. Ed. Mark Shermis and Jill Burstein. NY: Routledge, 2013: 281-97. Print.

Burstein, Jill. "Automated Essay Evaluation and Scoring." *The Encyclopedia of Applied Linguistics*, Ed. Carol Chapelle. Blackwell, OK, 2013. Print.

Burstein, Jill, Joel Tetreault, and Nitin Madnani. "Automated Essay Evaluation and Teaching Writing." *Handbook of Automated Essay Evaluation*. Ed. Mark Shermis and Jill Burstein. NY: Routledge, 2013: 55-67. Print.

Connolly, Thomas, Vladislav Veksler, and Wayne Gray. *Predicting Interest: Another Use for Latent Semantic Analysis*. Troy, NY: Department of Cognitive Science, Rensselaer Polytechnic Institute, 2009. Portable Document Format (PDF) File, Research Report.

Conard, M. "Aptitude Is Not Enough: How Personality and Behavior Predict Academic Performance." *Journal of Research in Personality* 40.3. 2006: 339-46. Print.

Deerwester, Scott, Susan Dumais, George Furnas, Richard Harshman, Thomas Landauer, Karen Lochbaum, and Lynn Streeter. Computer Information Retrieval Using Latent Semantic Structure. Bell Communications Research, Inc., assignee. Patent 4,839,853. 15 Sept. 1988. Print.

Dumais, Susan T. "Latent Semantic Analysis." *Annual Review of Information Science and Technology* 38.1, 2004: 188-230. Print.

Elliot, Norbert, and Andrew Klobucar. "Automated Essay Evaluation and Teaching Writing." *Handbook of Automated Essay Evaluation*. Ed. Mark Shermis and Jill Burstein. NY: Routledge, 2013: 16-35. Print.

"ETS Fast Facts." *About ETS: Factsheet*. Educational Testing Service, Web. 02 Oct. 2013.

Gallos, J. V. "Understanding the Organizational Behavior Classroom: An Application of Developmental Theory." *Journal of Management Education* 17.4, 1993: 423-39. Print.

Gosling, S. "A Very Brief Measure of the Big-Five Personality Domains." *Journal of Research in Personality* 37.6, 2003: 504-28. Print.

Hall, Mark, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian Witten. "The WEKA Data Mining Software: An Update." *Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) Explorations* 11.1, 2009 Print.

Jorge-Botana, Guillermo, Jose Leon, Ricardo Olmos, and Inmaculada Escudero. "Latent Semantic Analysis Parameters for Essay Evaluation Using Small-Scale Corpora*." *Journal of Quantitative Linguistics* 17.1, 2010: 1-29. Print.

Klobucar, Andrew, Norbert Elliot, Perry Deess, and Oleksandr Rudniy. "Automated Scoring in Context: Rapid Assessment for Placed Students." *Assessing Writing* 18.1, 2013: 62-84. Print.

Landauer, Thomas, Peter Foltz, and Darrell Laham. "An Introduction to Latent Semantic Analysis." *Discourse Processes* 25, 1998: 259-84. Print.

Layfield, Colin. *With LSA Size DOES Matter*. Msida: University of Malta: Department of CIS, 2012. Portable Document Format (PDF) File, Research Report.

Lenton, Alison P., Constantine Sedikides, and Martin Bruder. "A Latent Semantic Analysis of Gender Stereotype-Consistency and Narrowness in American English." *Sex Roles* 60.3-4, 2009: 269-78. Print.

Lottridge, Susan, Matthew Schulz, and Howard Mitzel. "Validity and Reliability of Automated Essay Scoring." *Handbook of Automated Essay Evaluation*. Ed. Mark Shermis and Jill Burstein. NY: Routledge, 2013: 233-50. Print.

Lowry, Richard. *VassarStats*. *Kappa as a Measure of Concordance in Categorical Sorting*. Richard Lowry. Web. 20 Sept. 2013.

Mayfield, Elijah. *Comma-separated Value* (*CSV)*. Test Data. Pittsburgh: LightSIDE Labs, MovieReviews.

Mayfield, Elijah. "LSA Algorithims in LightSIDE." Structured Telephone Interview about LightSIDE Processes. 16 Sept. 2013.

Mayfield, Elijah. "Student Inquiry about Using LightSIDE in a Research Project." Message to the Author. 10 Dec. 2012. E-mail.

Mayfield, Elijah, and Carolyn Rose. "Applications of Automated Essay Evaluation in West Virginia." *Handbook of Automated Essay Evaluation*. Ed. Mark Shermis and Jill Burstein. NY: Routledge, 2013: 124-35. Print.

Mayfield, Elijah, David Adamson, and Carolyn Rosé. *LightSIDE Research Workbench User's Manual*. Pittsburgh, PA: LightSIDE Labs, 2013. Portable Document Format (PDF) File, User Manual.

O'Connor, M., and S. Paunonen. "Big Five Personality Predictors of Post-secondary Academic Performance." *Personality and Individual Differences* 43.5, 2007: 971-90. Print.

Oberlander, Jon, and Scott Nowson. *Whose Thumb Is It Anyway? Classifying Author Personality from Weblog Text*. Sydney: Proceedings of the Coling/Association for

Computational Linguistics 2006 Main Conference Poster Sessions, July 2006. Portable Document Format (PDF) File, Conference Report.

Olney, Andrew. *Generalizing Latent Semantic Analysis*. Memphis, USA: Institute for Intelligent Systems University of Memphis, TN, 2009, Portable Document Format (PDF) File, Overview Report.

Pellegrino, James W., and Margaret L. Hilton. *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, D.C.: National Academies, 2012. Print.

Petersen, Michael, Morten Mørup, and Lars Hansen. *Latent Semantics as Cognitive Components*: 2nd International Workshop on Cognitive Information Processing, Lyngby, Denmark 2010. Portable Document Format (PDF) File, Conference Overview Report.

Reich, Justin. "Grading Automated Essay Scoring Programs- Part I." *Education Week*. Editorial Projects in Education, 14 Apr. 2013. Web. 2 Sept. 2013.

Rich, Changhua, Christina Schneider, and Juan D'Brot. "Applications of Automated Essay Evaluation in West Virginia." *Handbook of Automated Essay Evaluation*. Ed. Mark Shermis and Jill Burstein. NY: Routledge, 2013: 99-123. Print.

Ridgell, Susan, and John Lounsbury. "Predicting Academic Success: General Intelligence, "Big Five" Personality Traits, and Work Drive." *College Student Journal* 38.4 (2004) *Questia*. Web. 7 Aug. 2013.

Roberts, Brent. "How Conscientious Are You?" University of Illinois at Urbana-Champaign, 2009. Web. 10 Sept. 2013, Portable Document Format (PDF) File, Overview Report.

Roberts, Brent. "Personality, Leadership, and Self-Esteem." University of Illinois at Urbana-Champaign, IL, 2009. Article. Web. 10 Sept. 2013.

Rosé, Carolyn, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. "Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-supported Collaborative Learning." *International Journal of Computer-Supported Collaborative Learning* 3.3, 2008: 237-71. Print.

Srivastava, Sanjay. "The Five-Factor Model Describes the Structure of Social Perceptions. " *Psychological Inquiry*. Eugene, OR: University of Oregon, Portable Document Format (PDF) File, Research Report.

Sedlacek, William. "Employing Noncognitive Variables in the Admission and Retention of Nontraditional Students." *Achieving Diversity: Issues in the Recruitment and*

Retention of Underrepresented Racial/ethnic Students in Higher Education, 1993: 33-39. Print.

---. "Why We Should Use Noncognitive Variables with Graduate and Professional Students." *The Advisor: The Journal of the National Association of Advisors for the Health Professions* 24.2, 2004: 32-39. Print.

Shermis, Mark, Jill Burstein, and Sharon Bursky. "Introduction to Automated Essay Evaluation." *Handbook of Automated Essay Evaluation*. Ed. Mark Shermis and Jill Burstein. NY: Routledge, 2013: 1-15. Print.

Shermis, Mark, and Ben Hamner. *Contrasting State-of-the-Art Automated Scoring of Essays: Analysis*: The University of Akron, OH, 2012. Portable Document Format (PDF) File, Technical Report.

Stricker, Lawrence "'Test-wiseness' on Personality Scales." *Journal of Applied Psychology* 53.3, Pt.2, 1969: 1-17. Print.

Stricker, Lawrence, Gita Z. Wilder, and Brent Bridgeman. "Test Takers' Attitudes and Beliefs About the Graduate Management Admission Test." *International Journal of Testing* 6.3, 2006: 255-68. Print.

Stricker, Lawrence, Gita Wilder, and Donald Rock. "Attitudes about the Computer-based Test of English as a Foreign Language." *Computers in Human Behavior* 20.1, 2004: 37-54. Print.

Taboada, Maite. "Stages in an Online Review Genre*." *Text and Talk* 31.2, 2011: 247-69. Print.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. "Lexicon-Based Methods for Sentiment Analysis." *Computational Linguistics* 37., 2011: 268-308. Print.

Tkacik, Daniel. "Researchers Develop LightSIDE Program for Grading Essays." *The Tartan Online*. The Tartan, 23 Apr. 2012. Web. 19 Mar. 2013.

Williams, K.M., Hutchinson Opren, L.J. Walker, and B.D. Zumbo. *Personality, Empathy, and Moral Development: Examining Ethical Reasoning in Relation to the Big Five and the Dark Triad*. Calgary: University of British Columbia, Canada, 2006. Portable Document Format (PDF) File, Technical Report.

Williamson, David, Xiaoming Xi, and F. Jay Breyer. "A Framework for Evaluation and Use of Automated Scoring." *Educational Measurement: Issues and Practice* 31.1, 2012: 2-13. Print.

Wolfe, Michael, and Susan Goldman. "Use of Latent Semantic Analysis for Predicting Psychological Phenomena: Two Issues and Proposed Solutions." *Behavior Research Methods* 35, 2003: 22-31. Print.

Yarkoni, Tal. "Personality in 100,000 Words: A Large-scale Analysis of Personality and Word Use among Bloggers." *Journal of Research in Personality* 44.3, 2010: 363-73. Print.

Yang, Yongwei, Chad W. Buckendahl, Piotr J. Juszkiewicz, and Dennison S. Bhola. "A Review of Strategies for Validating Computer-Automated Scoring." *Applied Measurement in Education* 15.4, 2002: 391-412. Print.